A Practical Logic of Cognitive Systems
Volume 2

The Reach of Abduction
Insight and Trial

This Page is Intentionally Left Blank

A Practical Logic of Cognitive Systems
Volume 2

The Reach of Abduction
Insight and Trial

Dov M. Gabbay
Department of Computer Science
King's College London
Strand, London, WC2R 2LS, U.K.

and

John Woods
Philosophy Department
University of British Columbia
Vancouver, BC Canada, V6T 1Z1
&
Department of Computer Science
King's College London
Strand, London, WC2R 2LS, U.K.

2005

ELSEVIER

Amsterdam - Boston - Heidelberg - London - New York - Oxford - Paris
San Diego - San Francisco - Singapore - Sydney - Tokyo

ELSEVIER B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE
Amsterdam, The Netherlands

ELSEVIER Inc.
525 B Street
Suite 1900, San Diego
CA 92101-4495, USA

ELSEVIER Ltd.
The Boulevard
Langford Lane, Kidlington,
Oxford OX5 1GB, UK

ELSEVIER Ltd.
84 Theobalds Road
London WC1X 8RR
UK

Notice
No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

First edition 2005

Library of Congress Cataloging in Publication Data
A catalog record is available from the Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN: 0-444-51791-X

∞ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

In Memoriam

Raymond Reiter
12 June 1939–16 September 2002

Víctor Sánchez-Valencia
10 January 1951–7 June 2003

This Page is Intentionally Left Blank

Contents

Acknowledgements

The present work arises from conversations the authors had at the 1999 meeting of the Society for Exact Philosophy in Lethbridge, organized by Bryson Brown. Early results were presented to a seminar on abduction in June 1999 in the Department of Computing and Electrical Engineering, Heriot-Watt University, arranged by Fairouz Kamareddine, and to the Conference on the Frontiers of Logic at King's College, London in November 1999, organized by Dov Gabbay. Additional work formed the basis of presentations in May 2000 to the conference on Modelling at the University of Lethbridge in June 2000, presided over by Paul Viminitz. Further results were worked up at the University of British Columbia in the Spring of 2000, where Gabbay was Visiting Fellow of the Wall Institute for Advanced Study and Woods a visitor in the Noted Scholars Programme. Arrangements in Vancouver were organized by Mohan Matthen.

A first draft of the book served as the text for a course the authors taught at the European Summer School on Logic, Language and Information at the University of Birmingham in August 2000. Achim Jung was responsible for the organization of these meetings.

Subsequent results were discussed at the Conference on Formal and Applied Practical Reasoning, Imperial College, London in September 2000, organized by Jim Cunningham and Johan van Benthem; at meetings of the Ontario Society for Study of Argumentation at the University of Windsor in May 2001, organized by Hans Hansen and Christopher Tindale; and at the University of Groningen where Woods was the Vonhoff Professor in the Spring Term of 2001. Arrangements in Groningen were made by Erik Krabbe and Theo Kuipers. Further revisions were presented to the Conference on the Logic of Rational Agency, in Dagstühl in December 2001. Mike Woolridge and Wiebe van der Hoek were the organizers. Woods also acknowledges the support of Frans van Eemeren's Vakgroep Taalbeheersing, Argentatietheorie an Rhetorica in the University of Amsterdam where he was Visiting Professor in the Spring and Summer 1999, and most years since 1987.

Two lectures were presented to the Conference on Mistakes of Reason, at the University of Lethbridge in 2002. Kent Peacock and Andrew Irvine were in charge of arrangements. More recent developments were presented as a lecture to the Cognitive Science Society, in Chicago, August 2004, organized by Lorenzo Magnani, and to the De Morgan Workshop on Logic and Law at King's College London in November 2004, organized by the authors and their colleagues Andrew Jones and Matthias Baaz. Additional materials were presented in two lectures delivered to the Conference on Model-Based Reasoning in Science and Engineering: Abduction, Visualization and Simulation, in December 2004 in Pavia. Lorenzo Magnani was in charge of the Programme.

Alex Boston was Woods' research assistant at UBC, as were Dawn Collins and Brian Hepburn at the University of Lethbridge, and Carol Woods in Vancouver. Brian Hepburn also formatted the early drafts of the book in LaTeX. The project was supported in London by Jane Spurr, Publications Administrator to Dov Gabbay. The authors express their warm thanks to all.

Woods's research is supported by grants from the Social Science and Humanities Research Council of Canada; the University of Lethbridge Research Fund; the Engineering and Physical Sciences Research Council of the United Kingdom, of which he was Visiting Fellow in the academic years 2003-2004; the Department of Computer Science at King's College London; where he is the Charles S. Peirce Visiting Professor of Logic; and the former Dean of Arts and Science of the University of Lethbridge, Professor Bhagwan Dua, and his successor, Christopher Nicol; Professor Margaret Schabas, Head, Department of Philosophy and Professor Nancy Gallini, Dean of Arts, University of British Columbia, to whom the authors owe a particular debt of gratitude. Thanks are also due, and happily given, to our colleagues at Elsevier, Arjen Sevenster our Publishing Editor and Andy Deelen, Production Associate.

Gabbay's research for this book was partially supported by the European Community's PLATO PROJECT, a joint project with Edinburgh University. The objective of this project is to develop logical tools and a goal directed database for modelling some of the arguments of Plato. The mechanism of abduction plays a central role in such models. Further support is acknowledged from the Alexander von Humboldt Foundations which provided a generous fellowship for work on the logics of practical reasoning. We are also grateful to King's College London for a Sabbatical leave allowing for the research on practical reasoning mechanisms, and finally both authors are grateful for the EPSRC fellowship (Grant GR/R59564/01) supporting Woods's visits to King's College in 2002 and 2003–04.

Parts of chapter 2 are drawn or adapted from chapter 2 of *Agenda Relevance* [Gabbay and Woods, 2003a], as are section 6.7 and chapter 11. We are grateful to our publisher, Elsevier, for permission to reprint this material.

A word of special thanks to Atocha Aliseda, and Paul Bartha for the courtesy of allowing us to benefit from pre-publication versions of their respective books *Abductive Reasoning: Logical Investigations into the Process of Discovery and Evaluation*, and *Analogical Reasoning and Plausibility in the Sciences*; and to Nicholas Rescher for the stimulation afforded by his pioneering studies in logic and cognitive rationality.

In addition to those already mentioned, we are grateful for discussions bearing on our work to Tom Addis, Ardis Anderson, Bradley Armour-Garb, David Atkinson, Diderik Batens, John Beatty, Krystia Broda, Peter Bruza, Balakishnam Chandrasekaran, Jim Cunningham, Donald Davidson, Stephen Friesen, Artur Garcez, Donald Gillies, Alvin Goldman, Nick Griffin, Jaakko Hintikka, David Hitchcock, Dale Jacquette, Ruth Kempson, Shalom Lappin, Dominic McIvor Lopesz, Peter McBurney, Tom Maibaum, David Makinson, Joke Meheus, Rolf Nossum, Sami Paavolva, Jeanne Peijnenberg, Gabriella Pigozzi, David Poole, Graham Priest, Michael Rabin, Chris Reed, Ray Reiter, Odinaldo Rodriques, Victor Rodych, Menno Rol, Jan-Willem Romeyn, Patrick Rysiew, Victor Sánchez Valencia, Matti Sintonen, Roy Sorensen, Patrick Suppes, Jason Tangen, Alice ter Meulen, Alexander van den Bosch, Bas van Fraassen, Jan Albert van Laar, John Vokey, Joe Wells and Jon Williamson.

This Page is Intentionally Left Blank

Preface

Although *A Practical Logic of Cognitive Systems* exhibits some common themes, we have written the individual volumes with a view to their being read either as stand-alone works or as linked and somewhat overlapping items in the series, depending on the interests of particular readers. Relevance was our main theme in volume one; abduction will occupy us in the present volume; and volume three will concern itself with fallacious reasoning. Here too, we intend to honour the pledge of independent readability. Even so, certain continuities will also be evident in all volumes, of which the first and foremost is what we suggest about the structure of practical reasoning. In some cases, it will be unavoidable that we repeat a point made in a predecessor volume. Sometimes we will elaborate upon a prior point. On occasion, we will correct what we now see as a mistake.

In writing our predecessor volume on relevance, we were mindful of two approaches to the subject that had attained dominant purchase. One is the output of a generation's research on relevant logic, ensuing from the work of Alan Ross Anderson and Nuel D. Belnap, Jr., beginning in the late 1950s. The other is the theory of the communication theorists, Dan Sperber and Deirdre Wilson, whose influential pragmatic account appeared in 1986. We did not want to write a derivative book; neither were we much attracted by the prospects of polemical attack. We desired to take an approach that at once recognized the significance of the dominant views, while attempting to advance beyond them in substantial measure.

Abduction faces us with a somewhat different challenge. No less central a factor in practical reasoning than relevance, we trust that we give no offence in observing that the abductive landscape is not yet presided over by dominant theoretical presences, in the manner of relevance. A possible exception to this are the scattered contributions by the modern founder of abductive logic, Charles Peirce. Peirce's sallies are indeed seminal, and dotted with some brilliantly original insights. But unlike the cumulative record of modern relevant logicians and the detailed theoretical articulation of Sperber's and Wilson's account, Peirce left the logic of abduction in a comparatively undeveloped state. It is true that there is by now a large literature on abduction, created by an impressive number of authors from philosophy, cognitive psychology, computer science, artificial intelligence and, of course, logic. From philosophy alone it may be suggested that, contrary to our present suggestion, an important approach has indeed presented itself in the literature that has grown up around Gilbert Harman's significant paper from 1965 on inference to the best explanation. There can be no doubt that inference to the best explanation is an important idea which has been ably probed by a generally sophisticated literature. Even so, we are not quite ready to accede to a dominance that is more arguably to be found in the literature on relevance. There are three reasons for this reluctance. One is that various kinds of abductive practice have nothing to do with achieving explanations. Another is that even in those cases in which abduction has an explanationist character, the factor of explanation is but a part, albeit an important part, of the abductive pie. Thirdly, in some versions of it, inference to the best explanation is not abductive, surprising as that may strike us initially.

If we are correct in these observations, abduction is a more wide-open field than relevance. For the would-be theorist this is an advantage and a disadvantage. The advantage is that achieving a dominant position is, in principle, a target still to be aimed at. The disadvantage is that there are fewer stout shoulders on which the theorist might secure a purchase. Still, we don't wish to leave the impression that the abductive theorist's is a voice in a solitary wilderness. There is much good work that has already been published, of which three recent examples are [Aliseda, forthcoming; Magnani, 2001a] and [Meheus *et al.*, forthcoming].

The comparative openness of the logic of abduction makes a book such as this in like degree an enterprise of first words rather than last. Even in what we think we have already come to understand about abduction, there is ample discouragement of the idea that all of abduction can be gobbled up in a single try. Accordingly, the best we can hope for is new ground decisively broken in ways that portend favourably for the grand theory, whenever it appears.

Part I

A Practical Logic of Cognitive Systems

This Page is Intentionally Left Blank

Chapter 1

Introduction

It is sometimes said that the highest philosophical gift is to invent important new philosophical problems. If so, Peirce is a major star on the firmament of philosophy. By thrusting the notion of abduction to the forefront of philosophers' consciousness he created a problem which — I will argue — is the central one in contemporary epistemology.

Jaakko Hintikka,

The surprising fact C is observed. But if A were true, C would be a matter of course. Hence there is reason to suspect that A is true.

Charles S. Peirce

Abduction is our subject here. We meet it in a state of heightened theoretical activity. It is part of the contemporary research programmes of logic, cognitive science, AI, logic programming, and the philosophy of science. This is a welcome development. It gives us multiple places to look for instruction and guidance.

The approach that we take in this book is broadly logical. Any fears, even so, that this will be an over-narrow orientation may be allayed by our decision to define logic as the disciplined description of the behaviour of real-life logical agents. In this we command a theme that has played since antiquity: that logic is an account of how thinking agents reason and argue. Because we wish to give due attention to the process side of the process-product distinction, we propose a rapprochement between logic and psychology, with a special emphasis on developments in cognitive science. It would be foolish to suggest that the hugely profitable theoretical attainments of modern mathematical logic have no place in an agent-based, psychologically realistic account of abduction. The rich yield in the four pillars of post-Fregean logic — proof theory, set theory, model theory and recursion theory — are bench marks of intellectual achievement. But because these results bear on properties of linguistic structures independently of the involvement of an agent, or (as in proof theory) in ways involving only highly abstract intimations of agency, the theorems of standard mathematical logic tend not to have much in the way of *direct* application to what thinking agents do when they reason. Indirect relevance is another matter. Part of the story of an agent's discharge of his cognitive agenda will involve his taking into account (perhaps only tacitly at times) properties of linguistic structures, such as consistency and consequence. Accordingly, we propose to absorb the logic of linguistic structures into a more comprehensive logic of agency. In so doing, we are mindful of various ways in which to 'agentify' logics of linguistic structures, whether by way of natural deduction protocols, relevance constraints, Prolog with restart and diminishing resources, N-Prolog with bounded restart for intuitionistic logic, action and time modalities, labelled instructions to presumed agents in proof theoretic environments (e.g., labelled deductive systems), "soft" consequence (as in non-monotonic, default and defeasible logics), enriching semantic models (e.g., as in fibred semantics), and applications to fallacious reasoning (e.g., the analysis of *petito principii* by way of intuitionistic modal logic). Also important in this regard is the structuralist approach to logic ensuing from [Gentzen, 1935, pp. 176–210, 149–167], and given new impetus by [Scott, 1972]. Such logics have been imaginatively exploited by AI researchers in developing approaches to plausible reasoning [Kraus *et al.*, 1990], as well as more general frameworks for non-monotonic consequence relations [Gabbay, 1985b]. Structural logics have also been appropriated to advantage by the research community in dynamic semantics, where a variety of consequence relations have been analyzed non-monotonically [van Benthem, 1996; Port and van Gelder, 1995] and [Gochet, 2002]. It is these developments collectively to which the name of *the new logic* has been applied [Gabbay and Woods, 2001a]. It is a conception of logic especially suited to our purposes here.

In seeking a reconciliation with psychology, we risk the opprobrium of those who, like Frege, abjure psychologism in logic. It may surprise the reader to learn that in this we are entirely at one with Frege. If logic is confined to an examination of propositional structures independently of considerations of agency and contexts of use, then psychology has no place in logic. If, moreover, psychologism were to require that logic be thought of as just another experimental science, then, with Frege, we would part company with psychologism. If, however, it is legitimate to regard a logic as furnishing formal models of certain aspects of the cognitive behaviour of logical agents, then not only do psychological considerations have a defensible place, they cannot reasonably be excluded.[1]

[1] Psychologism is discussed in greater detail in [Gabbay and Woods, 2005]. See also [Brockhaus, 1991] and Jacquette [2003a; 2003c; 2003b].

If, as we think, the distinction between logic and psychology is neither sharp nor exhaustive, it is necessary to say what a logic, as opposed to a psychology, of abduction, or of any inferential practice of a logical agent, would be. In this regard, we see ourselves in the historical mainstream of logic. From its inception, logic has been an account of target properties of linguistic structures, or of relations between linguistic and set theoretic structures, which are then adapted for use in the larger precincts of reasoning and arguing. Aristotle invented syllogistic logic to serve as the theoretical core of a wholly general and psychologically real theory of argument and inference. Doing so led him to 'agentify' the defining conditions on the syllogistic logic of linguistic structures in ways that made it, in effect, the first ever intuitionistic, non-monotonic, relevant (hence paraconsistent) logic [Woods, 2001]. Aristotle developed the syllogistic constraints in ways that took explicit recognition of agency, action and interaction. The father of logic, therefore, ran the subject's broad programme by adapting the logic of linguistic structures to the more comprehensive logic of inference and argument. What made it appropriate to consider the larger project as logic was that it was a principled extension and adaptation of the theory of those core structures. It is something in this same sense that ours too is a work of logic. We seek an account of what abductive agents do, in part by extension and adaptation of accounts of properties of linguistic structures, which themselves have been set up in ways that facilitate the adaptation and anticipate the approximate psychological reality of the requisite formal models. Oddly, then, what we are calling the new logic is in its general methodological outlines not only the old logic, but logic as it has been with few exceptions throughout all of its long history.

For us there is a point of departure. Although we concede that key logical properties of propositional structures play a central role in the broader logic of argument and inference, it is also our view that this broader logic must also take into account the fact, attested to by science and intuition alike, that real-life reasoning in real time is not always a matter of conscious symbol-processing.

Cognitive science engages our logical interests in its two main branches. Like cognitive neuropsychology, the new logic has an interest in modelling natural cognition. Like artificial intelligence, the new logic seeks an understanding of model of artificial intelligence with extensions to robotics and artificial life. In both domains it remains a characteristic contribution of cognitive science to conceive of cognition as computational; so its emphasis is on computer modelling of cognition, whether artificial or biological.

Much of the impetus for theoretical work on abduction comes from the philosophy of science. Philosophy of science was, in effect, invented by the Vienna Circle and the Berlin Group in the first third of the century just past. For the thirty or forty years after the mid-thirties, philosophy of science was a foundational enterprise. It proposed a logic of science as giving the logical syntax of the language of (all) science. It was an approach that disclaimed as excessively psychological the context of discovery (as Reichenbach called it), and concentrated on the context of justification (also Reichenbach). It was an imperious rather than case-based orientation, in which the flow of enquiry is top-down, from logic to the philosophy of science, and thence to science itself. And its analytical method was that of the explication of concepts.

In the past decades, philosophy of science has evolved from its original stance to its present non-foundational or *neoclassical* [Kuipers, 2000]. In it a place is provided for the contest of discovery, and there is much less attention paid to the (presumed) logical syntax of science and more on the construction of mathematical models of the conceptual products of natural scientific theorizing. Accordingly, the flow of enquiry is bottom-up, from science, to the philosophy of science, to the logic of science.

As conceived of in the neoclassical orientation, the logic of science provides formal idealized models of the most general and central features of the conceptual content of the process of scientific reasoning. If we remove the qualification 'scientific', the logic which the new classical logic of science exemplifies resembles our own more general conception of logic, all the more so when the modelling dynamics are allowed to range over thinking processes themselves, rather than just the conceptual products of such processes.

We have remarked that a distinguishing feature of present-day cognitive science is its tendency to regard cognition as computational. It is not however an essential condition on cognitive science that it eschew formal modelling for computational articulation. This is a latitude that we intend to make something of. Although we shall enquire into computer implementation of accounts of abductive reasoning, our main technical thrust will be by way of formal models. In this respect we propose to retain a distinctive and methodologically powerful feature of mathematical logic, even though the reach of our models extends beyond sets of properties of agent-independent systems, as we have said.

This bears on a second feature of this book, namely, the necessity to specify inputs to our formalizing devices. In a rough and ready way, these can only be what is known informally about how agents actually reason. In a well-entrenched analytical tradition, inputs are made accessible by the theorist's intuitions about the subject at hand, which stand as his *data* for a theory of it. Intuitions are the conceptual data for theory, the inputs to the theory's formal models. We ourselves are disposed to take this same approach to inputs, but we do so only after taking to heart Bacon's admonition against the use of undisciplined observations in science, and similar advice from Suppes, who emphasizes the importance of presenting one's formalizing devices with *models* of the data [Suppes, 1962] and [Starmans, 1996, ch. 4]. Models of the data constitute the conceptual content of our theory of abduction, which serves in turn as inputs to formalization. Accordingly, models of the data can be conceived of as *conceptual models*. To achieve the requisite conceptualization, we employ a methodological tool which was characteristic of the old philosophy of science, and which has long served the broader purposes of analytic philosophy. This is the methodology of *concept explication* or *conceptual analysis*. Given that our project is a logic of abduction, concepts for explication or conceptual analysis include the obvious ones: abduction itself, explanation, relevance, plausibility, presumption and analogy, among others. In treating of these notions, the explication should leave the explicated concept recognizably present; it should preserve clear cases; it should be precisely specified; it should be as straightforward as the most important of the actual complexities allow; and where possible it should stimulate new questions and motive new research. The process of concept explication produces an *informal theory* of abduction. It stands midway between the raw data of which they are a conceptual model, in the manner of Suppes and Starmans, and aggressively specified formalizations.

Formal modelling is a central part of our project. Models call for a certain caution, concerning which see [Gabbay and Woods, 2004] and [Gabbay and Woods, 2003b]. It is not too far off the mark to say that formalized models are empirically false representations of their theoretical targets. Not only do models lose information; they exchange accuracy and detail for precision and systematicity. The trade-off is economically advantageous when an agent's actual behaviour can fairly be called an approximation to the deviances sanctioned in the model (its norms, so to speak). But not every difference between fact and idealized norm is an approximation; so care needs to be taken.

Some ideal modelers are of the view that a model's norms attain their legitimacy and their authority because they achieve requisite standards of prescriptiveness. Thus norms tell us how we *should* reason, and actual practice reveals how we *do* reason, and therefore how close actual reasoning is to correct reasoning. We ourselves have little confidence in this picture. We are not of the view that prescriptive legitimacy is self-announcing. Accordingly we favour a more circumspect approach, by way of what we call the

> *Actually Happens Rule*: To see what agents should do, look first to what they actually do. Then repair the account if there is particular reason to do so.

In espousing the *Actually Happens Rule* we find ourselves loosely in the company of cognitive scientists who take the so-called "panglossian" approach to the analysis of human rationality [Stanovich, 1999]. (We return to this point in section 3.1 below.)

It is said that the neoclassical approach to the philosophy of science is neutral on the realism-instrumentalism question [Kuipers, 2000]. The ambiguity of the claim calls for a clarification. If it means that neoclassical philosophy of science doesn't adjudicate the rivalry between realism and instrumentalism in science, then the claim is correct. If it means that neoclassical philosophy of science doesn't acknowledge realist and/or instrumentalist manoeuvres discernible in actual scientific practice, then it is a claim from which we demur. We take the scientist and the everyday researcher as they come. We grant that human reasoning in general is embedded in realist presumptions, and yet we also acknowledge important episodes of self-proclaimed instrumentalist reasoning. (The early Gell-Mann was an unfettered instrumentalist about quarks. Later on, he lightened up considerably, encouraged by empirical developments. See here [Johnson, 1999].) The logic we want to construct is a logic for the human reasoner as he presents himself to and engages with the world. We leave philosophical validation of that stance to the epistemological experts. Even so, the logic of abduction is tied in an essential way to the kindred distinction between cognitivism and instrumentalism. As we shall be at pains to show, there is an important sense in which abductive reasoning instantiates non-cognitivist conditions on rational performance. Another, somewhat more paradoxical-sounding, way of saying the same thing is that when he performs abductively, there is an important sense in which the cognitive agent is required to operate in ways that are *epistemically sub-par*.

Concerning matters epistemological, we ourselves favour a naturalized non-foundationalist perspective, which can be seen as echoing the non-foundationalism of neo-classical philosophy of science.[2] Ours is a generically reliabilist approach to epistemology, in which what passes for knowledge is belief induced by the appropriate cognitive devices when functioning properly. It is not a view for which we here argue; others (e.g., [Millikan, 1984; Goldman, 1986]) have done so before us, in ways adequate for our present purposes. (But see also [Wouters, 1999].)

In the philosophy of science, perhaps constructive empiricism is the most natural *entré* to the logic of abduction. The principal epistemological tenet of constructive empiricism is that well established science makes true only the observational content of scientific theories. As for the rest – the theoretical postulates of such theories – the best that can be said of them is that they are integral to theories that are empirically adequate, and that the best that can be said about them is good enough for scientific legitimacy [van Fraassen, 1980]. Constructive empiricism has been vigorously attacked on grounds that it is incompatible with scientific realism. This is a widely entrenched criticism voiced by a large literature. For all its apparent appeal, this is not a criticism that we share. The failure of physics to verify the existence of quanta is insufficient to show that physics cannot reasonably suppose quanta to be real or that the other theoretical claims of quantum mechanics cannot reasonably be supposed to be true. There is a difference between what can be shown true and what can reasonably be supposed true. This is a difference

[2] In the manner of Quine's "Epsitemology Naturalized" [Quine, 1969a]. We note in passing that Quine's original subtitle for this celebrated paper is "The Case for Psychologism".

that matters for the logic of abduction in a central way. But it is not a difference that should discourage either the theoretical scientist or the abductive logician.

This Page is Intentionally Left Blank

Chapter 2

Practical Logic

> . . . for all the proclaimed rationality of modern humans and their institutions, logic touches comparatively little of human practice.
>
> Richard Sylvan

> [T]he limit on human intelligence up to now has been set by the size of the brain that will pass through the birth canal .... But within the next few years, I expect we will be able to grow babies outside the human body, so this limitation will be removed. Ultimately, however, increases in the size of the human brain through genetic engineering will come up against the problem that the body's chemical messengers responsible for our mental activity are relatively slow-moving. This means that further increases in the complexity of the brain will be at the expense of speed. We can be quick-witted or very intelligent, but not both.
>
> Stephen Hawking

2.1 First Thoughts on a Practical Logic

The theory of abduction that we develop in this volume is set up to meet two conditions. One is that it show how abduction plays within a practical logic of cognitive systems. The other is that, to the extent possible, it serve as an adequate stand-alone characterization of abduction itself. In the first instance we try to get the logic of cognitive systems right, though with specific attention to the operation of abduction. In the second instance, we try to get abduction right; and we postulate that our chances of so doing improve when the logic of abduction is lodged in this more comprehensive practical logic.

We open this chapter with a brief discussion of what we take such a logic to be. Readers who wish a detailed discussion can consult chapters 2 and 3 of the companion volume, *Agenda Relevance: A Study in Formal Pragmatics*. Other readers, who may be eager to get on with abduction without these prefatory remarks, can go directly to section 3.1.

In the prequel to this book we adopted a convention for flagging the more important of the claims and ideas advanced by our conceptual model of the relevance relation. Key claims that we were prepared to assert with some confidence we flagged as (numbered) definitions or propositions. Ideas that called for a greater tentativeness we flagged as (numbered) propositions prefixed with the symbol ♡. We here follow that same practice for abduction.

2.1.1 A Hierarchy of Agency Types

We take the position that reasoning is an aid to cognition, a logic, when conceived of as a theory of reasoning, must take this cognitive orientation deeply into account. Accordingly, we will say that a *cognitive system* is a triple of a cognitive agent, cognitive resources, and cognitive target performed in real time. (See here [Norman, 1993; Hutchins, 1995].) Correspondingly, a logic of a cognitive system is a principled description of conditions under which agents deploy resources in order to perform cognitive tasks. Such is a practical logic when the agent it describes is a *practical agent*. So, then,

Definition 2.1 (Cognitive systems) *A cognitive system CS is a triple X, R, A of a cognitive agent X, cognitive resources R, and a cognitive agenda A executed in real time.*

Definition 2.2 (Practical logics, a first pass) *A practical logic is a systematic account of aspects of the behaviour of a cognitive system in which X is a practical agent.*

A practical logic is but an instance of a more general conception of logic. The more general notion is reasoning that is target-motivated and resource-dependent. Correspondingly, a logic that deals with such reasoning is a Resource-Target Logic (*RT*-logic). In our use of the term, a practical logic is a *RT*-logic relativized to practical agents.

How agents perform is constrained in three crucial ways: in what they are disposed towards doing or have it in mind to do (i.e., their *agendas*); in what they are capable of doing (i.e., their *competence*); and in the means they have for converting competence into performance (i.e., their *resources*). Loosely speaking, agendas are programmes of action, exemplified by belief-revision and belief-update, decision-making and various kinds of case-making and criticism transacted by argument. For ease of exposition we classify this motley of practices under the generic heading "cognitive", and we extend the term to those agents whose practices these are.[1]

An account of cognitive practice should include an account of the *type of cognitive agent* involved. Agency-type is set by two complementary factors. One is the degree of command of resources an agent needs to advance or close his (or its) agendas. For cognitive agendas, three types of resources are especially important. They are (1) *information*, (2) *time*, and (3) *computational capacity*. The other factor is the height of the cognitive bar that the agent has set for himself. Seen this way, agency-types form a hierarchy H partially ordered by the relation C of commanding-greater-resources-in-support-of-higher-goals-than. H is a poset (a partially ordered set) fixed by the ordered pair $\langle C, X\rangle$ of the relation C on the set of agents X.

Human agency ranks low in H. If we impose a decision not to consider the question of membership in H of non-human primates, we could say that in the H-space humans are the lowest of the low. In the general case the cognitive resources of information, time and computational capacity are for human agents comparatively less abundant than for agents of higher type, and their cognitive goals are comparatively more modest. For large classes of cases, humans perform their cognitive tasks on the basis of less information and less time than they might otherwise like to have, and under limitations on the processing and manipulating of complexity. Even so, paucity must not be confused with scarcity.[2] There are cases galore in which an individual's resources are adequate for the attainment of the attendant goal. In a rough and ready way, we can say that the comparative modesty of an agent's cognitive goals inoculates him against cognitive-resource scarcity. But there are exceptions, of course.

Institutional entities contrast with human agents in all these respects. A research group usually has more information to work with than any individual, and more time at its disposal; and if the team has access to the appropriate computer networks, more fire-power than most individuals even with good PCs. The same is true, only more so, for agents placed higher in the hierarchy — for corporate actors such as NASA, and collective endeavours such as quantum physics since 1970. Similarly, the cognitive agendas that are typical of institutional agents are by and large stricter than the run-of-the-mill goals that motivate individual agents. In most things, NASA aims at stable levels of scientific confirmation, but, for individuals the defeasibly plausible often suffices for local circumstances.

These are vital differences. Agencies of higher rank can afford to give maximization more of a shot. They can wait long enough to make a try for total infor-

[1] Agendas are discussed at greater length in [Gabbay and Woods, 2003a].

[2] We have been guilty of this confusion in previous writings, notably in[Gabbay and Woods, 2003a].

mation, and they can run the calculations that close their agendas both powerfully and precisely. Individual agents stand conspicuously apart. For most tasks, the human cognitive agent is a satisficer. He must do his business with the information at hand, and, much of the time , sooner rather than later. Making do in a timely way with what he knows now is not just the only chance of achieving whatever degree of cognitive success is open to him as regards the agenda at hand; it may also be what is needed in order to avert unwelcome disutilities, or even death. (We do not, when seized by an onrushing tiger experience, wait before fleeing for a refutation of skepticism about the external world or a demonstration that the approaching beast is not an hallucination.)

Given the comparative humbleness of his place in H, the human individual is frequently faced with the need to practise cognitive economies. This is certainly so when either the loftiness of his goal or the supply of drawable resources create a cognitive strain. In such cases, he must turn *scantiness* to *advantage*. That is, he must (1) deal with his resource-limits and in so doing (2) must do his best not to kill himself. There is a tension in this dyad. The paucities with which the individual is chronically faced are often the natural enemy of getting things right, of producing accurate and justified answers to the questions posed by his agenda. And yet not only do human beings contrive to get most of what they do right enough not to be killed by it, they also in varying degrees prosper and flourish.

This being so, we postulate for the individual agent *slight-resource adjustment strategies (SRAS),* which he uses to advantage in dealing with the cognitive limitations that inhere in the paucities presently in view. We make this assumption in the spirit of Simon [1957] and an ensuing literature in psychology and economics. At the heart of this approach is the well-evidenced fact that, for ranges of cases, "fast and frugal" is almost as good as full optimization, and at much lower cost [Gigerenzer and Selten, 2001]. We shall not take time to detail the various conditions under which individuals extract outcome economies from resource limitations and target modesty, but the examples to follow will give some idea of how these strategies work.[3] We note that the hierarchical approach to agency gives us a principled hold on the distinction between *practical* and *theoretical* agents and, correspondingly, between practical and theoretical *reasoning*. Practical reasoning is reasoning done by a practical agent. An agent is a practical agent to the extent that it commands comparatively few cognitive resources in relation to comparatively modest cognitive goals. Theoretical reasoning is reasoning done by a theoretical agent. An agent is theoretical to the extent that it commands comparatively much in the way of cognitive resources, directed at comparatively strict goals. We have it, then, that

[3] For a fuller discussion, see also [Woods *et al.*, 2002] and [Gabbay and Woods, 2003a]; *cf.* [Sperber and Wilson, 1986].

Definition 2.3 (Hierarchy of agency types) *H is a hierarchy of agency types when H is a set of cognitive agents partially ordered by the (complex) relation of commanding more cognitive resources R in relation to higher cognitive goals than.*

Definition 2.4 (Practical agency) *A cognitive agent is a practical agent to the extent that he (or it) ranks low in H.*

Definition 2.5 (Theoretical agency) *A cognitive agent is a theoretical agent to the extent that it ranks high in H.*

Our interpretation of the practical-theoretical dichotomy may strike the reader as nonstandard, if not eccentric; on the face of it, there is no natural non-negative antonym of our use of the word "practical". We ourselves are prepared to put up with the nonstandardness in return for conceptual yield.

We have cautioned against the equation of resource-paucity with resource-scarcity. It is, of course, quite true that in some sense practical agents operate at a cognitive disadvantage. But it is advisable not to make too much of this. What should be emphasized is that in relation to the cognitive standards that an institutional agent might be expected to meet, the resources available to a practical agent will typically not enable him (or it) to achieve that standard. Whether this constitutes an unqualified disadvantage depends on the nature of the task the individual has set for himself and the cognitive resources available to him. For a practical agent to suffer an unqualified disadvantage, two factors must intersect in the appropriate way: his resources must be inadequate for the standard he should hit, in relation to a goal that has reasonably been set for him. So, then, the measure of an agent's cognitive achievement is a function of three factors: his cognitive *goal*; the *standard* required (or sufficient) for achieving that goal; and the cognitive *wherewithal* on which he can draw to meet that standard.

In discharging his cognitive agendas, it fits neither the relevant resource contingencies, the intellectual design or the imperatives of closure that a practical agent conduct his affairs on the model of axiomatic set theory, particle physics or welfare economics. An individual makes do with lesser ambitions because in general they are all that he need fulfill and all that he can afford. We see in this an essential equilibrium. The practical agent tends to set goals that he can attain and is stocked with the wherewithal that makes attainment possible (and frequent). In the matter of both goals set and the execution of standards for meeting them, the individual is a satisficer rather than an optimizer. There are exceptions, of course; a working mathematician won't have a solution of Fermat's Last Theorem unless he has a full-coverage proof that is sound (and, as it happens, extremely long).

Proposition 2.6 (Interaction of goals and resources) *The resources and goals of a practical agent exert a reciprocal influence. By and large, a practical agent's* *cognitive goals are comparatively modest. Accordingly, plausible beliefs defeasibly held are a practical agent's stock-in-trade. For most of what presses for his cognitive response, neither mathematical nor scientific certainty is either required or possible.*

The tendency to satisfice rather than maximize (or optimize) is not what is *distinctive* of practical agency. This is a point to emphasize. In most of what they set out to do and end up achieving, institutional agents exhibit this same favoritism. What matters — and sets them apart from the likes of us — is not that they routinely optimize but that they satisfice against loftier goals and tougher standards.

It is necessary to have a brief further word about the proposed concurrence between the distinction between practical and theoretical agents and that between individuals and institutions. As we conceive of the first of this pair, it is true by definition these that practical agents rank comparatively low and theoretical agents comparatively high in the hierarchy H of agent types. It is not a matter of definition that there is a concurrence of sorts between practical agents and individuals and between theoretical agents and institutions. These concurrences are matters of fact, and are so with a certain looseness. Some of the tasks open to an individual thinker, as well as the resources available for their completion, enable him to function as a theoretical agent in our technical sense. Again, someone who seeks a proof of Fermat's Last Theorem may find that he can proceed without press of time, short of his own mortality. Similarly, an institution that wishes to acquaint itself with the postal code of one of its clients can do so if a solitary employee takes the half-minute to consult the Post Office's directory. Our proposal concerning these loose confederacies amounts to little more than this. If we were to take the union of individual and institutional cognitive agents and applied to it the ordering of greater-resources-in-application-to-stricter goals, we would see, as a matter of fact, that

Proposition 2.7 (Individual and practical agents) *It is* typical *of the cognitive behaviour and cognitive circumstances of individual agents that they tend to rank comparatively low in H.*

Corollary 2.7(a) *It is* typical *of individuals to function as practical agents.*

Likewise,

Proposition 2.8 (Institutions and theoretical agents) *It is* typical *of the cognitive behaviour and cognitive circumstances of institutional agents that they tend to rank comparatively high in H.*

Corollary 2.8(a) *It is* typical *of institutions to function as theoretical agents.*

It is also useful to emphasize that the type of agency involved in these concurrences is *cognitive* agency. For the much broader notion of agency that carries the meaning of *ability to do*, the differences between what individuals are able to do and what institutions are able to do is not well-captured by where they rank in H. Cognitive agency is another matter, enough so that it lends requisite plausibility to the claims of approximate concurrence.

The approximation of the concurrence harbours another fact of interest (and intuitive plausibility). Given that the difference between practical and theoretical agency is a matter of numbers and amounts of resources and degrees of ambitiousness of cognitive goals, then any cognitive task solvable with resources of that type is such that if a practical agent can solve it, so too can can theoretical agent solve it. Anol ther way of saying this is that for large ranges of cases, whatever can be known by an individual agent can also be known by an institutional agent; not conversely, however.[4]

2.1.2 Peculiarities of Institutional Agents

It is true that practical agents often struggle with the problem of straitened cognitive resources. In giving this matter its due emphasis, we would not wish to leave the impression that institutional agents are invariably better off in this regard. Consider the case of the empirical sciences. Everything that is currently known of the history of science attests to the presence of an endemic discouragement, known as the *law of logarithmic returns*. The law provides that science acquires new information at a rate more or less proportional to what is spent to acquire it, but that the rise of new knowledge is proportional not with the quantity of new information, but rather with its logarithm. That is, putting K^I as the quantity of knowledge embedded in a body of information, $K^I = logI$. Accordingly,

Proposition 2.9 (The law of logarithmic return) *Informational flow-through of eponentially increasing magnitude is required for cognitive outputs at arithmetically increasing levels* [5]

In our brief treatment of them here, we have been concentrating on what theoretical or institutional agents have in common. But there are differences which matter greatly for their respective cognitive wherewithal. We have the space to make glancing reference to just one of these important differences. This is the difference between *markets* and *committees*. Neither of these is immune from cor-

[4] The qualification, " for large ranges of cases" is necessary. Exceptions include experiential knowledge. Only individuals can know what it is like to have arthritis or what it is like to be married to an arthritic. Conversely, both individuals and institutions alike may, absent the requisite experiences, have some knowledge of what arthritis is like or what being married to an arthritic is like.

[5] As Nicholas Rescher points out, proposition 2.9 is the cognitive analogue of the Weber-Fechner law of psychophysics, according to which stimulus imputs of geometrically increasing magnitude are required for arithmetically increased levels of *perception*. [Rescher, 1996, p. 79]

rosions that taint their cognitive behaviour. But in a rough and ready way, markets outperform committees.

When markets operate effectively, they are large, diverse groups whose collective decisions are outcomes of disagreement and contention, rather than consensus. Accordingly, markets possess mechanisms of aggregation that produce judgements that need not, and often are not, held by any given member of the group. In lots of cases, markets cognitively outperform any of their individual members. A further trait of markets is the very large degree of interpersonal independence of its individual members. There was a time when the received wisdom about groups could be summed up on Carlyle's sentiment, "I do not believe in the collective wisdom of individual ignorance" [6] or in Nietzsche's pungent observation, "Madness is the exception in individuals but the rule in groups" [Nietsche, 1966, p. 90]. But common experience and social scientific research belie the conventional wisdom as it relates to markets.

A case in point is the destruction in 1986 of the space shuttle *Challenger*. We here follow the exposition of [Maloney and Mulherin, forthcoming]. Eight minutes after the explosion the story was on the Dow Jones News Wire. Within minutes of that, investors started selling off the stocks of the four principal contractors involved in the Challenger launch. Rockwell International had manufactured the shuttle and its main engines. Martin Marietta had built the shuttle's external fuel tank. Lockheed had been in charge of ground support. Morton Thiokol had manufactured the shuttle's solid-fuel booster rocket. In less than a half an hour after the explosion, the stock of the first three of these companies was down by between three and 6%. The fourth company's stock, that of Morton Thiokol, was hit much harder, and buyers could not be found for the dumped shares. This very quickly precipitated a trading halt. When it was lifted, about an hour fter the explosion, Morton Thiokol's stock had fallen by 6% and by the end of the trading day was down by 12%. Meanwhile, the losses of the other three had improved and evened out at about 3%.

Within hours of the blast the stock market had determined that the responsibility for it rested with Morton Thiokol. In reaching this collective determination, thousands of individuals made buy-sell decisions independently of one another, and without benefit of any yet published speculation about where the responsibility lay. (According to an article the next morning in the *New York Times*, there were no clues as to the cause of the accident). Six months after, the Presidential Commission on the *Challenger* determined that the accident was caused by structural deficiencies in the shuttle's O-rings on the booster rocket made by Thiokol.

How did the market manage to be right in its determination of fault? Maloney and Mulherin determined that there was no dumping of stock by company officers on the day of the crash. Nor had any of Thiokol's competitors shorted the stock.

[6] Quoted from [Menschel, 2002, p. 136]

There was no evidence that anyone dumped Thiokol and purchased competitors' shares. In short Maloney and Mulherin report no evidence of insider trading.

According to [Surowiecki, 2004],

> the market was smart that day because it satisfied the four conditions that characterize wise crowds: *diversity of opinion* ..., *independence* ..., *decentralization* ... and *aggregation* (p. 10. Emphases added)

Accordingly,

Proposition 2.10 (Wisdom of crowds) *"If you ask a large enough group of diverse, independent people to make a prediction ..., and then average those estimates, the errors each of them makes ... will cancel themselves out. Each person's guess ... has two components: information and error. Subtract the error, and you're left with information." [Surowiecki, 2004, p. 10]*

It is striking that the market in this case outperformed all experts who got the answer wrong, and, in the sheer speed of its determination, also very significantly outpriced those experts who eventually got the right answer. Accordingly,

Proposition 2.11 (Markets and experts) *When they work best, markets outperform each individual member, including expert members.*

Markets are far from perfect collective cognizers. Sometimes they go badly awry. One difficulty for collective cognition is the so-called *information cascade*, which is a form of herding. [Bikhchandani *et al.*, 1992] and [Bikhchandani *et al.*, 1998] It is supremely ironic that the market got the later disaster of the *Columbia* in 2003 entirely wrong. After the *Columbia* suffered re-entry disintegration, the stock of Alliant Techsystems was hammered. Alliant Techsystems owns Thiokol, and Thiokol continues as builder of booster rockets for the shuttle program. Although we now have it that the cause of the accident was damage to a wing when struck by insulation foam, what appears to have occurred is that traders took the 1986 event as a kind of precedent, inferring that what was true of the *Challenger* incident in 1986 was also true of the *Columbia* incident of 2004. Since any holder of Alliant in 2004 could be expected to have some notion of Thiokol's fate in 1986, what we have here is a significant loss of market independence.

Committees are unlike markets in all the key respects. They tend to be small, rather than very large. They aim at consensus, rather that tolerating (indeed requiring) even high markets, committees also have feedback mechanism that actually obliterates the independence of markets. The one point on which there exists a structural similarity are mechanisms of aggregation. Even so, the aggregation of a committee's judgement is subject to difficulties that markets don't have.

A case in point is the jury in criminal trials where a unanimous vote is necessary for a finding of guilty. The principal difference between a jury and a market is that juries are highly personal in their operation and markets highly impersonal. On a jury each member has knowledge of how the others feel about the case. Their discussions are often fractious and exhausting. Given the press of externalities — e.g., that judges are extremely unencouraging about prospects of a hung jury, that there is a requirement of unanimity for conviction, and that juries are expected to render their verdicts in a timely way — there is ample occasion for herding, especially the so-called *bandwagon effect.* It might strike us as exceedingly imprudent that, in light of the cognitive corruptions to which they are prey, determinations of criminal guilt is left to juries. A partial answer to this challenge and the constraints imposed upon how juries reach their findings. One is that they must determine the facts of the case on the basis of evidence led and crossed at trial, and on nothing else. Another is that the evidence must sustain a finding of proved beyond a reasonable doubt. (We return to this point in chapter 8.)

In some ways, juries are special cases of committees. Jurors are carefully selected to exclude persons who may have expert knowledge of matters bearing on the case at hand. Furthermore, in reaching their findings, jurors are required to reasons in the manner of ordinary reasoners. Although expert opinion may be led at trial, it is clear that in its overall orientation, the criminal justice system favours untutored judgement over the expert's command. And, unlike the case of markets, the reason for this is not that juries of ordinary reasoners cognitively outperform criminologists and professional investigators, but rather it is added protection for the accused against wrongful conviction. In non-jury committees, individual expertise (when it can be recognized) is often given pride of place, and not infrequently is the trigger of a bandwagon effect. So we may say,

Proposition 2.12 (Committees and experts) *It is not in general the case that when a committee functions at its best, it outperforms the most expert of its members.*

2.1.3 Normativity

It is not infrequently supposed that it is intrinsic to logic to articulate standards of normative correctness, and that this separates logic from the domain of empirical enquiry. For all its substantial provenance, this, we think, is not a supportable view of logic.[7] Our *Actually Happens Rule* raises the question of the extent to which a RT-logic sets normative standards for rational cognitive practice. We have already said that this rule bears some affinity to a position in psychology called "panglossism". We should say something more about this. Contemporary cognitive science marks a distinction among three different models of cognitive

[7] For particularly blatant (though not untypical) expressions of this view, see[Walton, 2002, p. 474] and [Simon, 1977, p. 265].

performance. They are the normative model N which specify sets standards of optimal rational performance, irrespective of the costs — especially the computational costs — of compliance. The prescriptive model P attenuates these standards so as to make them computationally executable by beings like us. The descriptive model D gives a law-governed account of how beings like us actually perform. Following Stanovich [1999], it is possible to discern three different positions concerning how the three models, N, P, and D, are linked. The principle of linkage is nearness to the goal of *good reasoning*. On the panglossian approach, there is little or nothing to distinguish N, P, and D in relation to the good reasoning goal. At the opposite extreme is the "apologist" position in which N meets the goal and both P and D fail it, and do so both seriously and next-to-equally. The "meliorist" position takes up the middle position. N meets the goal. P fails it, but not so badly as to preclude *approximate* realization. D, on the other hand, fails it significantly.

It is not our intention to deal with the panglossian-meliorist-apologist rivalries at length. If we were forced to choose among the three, we would opt for the panglossian position. In fact, however, we find ourselves attracted to a fourth position, which the panglossian position somewhat resembles, but which is also importantly different. Baldly stated, we are disposed to *reject* the normative model, and to *reject* its prescriptive submodel. Thus our own position is vacuously panglossian: D reflects good reasoning rather well, and no other model reflects it better (since there are none). What so inclines us is the failure of those who espouse the N-P-D trichotomy to demonstrate that N provides objectively correct standards for optimal rationality and that P provides objectively correct standards for a computationally realizable optimal rationality. (Sometimes this is called "optimization under constraint". See [Stigler, 1961].) We will not debate this issue here (but see, e.g., [Woods, 2003, ch. 8]). But perhaps an example would help explain our reluctance. It is widely held that a system of standard logic is a normative model of good reasoning, because it contains provably sound rules for the valid derivation of conclusions from premisses (or databases). This presupposes that the hitting of validity-targets is invariably an instance of good reasoning. The truth is that in lots of situations valid deduction from premisses or a database is not the way in which good reasoning proceeds, well attested to by the example of *ampliative* reasoning.

But what if the reasoner's task were such as to require the use of deduction rules? Wouldn't a system of logic whose deduction rules were provably sound and complete be a convincing model of that particular sort of good reasoning? No. Good reasoning is always good in relation to a goal or an agenda (which may be tacit). Reasoning of the sort in question is good if it meets its goal or closes its agenda using only valid deduction rules. Reasoning validly is never *itself* a goal of good reasoning; otherwise one could always achieve it simply by repeating a premiss as conclusion, or by entering a new premiss that contradicts one already present.

Suppose, finally, that the would-be deductive reasoner had ready to hand a sound and complete set of deduction rules and a set of heuristic principles that, for any goal attainable through deductive reasoning, guided the reasoner in the selection of particular rules at each step of the deduction process. Wouldn't those deduction rules together with those heuristic rules serve as a normative model of the sort of reasoning our reasoner has set out to do? And, given that no such heuristics would work if they weren't actually used by real-life deducers, isn't there reason to think that we have here a case in which a normative and a descriptive model converge in a panglossian way?

Yes, but with a clarification and a caveat. The clarification is that we do not despair of the idea of normative cognitive performance. Rather, there is no reliable way to capture this normativity save by attending to what beings like us actually do. Our normativity is descriptively immanent, rather than transcendent. The caveat that we would make is that no such set of deduction rules supplemented by the requisite heuristic rules suffice for the rationality of the *goal* that our reasoner may wish to achieve on any given occasion. This is the spirit in which we proposed the *Actually Happens Rule.* The *Actually Happens Rule* is rooted in what, for us, is a primary datum. It is that the reasoning actually performed by individual agents is sufficiently reliable as not to kill them. It is reasoning that precludes neither security nor prosperity. This is a fact of fundamental importance. It helps establish the fallibilist position that it is not unreasonable to pursue modes of reasoning that are known to be imperfect. Suffice it to say that a logic of reasoning must preserve this aspect of reasonableness.

Before leaving this matter, it would be well to take note of two prominent arguments on behalf of the existence of normative models (and, by extension, of prescriptive models) of human cognitive performance. Let K be a kind of cognitive performance, and let $S = \{S_1, \ldots, S_n\}$ be the set of standards for K that are sanctioned in a normative model N. According to those who favour the normative model approach, there are two reasons for supposing that the standards S_i are normative for us ground-zero reasoners.

1. *The analyticity rationale.*
   The S_i are normative for us because they are analytically true descriptions of what it is for K to be rational.

2. *The reflective equilibrium rationale.*
   The S_i are normative for us because they are in reflective equilibrium with what is taken to be rational K-practice.

We reject both these rationales. The analyticity rationale stands or falls with the theoretical utility of the concept of analyticity (or truth solely by virtue of the meaning of constituent terms). There is a large literature, with which it is appropriate to associate the name of Quine, that counsels the rejection of analyticity.

(See, e.g., [Quine, 1953]; *cf.* [Woods, 1998].) These criticisms will be familiar to many readers of this book, and we will not repeat them here. (We are not so much attempting to prove a case against the normative models approach as to indicate to the reader why it is that we do not adopt it.) Still, here is an instructive example. Until 1902 it was widely held that the axioms of intuitive set theory were analytic of the concept of set. Then the Russell paradox put an end to any such notion.

The reflective equilibrium rationale can briefly be characterized as a balancing act between conservatism and innovation. Consider a proposed new K-principle. It should be rejected if it contradicts received opinion about K-rationality. Equally, consider a piece of K-behaviour. It should be rejected if it contradicts the established K-principles. But a new principle can be adopted if it attracts the requisite change in accepted K-behaviour.

The doctrine of reflective equilibrium pivots on the fact that the K-theorist, like everyone else, begins in *medias res*. He cannot proceed except on the basis of what he already believes about K-hood, and he cannot proceed unless what he initially believes about K-hood is also believed by others who are relevantly situated. The qualification "relevantly situated" is important. If K-theory is a relatively technical matter, then the relevantly situated others could be the community of researchers into K. Their beliefs, and our K-theorist's starting point, are a blend of individual K-judgements and sets of K-principles that are now in reflective equilibrium.

We accept this as a descriptively adequate representation of how a K-theorist actually proceeds. We concede that at an operational level there is no other way for the K-theorist *to* proceed. These ways of proceedings are an indispensable heuristic for the would-be K-theorist. But it is a mistake to think that, because this is the way that he must proceed it must follow that the reflective equilibrium from which he does proceed is epistemically or normatively privileged. It is a mistake of a sort that we call the *Heuristic Fallacy*.

Definition 2.13 (The heuristic fallacy) *The heuristic fallacy is the mistake of determining that any belief that is indispensable to the thinking up of a theory is a belief that must be formally derivable in the theory itself.*

We have it, then, that

Proposition 2.14 (Normativity) *In a practical logic of cognitive systems, normativity is implicit in standard practice. Even so, the reflective equilibrium model of normative rationality is untenable. The anyticity model is also unsound.*

Corollary 2.14(a) *Human reasoning is for the most part right enough about the right things. Right things are those things, if got wrong, carry high costs. Accordingly, a description of what reasoners are up to when they reason about such matters will reveal the norms of right reasoning such as they may be.*

It is necessary to say a word about *ideal* models. In the later chapters of this book, we develop formal models of abduction. A formal model is an idealized description of what an abductive agent does. As such, it reflects some degree of departure from empirical accuracy. Thus an ideal model I is distinct from a descriptive model D. But isn't this tantamount to a capitulation to normativity? It is not. Ideality is not normativity. Ideality is abstraction in quest of systemacity. The laws of frictionless surfaces are idealizations of the the slipperiness of real life. The laws that hold in the ideal model are approximated to by the pre-game ice at Maple Leaf Gardens. But no one thinks that there is something *defective* about the Gardens' ice conditions. We do not want to minimize the difficulty in getting the right ideal models of K-behaviour. All that we say now is that it is no condition on the selection of ideal K-models that they hit pre-set standards tricked out in a putatively normative model N. [8]

Ideal modellers have long recognized that reasoners in the rough, that is, reasoners operating in real time in the give-and-take of actual circumstance operate in ways that deviate from the modeller's putative norms. There is little inclination among such theorists to dismiss such performance-levels out of hand. But there is a near-universal disposition to regard them as subpar, as less than best. So conceived, real-life performance approximates to the performance called for by higher standards; and in this presumed suboptimality lies its subparness; it is, at best, approximate success. There is a name for such performance standards. They go under the collective designation of *heuristics*. We have nothing to say against the concept of heuristics, except this:

Proposition 2.15 (Heuristics) *It is simply a mistake to suppose that a heuristically successful performance is, just so, a subpar performance.*

2.1.4 Mathematical Models

Given its extensive absorption by mathematics, it is hardly surprising that modern mainstream logic has turned to mathematics for its working notion of model. Such in turn has also been the choice of the natural sciences, certainly of those of them for the expression of whose laws mathematics is indispensable. The more complex of the natural sciences have fared less well in capturing its essential insights in mathematical formalisms. This, too, is not surprising, given the comparative messiness and lack of generality of, say, the life sciences. Biology is an interesting test case for the would-be practical logician. There is a use of the word "theory" in which a scientific accounts theoretical component is that which falls beyond the ambit of observation. In many cases, a theory is little more than a mechanical device that computes or predicts output from a system's inputs. In biology, perhaps

[8] Our position on normative models is set out in greater detail in[Gabbay and Woods, 2003b]

the classic example is theoretical population and evolutionary genetics. Here all the basic processes are quite well known. These include the operations of inheritance, the facts of mutation, migration, and the mechanisms of natural selection under varying conditions of survival and fertility. Thus

> Theoretical evolutionary genetics assembles all these phenomena into a formal mathematical structure that predicts changes in the genetic composition of populations and species over time as a function of the numerical values of these elementary processes [Lewontin, 2003, p. 40].

Here the formalization works. It works because the underlying mechanisms are known. There are lots of cases in which this is not so. There the formal modelist has, apart from desistance, no option but to fly high. For, in its pure form, the mechanical formalities are posited without any direct connection to underlying material data. This makes the theorist's formal model an empirically unsupported place-holder for the actual dynamical details once they become known. An especially extreme, and failed, example of this theoretical high-flying was the Rashevsky school of mathematical biophysics, which operated in the late 1930s at the University of Chicago. Within three decades the movement was dead, made so by the extreme over-idealization of its physical models, so radical as to make them empirically inert. The Rashevsky collapse teaches an important lesson. It is that certain biological processes may not admit of accurate mathematical expression. There are still other cases in which postulated mathematical expressions of biological processes turn out to be right, but a good deal less than optimal even so. This we see in the case of Turing's conjecture that early embryonic development could be understood as the result of different concentrations of (observationally undetermined) molecules, distributed differentially within the embryo. This is right as far as it goes. But developmental genetics owes nothing to Turing's model. What it achieves in accuracy it pays for in over-simplification [Lewontin, 2003, p. 41].

These are lessons for the practical logician to take to heart. For one thing, the behaviour of human animals exceeds in complexity any grasp we have of the fruit fly, no matter how exhaustive. Apart from that, there is the charming problem of "down below", which is where much of a practical agent's cognitive agenda is transacted. The republic of down below is ringed by unwelcoming borders. Not only is much of what goes on there inaccessible to introspection, and experimental probes are heavily constrained by the ethical requirement to do no harm.

2.1.5 Slight-resource Adjustment Strategies

Slight-resource adjustment strategies lie at the crux of the economy of effort, as Rescher calls it [Rescher, 1996, p.10]. They instantiate a principle of least effort, and they bear on our tendency to minimize the expenditure of cognitive assets. [9]

2.1.6 Hasty Generalization

Individual cognitive agents are hasty generalizers. Hasty generalization is a *SRAS*, i.e., a slight-resource adjustment strategy. In standard approaches to fallacy theory and theories of statistical inference, hasty generalization is a blooper; it is a serious sampling error. This is the correct assessment if the agent's objective is to find a sample that is guaranteed to raise the conditional probability of the generalization, and to do so in ways that comport with the theorems of the applied mathematics of chance. Such is an admirable goal for agents who have the time and know-how to construct or find samples that underwrite such guarantees. But as Mill shrewdly observes, human individuals often lack the wherewithal for constructing these inferences. The business of sample-to-generalization induction often exceeds the resources of individuals and is better left to institutions. (See [Woods, 2004].) A related issue, even supposing that the requisitely high inductive standards are meetable in a given situation in which a practical agent finds himself, is whether it is necessary or desirable for him (or it) to meet that standard. Again, it depends on what the associated cognitive goal is. If, for example, an individual's goal is to have a reasonable belief about the leggedness of ocelots is, rather than to achieve the highest available degree of scientific certainty about it, it would suffice for him to visit the ocelot at the local zoo, and generalize hastily "Well, I see that ocelots are four-legged".

2.1.7 Generic Inference

Often part of what is involved in a human reasoner's facility with the one-off generalization is his tendency to eschew generalizations in the form of universally quantified conditional propositions. When he generalizes hastily the individual agent is often making a *generic* inference. In contrast to universally quantified conditional propositions, a generic claim is a claim about what is characteristically the case. "For all *x*, if *x* is a ocelot, then *x* is four-legged" is one thing; "Ocelots are four-legged" is quite another thing [Krifka *et al.*, 1995]. The first is felled by any true negative instance, and thus is *brittle*. The second can withstand multiples of true negative instances, and thus is *elastic*. There are significant economies in this. A true generic claim can have up to lots of true negative instances. So it is

[9] See here the classic work of George Zipf. [Zipf, 1949]

true that ocelots are four-legged, even though there are up to lots of ocelots that aren't four-legged. The economy of the set-up is evident: With generic claims, it is unnecessary to pay for every exception.

Generic claims are a more affordable form of generalization than the universally quantified conditional. This is part of what explains their dominance in the generalizations that individual agents tend actually to make (and to get right, or some near thing). It must not be thought, however, that what constitutes the rightness (or some near thing) of an individual's hasty generalizations is that when he generalizes thus he generalizes to a generic claim. Although part of the story, the greater part of the rightness of those hasty generalizations arises from the fact that, in making them, an individual typically has neither set himself, nor met, the standard of inductive strength. This, together with our earlier remarks about validity, is telling.

Proposition 2.16 (Validity and inductive strength) *Given the cognitive goals typically set by practical agents, validity and inductive strength are typically not appropriate (or possible) standards for their attainment.*

Corollary 2.16(a) *This, rather than computational costs, is the deep reason that practical agents do not in the main execute systems of deductive or inductive logic as classically conceived.*

2.1.8 Natural Kinds

Our adeptness with generic inference and hasty generalization is connected to our ability to recognize *natural kinds* [Krifka *et al.*, 1995, pp.63–95]. Natural kinds have been the object of much metaphysical skepticism of late [Quine, 1969b], but it is a distinction that appeals to various empirical theorists. The basic idea is evident in concepts such as *frame* [Minsky, 1975], *prototype* [Smith and Medin, 1981], *script* [Schank and Abelson, 1977] and *exemplar* [Rosch, 1978]. It is possible, of course, that such are not a matter of metaphysical unity but rather of perceptual and conceptual organization.

It goes without saying that even when the goal is comparatively modest — say, what might plausibly be believed about something at hand — not every hasty generalization that could be made comes anywhere close to hitting even that target. The (defeasible) rule of thumb is this: The hasty generalizations that succeed with these more modest goals are by and large those we actually draw in actual cognitive practice. We conjecture, in the spirit of the *Actually Happens* principle, that the comparative success of such generalizations is that they generalize to generic propositions, in which the process is facilitated by the agent's adeptness in recognizing natural kinds.

2.1.9 Defaults

Generic inference tolerates exceptions, but it is not *ex cathedra*. The cognitive economy of individual agency is a highly fallibilist one. It is an economy characterized by *defaults*. A default is something taken as true in the absence of indications to the contrary [Reiter, 1980]. It is characterized by a process of reasoning known as "negation-as-failure" [Geffner, 1992]. For example, Harry checks the departure times for a direct flight from Vancouver to London early Saturday afternoon. Finding none posted, he concludes that there is no such flight at that time on Saturday. A great many defaults arise by instantiation from generic claims. It is generically true that crows fly. It is a default that Jasper the crow flies. Default inference does not preserve the comparative immunity from counterexample possessed by generic propositions. If Jasper doesn't in fact fly, that falsifies the default that says he does, but does not falsify the generic claim that says that crows fly. (See below, chapter seven).

Default reasoning is inherently conservative and inherently defeasible, which is the cognitive price one pays for conservatism. Conservatism is, among other things, a method for collecting defaults D. One of the principles, of collection, as we saw, is "Derive D by instantiation of generic truth". Another, both broader and overlapping, is "Derive D from common Knowledge". The economies achieved are avoidance of the costs of fresh-thinking. (Descartes' epistemological project would be costly beyond price for an individual to execute.)

2.1.10 Discourse Economies

Further economies are evident in regularities affecting conversation. One such has been called

> *The Reason Rule*: One party's expressed beliefs and wants are a *prima facie* reason for another party to come to have those beliefs and wants and, thereby, for those beliefs and wants to structure the range of appropriate utterances that party can contribute to the conversation. If a speaker expresses belief X, and the hearer neither believes nor disbelieves X, then the speaker's expressed belief in X is reason for the hearer to believe X and to make his or her contributions conform to that belief ([Jacobs and Jackson, 1983, p. 57] and [Jackson, 1996, p. 103]).[10]

A corollary to the reason rule is the *ad ignorantiam* rule:

[10] The reason rule reports a *de facto* regularity between real-life discussants. When the rule states that a person's acceptance of a proposition is reason for a second party to accept it, "reason" means "is *taken* as reason" by the second party.

> *Ad Ignorantiam Rule*: Human agents tend to accept without challenge the utterances and arguments of others except where they know or think they know or suspect that something is amiss [Gabbay and Woods, 2002].

Here, too, factors that trigger the *ad ignorantiam* rule are dominantly economic. Individuals lack the time to fashion challenges whenever someone asserts something or advances a conclusion without reasons that are transparent to the addressee. Even when reasons are advanced, social psychologists report that addressees tend not to appraise them before accepting the conclusions they purport to underwrite. Addressees tend to do one or other of two different things before submitting such reasons to critical scrutiny. They tend to accept a person's conclusion if they find it *plausible*. They also tend to accept the other party's conclusion if it seems to them that this is a conclusion which is within that person's competence to make; that is, if he is judged to be in a position to know what he is talking about, or if he is taken as having the appropriate *expertise* or *authority*. (See, e.g., [Petty and Cacioppo, 1986; Eagly and Chaiken, 1993][Petty *et al.*, 1981; Axsom *et al.*, 1987; O'Keefe, 1990], and the classic paper on the so-called *atmosphere effect*, [Woodworth and Sells, 1935]. But see also [Jacobs *et al.*, 1985].)

2.1.11 Consciousness

A further important respect in which individual agency stands apart from institutional agency is that human agents are conscious. (The consciousness of institutions, such as may be figuratively speaking, supervenes on the consciousness of the individual agents who constitute them.) Consciousness is both a resource and a limitation. Consciousness has a narrow bandwidth. This makes most of the information that is active in a human system at a time consciously unprocessible at that time. In what the mediaevals called the *sensorium* (the collective of the five senses operating concurrently), there exist something in excess of 10 million bits of information per second; but fewer than 40 bits filter into consciousness at those times. Linguistic agency involves even greater informational entropy. Conversation has a bandwidth of about 16 bits per second.[11]

The narrow bandwidth of consciousness bears on the need for cognitive economy. It helps elucidate what the scarcity of information consists in. We see it explained that at any given time the human agent has only slight information by

[11] [Zimmermann, 1989]. Here is John Gray on the same point: "If we do not act in the way we think we do, the reason is partly to do with the bandwidth of consciousness — its ability to transmit information measured in terms of bits per second. This is much too narrow to be able to register the information we routinely receive and act on. As organisms active in the world, we process perhaps 14 million bits of information per second. The bandwidth of consciousness is around eighteen bits. This means that we have conscious access to about a millionth of the information we daily use to survive" [Gray, 2002, p. 66].

the fact that if it is consciously held information there is a bandwidth constraint which regulates its quantity. There are also devices that regulate consciously processible information as to *type*. A case in point is informational relevance. When H.P. Grice issued the injunction, "Be relevant", he left it undiscussed whether such an imperative could in fact be honoured or ignored by a conscious act of will. There is evidence that the answer to this question is "No"; that, in lots of cases, the mechanisms that steer us relevantly in the transaction of our cognitive tasks, especially those that enable us to discount or evade irrelevance, are automatic and prelinguistic [Gabbay and Woods, 2003a]. If there is marginal capacity in us to heed Grice's maxim by consciously sorting out relevant from irrelevant information, it is likely that these informational relevancies are less conducive to the closing of cognitive agendas than the relevancies that operate "down below". Thus vitally relevant information often can't be processed consciously, and much of what can is not especially vital.[12]

Consciousness can claim the distinction of being one of the toughest problems, and correspondingly, one of the most contentious issues in the cognitive sciences. Since the agency-approach to logic subsumes psychological factors, it is an issue to which the present authors fall heir, like it or not. Many researchers accept the idea that information carries negative entropy, that it tends to impose order on chaos.[13] If true, this makes consciousness a thermodynamically expensive state to be in, since consciousness is a radical suppressor of information. Against this are critics who abjure so latitudinarian a conception of information [Hamlyn, 1990] and who remind us that talk about entropy is most assured scientifically for closed systems (and that ordinary individual agents are hardly *that*).

The grudge against promiscuous "informationalism" in which even physics goes digital [Wolfram, 1984] is that it fails to explain the distinction between energy-to-energy transductions and energy-to-information transformations [Tallis, 1999, p. 94]. Also targeted for criticism is the view that consciousness arises from or inheres in neural processes. If so, "[h]ow does the energy impinging on the nervous system become transformed into consciousness?" [Tallis, 1999, p. 94].

In the interests of economy, we decline to join the metaphysical fray over consciousness. The remarks we have made about consciousness are intended not as advancing the metaphysical project but rather as helping characterize the economic limitations under which individual cognitive agents are required to perform.

[12] Consider here taxonomies of vision in which implicit perception has a well-established place [Rensink, 2004].

[13] Thus Colin Cherry: "In a descriptive sense, entropy is often referred to as a 'measure of disorder' and the Second Law of Thermodynamics as stating that 'systems can only proceed to a state of increased disorder; as time passes, entropy can never decrease.' The properties of a gas can change only in such a way that our knowledge of the positions and energies of the particles lessens; randomness always increases. In a similar descriptive way, information is contrasted, as bringing order out of chaos. Information, then is said to be 'like' negative energy"[Cherry, 1966, p. 215].

Consciousness is tied to a family of cognitively significant issues. This is reflected in the less than perfect concurrence among the following pairs of contrasts.

1. conscious v unconscious processing
2. controlled v automatic processing
3. attentive v inattentive processing
4. voluntary v involuntary processing
5. linguistic v nonlinguistic processing
6. semantic v nonsemantic processing
7. surface v depth processing

What is striking about this septet of contrasts is not that they admit of large intersections on each side, but rather that their concurrence is approximate at best. For one thing, "tasks are never wholly automatic or attentive, and are always accomplished by mixtures of automatic and attentive processes" [Shiffrin, 1997, p. 50]. For another, "depth of processing does not provide a promising vehicle for distinguishing consciousness from unconsciousness (just as depth of processing should not be used as a criterial attribute for distinguishing automatic processes ..." [Shiffrin, 1997, p. 58]). Indeed "[s]ometimes parallel processing produces an advantage for automatic processing, but not always .... Thoughts high in consciousness often seem serial, probably because they are associated with language, but at other times consciousness seems parallel ..." [Shiffrin, 1997, p. 62].

It is characteristic of agents of all types to adjust their cognitive targets upwards as the cognitive resources for attaining them are acquired. A practical agent may take on commitments previously reserved for agents of higher rank if, for example, he is given the time afforded by a tenured position in a university, the information stored in the university's library and in his own PC, and the fire-power of his university's mainframe. In like fashion, institutional agents constantly seek to expand their cognitive resources (while driving down the costs of their acquisition, storage and deployment), so that even more demanding targets might realistically be set. Accordingly,

Proposition 2.17 (Asset enhancement) *Agents tend toward the enhancement of cognitive assets when this makes possible the realization of cognitive goals previously unattainable (or unaffordable).*

Corollary 2.17(a) *Asset enhancement is always tied to rising levels of cognitive ambition. In relation to cognitive tasks adequately performed with present resources, an interest in asset enhancement is obsessive beyond the range of what would count as natural and proportionate improvements upon what is already adequately dealt with.*

2.2 Practical Logic

Is there really such a thing as a practical logic? Is a practical logic even possible? One standard philosophical view is that these questions should be answered negatively, since practical inference is about actions, whereas a would-be logic of practical inference is actually a theory of belief modification, and hence is theoretical.[14]

Joseph Raz has an interesting answer to this objection. He argues as follows:

1. Practical reasoning is reasoning about what actions to perform.
2. A logic of reasoning of any kind is, as such, a theory of theoretical inference.
3. So, a practical logic is a logic of theoretical reasoning *when performed in ordinary ways*, i.e., by beings like us in everyday circumstance [Raz, 1978, p. 8].

As it stands, Raz's argument is a non-sequitur. Its repair is possible by addition of the premiss, "Practical logics are possible". But this freedom from non-sequitur is bought at the cost of begging the question against the very critic for whose benefit Raz constructed his argument in the first place.

It is possible that Raz has misstated what he intended his answer to be. Perhaps what he had in mind is this:

1. Suppose we agree that any logic worthy of the name is, or subsumes, a theory of belief-modification
2. Suppose also that we agree that theories of belief-modification are theories of theoretical reasoning
3. Let it be a point of additional agreement that practical reasoning is always reasoning about what to do
4. If a logic of practical reasoning is possible, it is necessary and sufficient that in reasoning about what to do, reasoners modify (delete, add, intensify, etc.) their beliefs about what to do
5. Since it is obviously possible for people, in reasoning about what to, to modify their beliefs about what to do, a practical logic is possible. It is a logic of belief-modification (hence a logic of theoretical reasoning) concerning beliefs about what to do (hence a logic of practical reasoning).

[14] We note in passing, the oddity of supposing that belief-change is intrinsically a theoretical enterprise. But since we ourselves use the term "theoretical" in a somewhat nonstandard way, we can hardly complain of this other usage on grounds of nonstandardness, different as it is in other respects from our own.

6. What is more, it is not necessary for a conclusion of practical reasoning to be an updated belief about what to do, or that the premisses always be beliefs. If it is possible to reason directly from a desire rather from a the belief that the desire exists, then the present claim is well-justified.

We have drawn the reader's attention to Raz's interesting, though bungled, answer to a common objection to practical logic, not because we think that Raz's mistake is all that important. We have already said why we think the identification of practical reasoning with reasoning about what to do seems to us less than well-advised, for it leaves the other side of the implied contrast strikingly bereft of members. And we have explained why we think that we get a robust and principled distinction between the practical and the theoretical by relativizing its relata to different degrees of command of the requisite cognitive resources in pursuit of targets of differing conditions of strictness. Even so — and apart from our reservations about the case he makes for practical logic — we are rather taken with Raz's observation that "practical reasoning is but ordinary theoretical reasoning" [Raz, 1978, p. 8]. As we saw, Raz probably means by this that a logic of belief-modification is capable of dealing with *beliefs* about what to do, hence can be at once a theory of theoretical and practical reasoning. But Raz's words also fit our own conception of this distinction. Seen thus, theoretical reasoning is reasoning done with comparatively abundant resources aimed at comparatively ambitious targets, and practical reasoning is reasoning done with comparatively scant resources aimed at comparatively modest targets. Drawn in this way, it is unnecessary (and undesirable) to see the difference between theoretical and practical reasoning as ontologically stark. It is not that there is a sharp and deep difference in kind between the two, but rather a difference in cognitive reach and enabling wherewithal. What is more, if we were to take Raz's unexplained reference to "ordinary" reasoning as reasoning done by practical agents, i.e., agents with comparatively scant resources, then the words "practical reasoning is but ordinary theoretical reasoning" say something true about our conception of these things, in which practical reasoners use the same resources as theoretical reasoners, but fewer of them and in lesser quantities. Accordingly, we find it justified to persist with the view that

Proposition 2.18 (Practical logic revised) *A practical logic is a principled description of the belief and decision dynamics of a practical agent, that is, of an agent ranking comparatively low in the hierarchy H of agency-types.*

Whatever the details of an ideal models approach to logic, it is necessary that we not lose sight of the fact that

Proposition 2.19 (Approximation) *If an ideal model of a certain kind K of human performance is to have elucidatory value, it is necessary that an appropriate* *approximation relation be definable in principle between actual behaviourial K-competence and the model's idealized behaviour.*

The heart and soul of any theoretical approach to practical reasoning is that it takes due note of resource-limitations and cognitive target-modesty. It would be illuminating if there were a coherent connection between the methodological factor of approximation and the logical factors of practicality. It may be that such a connection exists and that it takes the following form:

♡ **Proposition 2.20 (Approximation and practicality)** *There exist systems of so-called* approximate reasoning *which are themselves approximations of classical logic. This suggests (a) that the factor of practicality in practical reasoning might be modelled as approximation to classical reasoning, and (b) that as the approximation converges on classical limits, the factors of practicality recede from the model. (See, for example, [Finger and Wasserman, to appeara; Finger and Wasserman, to appearb; Schaerf and Cadoli, 1995].)*

2.3 Connectionist Logic

There is a large literature — if not a large consensus — on various aspects of non-symbolic, subconscious cognition. If there is anything odd about our approach, it can only be the proposal to include such matters in the ambit of logic. Most, if not all, of what people don't like about so liberal a conception of logic is already present in the standard objections to psychologism, which we have already discussed. Strictly speaking, there is room for the view that, while psychologism is not intrinsically hostile to logic, psychologism about the unconscious and the prelinguistic simply stretches logic further than it can go, and should therefore be resisted.

This is an admonition that we respect but do not intend to honour. In this we draw encouragement from work by Churchland and others [Churchland, 1989; Churchland, 1995] on subconscious abductive processes. As Churchland observes, "... one understands at a glance why one end of the kitchen is filled with smoke: the toast is burning!" [1989, p. 199]. Churchland proposes that in matters of perceptional understanding, we possess "... an organized 'library' of internal representations of various perceptual situations, situations to which prototypical *behaviors* are the computed output of the well-trained network" [1989, p. 207]. Like Peirce [1931–1958, p. 5.181], Churchland sees perception as a limit of explanation, and he suggests that all types of explanation can be modelled as prototype activation by way of "... vector coding and vector-to-vector transformation" rather than linguistic representation and standardly logical reasoning. On this approach the knowledge that comes from experience is modelled in the patterning of weights in the subject's neural network, where it is seen as a disposition of the system to assume various activation configurations in the face of various inputs. Thus, as Robert Burton puts it, Churchland is drawn to the view that "inference to the best explanation is simply activation of the most appropriate available prototype vector" [Burton, 1999, p. 261].

♡ **Proposition 2.21 (Connectionist logic)** *Abductive logic has, in part, the structure of a connectionist logic.*

The suggestion that abduction involves a connectionist logic is attractive in two particular ways. One is that, unlike every other logic of explanation, connectionist explanation has a stab at being psychologically real. The other, relatedly, is that a connectionist logic is no enemy of the subconscious and prelinguistic sectors of cognitive practice. It is no panacea, either. There is nothing in the connectionist's prototype-library that solves the problem of the deployment of wholly new hypotheses, as, for example, in the case of Planck's postulation of quanta. On the other hand, the same is true of computer systems such as PI [Thagard, 1988], which mimic simple, existential, rule-forming and analogical genres of abduction. (See here [Burton, 1999, p. 264]. We discuss systems such as PI in chapter 5 below.) For, again, beyond that, we should not want to say that serial processing *requires* consciousness:

> Thoughts high in consciousness often seem serial, probably because they are associated with language, but at other times consciousness seems parallel, as when we attend to the visual scene before us. So the distinction between parallel and serial processing does not seem to map well onto the distinction between the conscious and the unconscious [Shiffrin, 1997, p. 62].

We shall resume the discussion of connectionist logics in chapters 5 and 8. Other candidates for the logic of down below are briefly considered in [Gabbay and Woods, 2003a, pp. 62–68].

2.3.1 Fallacies

Before leaving the issue of an individual's cognitive economics we touch briefly on some objections that might be brought against it. On the account sketched here, the individual is an inveterate fallacy-monger, whether by way of hasty generalization, *ad verecundiam* or *ad ignorantiam*, among others. In fuller accounts of the cognitive economy of individuals, the appearance of inveterate fallaciousness is even more widely evident [Woods *et al.*, 2002; Gabbay and Woods, 2003a]. It is not impossible that the human agent runs amok with fallacy, but we ourselves are disinclined to say so. The charge may be rebutted in one of two ways.

> (1) The practice would be a fallacy if interpreted in a certain way. But under more realistic construal, it doesn't fit with the fallacy in question.
>
> (2) The practice in question even under realistic interpretation qualifies as a fallacy by the lights of a certain standard, but does not qualify as a fallacy under a lesser standard, and it is the lesser standard that has the more justified application in the context in question.

If we go back to the example of hasty generalization, if the generalization is inference held to the standard of inductive strength, then it is a standard that in our haste is lost. But if the generalizer's cognitive goal is such as to make the standard of inductive strength more than its attainment requires, the generalization can hardly be faulted for failing a standard it omitted to set for itself, for failing to hit what it did not aim at.

The individual agent also economizes by unreflective acceptance of anything an interlocuter says or argues for, short of particular reasons to do otherwise. This outrages the usual ban on the *ad verecundiam*, according to which the reasoner accepts his source's assurances because he is justified in thinking that the source has good reasons for them. (The fallacy, then, would be the failure to note that the source is not suitably situated to have good reasons for his assurances.) Empirical findings indicate that this is not the standard which real-life individuals aim at. They conform their responses to a weaker constraint: If you have reason to believe that your source lacks good reasons for his assurances, then do not accept his assurances. The default position of *ad verecundiam* reasoners is that what people tell one another is such that incorporating it into one's own database or acting on it then and there is not in the general case going to badly damage one's cognitive agendas, to say nothing of wrecking one's life. We see in this a (virtual) strategy of cooperative acceptance, tentative though it is and must be, rather than a strategy for error-avoidance or error-minimization. Judged by the requisite standard, such expectations are in general neither misplaced nor fallacious. A fallacy is always a fallacy in relation to a contextually appropriate standard.

Ad ignorantiam is our final example. In its most basic form it is an inference in the form

1. It is not known that P
2. So, not-P.[15]

In that form there is not much to be said for it. But no one argues just by way of argument forms. In requisitely incarnate arrangements we sometimes get rather

[15] We are discussing the modern form of the *ad ignorantiam*, not Locke's conception, which in turn is a variant of Aristotle's *ignoratio elenchi* [Woods, 1999; Woods, 2004].

good arguments, such as negation-as-failure arguments. In their turn, negation-as-failure arguments are variations of *autoepistemic* arguments, such as:

1. If there were a Department meeting today, Harry would know about it.
2. But he doesn't,
3. So there isn't.

Or, as in the departure-announcement example,

1. If there were a direct flight from Vancouver to London early Saturday afternoon, the schedule would make that known to Harry.
2. But it doesn't.
3. So there isn't.

Autoepistemic inferences are inferences to a default. Harry's default position is that there is no such meeting and is no such flight. Such inferences are non-monotonic. New information might override these defaults. Here, too, there are fallacious cases of the *ad ignorantiam* depending on what the relevant standard is. Nobody thinks that the *ad ignorantiam* is truth-preserving.[16] For agents who are constitutionally and circumstantially bound to transact their cognitive agendas on the cheap (fast and frugal), who will say that the standards of default reasoning are inappropriate?

Let us say in passing that the variabilities that inhere in the hierarchy of agency-types suggest a general policy for the fallacy-attribution. It is roughly this. A fallacy is a mistake made by an agent. It is a mistake that seem not to be a mistake. Correspondingly, it is a mistake that is naturally made, commonly made, and not easy to repair (i.e., to avoid repeating) [Woods, 2004, ch. 1]. An inference or a move in a dialogue, or whenever else, is a fallacy relative to the type of agent in question and the resources available to agents of that type, and to the performance standards appropriate thereto. Given that individuals operate with scant resources, given the economic imperatives that these paucities impose and given the comparative modesty of their cognitive goals, what may have the look of fallacious practice lacks the cognitive targets and the performance standards against which fairly to judge inferences or moves in fulfillment of such targets as fallacious. On the other hand, for agencies of a type that occurs higher up in H — NASA, for example —

[16] An exception:

1. If I had a throbbing headache I would know it.
2. But I don't,
3. So I haven't.

cognitive targets are different (and more expensive) resources are abundant, and standards for the assessment of performance are correspondingly higher. Relative to those targets and those standards, cognitive practices having this appearance of fallaciousness are much more likely to *be* fallacious.

This helps motivate the traditional idea of a mistake that seems not to be a mistake. At the appropriate level, a cognitive practice *is* a mistake and may not appear to be a mistake, because at lower levels of the hierarchy it is *not* a mistake. Similarly, at least at the level of individual agency, we have an unforced explanation of why practices, which higher up would be fallacies, are lower down natural common and hard to change. It is because they are evolutionarily and experientially the best ways for individuals to manage their resource-strapped cognitive economies.[17]

Proposition 2.22 (Fallacies) *As standardly conceived of, fallacies are in the main wrongly attributed to practical agents. Either they are not patterns of reasoning that practical agents implement, or, when they are, they are directed to goals whose comparative modesty calls for standards that the instantiated cognitive behaviour in question manages to meet. In some cases, the so-called fallacies are successful* SRAS — *i.e., they are scant-resource adjustment strategies.*

This ends our foray into a practical logic of cognitive systems. Brief as the discussion has been, it may be hoped that the reader now has an orientation which will help to motivate the theory of abduction that we shall now begin to develop.

[17] This resource-based approach to fallacies can only be lightly sketched here. The fuller story can be found in our *Seductions and Shortcuts:Fallacies in the Cognitive Economy*, forthcoming in 2006.

Part II

Conceptual Models of Abduction

This Page is Intentionally Left Blank

Chapter 3

The Structure of Abduction

> The action of thought is excited by the irritation of doubt and ceases when belief is attained; so that the production of belief is the sole function of thought.
>
> Charles Peirce

3.1 Introductory Remark on Abduction

In our way of proceeding, conceptual accounts are inputs to formalization. Outputs are formal models of this conceptual content, in which the goals of greater precision and systematicity are realized. Conceptual models are sometimes called intuitive theories. A conceptual model of the behaviour of a practical agent is sometimes called pragmatic. So we begin with an intuitive pragmatics for abduction.[1]

The term 'abduction' was introduced into logical theory by Charles Peirce in the late 19th century. The introduction was not wholly original, since 'abduction' is a passable translation of Aristotle's *apagogē*, which is also translated as 'reduction' and was given the Latin rendering *abductio* by Julius Pacius. For Aristotle, an abduction is a syllogism, from a major premiss which is certain and a minor premiss which is merely probable, to a merely probable conclusion (*Prior Analytics* 2.25 69^a20–36). An important modern development cited the importance of reasoning from causes to effects. An insightful discussion can be found in Laplace's *Mémoires* [Laplace, 1904].

[1] The idea of cognitive economics is also the subject of important research in such disciplines as political science and marketing. See, for example, [Simon, 1982; Lilien *et al.*, 1992; Stigler, 1961; Shugan, 1980]. Still, there are important differences between Simon's approach to scarce-resource (or bounded) rationality and that of, say, Stigler.

In his early attempts to characterize abduction, Peirce also takes a syllogistic approach. Later on, he saw abduction as a form of reasoning in which a new hypothesis is provisionally accepted on the grounds that it explains the available data.[2]

It is necessary to note at the outset some significant ambiguities in the concept of abduction. In its most general sense, abduction is a process of justifying an assumption, hypothesis or conjecture for its role in producing something in which the abducer has declared an interest. Even within this category various distinctions press for recognition. For example, the hypothesis might help *explain* a given set of data or some phenomenon; or, it might facilitate the generation of observationally valid *predictions*; or it might permit the *elimination of other hypotheses*, thus providing the theorist with a *simpler* and *more compact* account or it might permit the *unification of disparate laws*. Here, then, are four distinct reasons which an abducer might offer as a justification for using a given hypothesis or for making a given conjecture. Abductions of this sort have an unmistakably pragmatic character. They are justifications of use without being evidence of the truth of the hypotheses in question.[3]

3.2 The Elementary Structure of Abductive Logic

Abduction offers two faces for the investigator's scrutiny. One is abduction the process, the other is abduction the product. In a rough and ready way, abductive products are investigated by way of properties possessed by the requisite linguistic structures or of linguistic structures in relation to abstract set theoretic structures. Abductive processes are investigated by way of conditions on the success or failure of the abductive behaviour of cognitive agents in actual practice. Both product and process are important foci of the investigator's probes; but in the approach taken in this book, considerations of process are given dominant place.

As we saw things in *Agenda Relevance*, the fundamental conceptual fact about relevance is that *information is relevant when it is helpful*. As we see things here, the fundamental conceptual fact about abduction is that abduction is *ignorance-preserving reasoning*. Nearly everyone agrees that all non-demonstrative reasoning occurs under conditions of uncertainty. But uncertainty, as standardly conceived, means *demonstrative* uncertainty. Abduction stands apart. The abducer's uncertainty extends well beyond the failure to be convinced by demonstrative means. It both arises in (indeed is prompted by) the abducer's ignorance more comprehensively conceived of; and even when it finishes successfully, the abduction leaves the reasoner in ignorance.

[2] A good short overview of Peirce on abduction is [Kraus, 2003].

[3] Newton, for example, accepted the action-at-a-distance theorem, but he was firm in thinking it unbelievable. (See below, chapter 4.)

In its barest form, abduction is a reaction of a certain kind to a *cognitive irritant*. As Rescher nicely observes, "The discomfort of unknowing is a natural component of human sensibility". [Rescher, 1996, p. 5]. The irritation is occasioned by the inability to hit some cognitive target with present epistemic resources. The cognitive target is in its turn constituted by some or other state of affairs. Putting the occasioning state of affairs as S, the set of our present cognitive resources (or knowledge-base) as K, the cognitive target occasioned by S as T, and then (as a first, and less than adequate, pass) the basic form of an *abductive trigger*[4] is

1. S obtains
2. S occasions T
3. K does not attain T

It is important to repeat:

Proposition 3.1 (The variability of abduction) *The parameters 'S', 'T' and 'attain' admit of variable instantiation.*

In one set of circumstances, S may be a newly discovered fact that cannot be explained by what is currently known. In that case, the abductive trigger is that the cognitive target T occasioned by S (the desired explanation) is not hit by what we currently know. In other cases, the unmet target associated with an abductive trigger can be entirely non-explanationist in character. If this is right, the theories such as those of [Thagard, 1989; Aliseda-LLera, 1997; Magnani, 2001a] and [Aliseda, forthcoming] which are explanationist accounts, canot qualify as general theories of abduction.[5] We briefly sketch a non-explanationist example. Let the state of affairs in question be one in which a set of proof rules implies a result which is thought to be unacceptable. Suppose further that the proof in question is not taken as a *reductio*. So the fact that its conclusion is unacceptable establishes (for those who think it so) that there is something wrong with the proof. If we assume that the proof misapplies none of its proof rules, then those who find the proof defective in this way must reject one or other of the rules used. This, is the situation of a "proof" "proving" the wrong thing. The target is finding the defective rule. But since the current rules encode what is currently known about these proof-structures, that target is not hittable from K.

[4] We borrow this attractive metaphor from Aliseda-LLera [1997].

[5] Thus Magnani: abduction is "inference to an explanatory hypothesis" [Magnani, 2001b, p. xi] and Aliseda: abduction is "reasoning from an observation to its possible explanations" [Aliseda, forthcoming, p. 8]; and [Meheus *et al.*, forthcoming]: ... logics for abductive reasoning enable one "to generate explanations for novel facts ... as well as for anomalous facts ...[Meheus *et al.*, forthcoming, p. 2].

3.3 Expanding the Schema

Our purpose is to expose something of the elementary logical structure of abductive reasoning, and to do so in a way that helps orient theorists to the various tasks that a logic of abduction should concern itself with. We are mindful of criticisms that have been leveled against the very idea of a logic of abduction [Reichenbach, 1938; Kapitan, 1992]; so we think it prudent to proceed with a certain diffidence. That our own account of abduction is itself abductive is methodological expression of this diffidence.[6]

We introduce the idea of an ignorance problem (IP)

Definition 3.2 (Ignorance problems) *An IP exists for a cognitive agent X iff X has a cognitive target T that cannot be attained from what he currently knows (or equivalently from K, his current knowledge base).*

IPs present cognitive agents with two options. One is to acquire new information that X will enable T to be attained. Accordingly, for an agent X,

> *IP-option (1) (X overcomes his ignorance)* X extends K to some successor knowledge-base K^* such that K^* attains T.

Another option is to acknowledge that the pair $\{K, T\}$ constitutes for X an insolubilium.

> *IP-option (2) (X's ignorance overcomes him)* Unable to succeed with option (1), X capitulates.

It is well to note the dynamic character character of this pair of options. For example, at time t_1, X might try and fail to exercise option (1). At t_2 he might acquiesce to option (2). Yet at t_3 he might recur to option (1) with good results.

It is commonly held that, when an agent is confronted with an ignorance-problem, (1) and (2) exhaust his option space. In fact, there is a third option. *It is the founding datum of abduction.*

> *IP-option (3) (Presumptive attainment)* X finds an H which, if he knew it, would together with K solve his IP; and from that fact he conjectures that H.

Option (3) incorporates the element of conjecture in an essential way. This is obvious in the case of H itself, but what is often overlooked is that this does not

[6] Anyone interested in whether this lands us in a "Hume" problem for for abduction might consult [Woods, 2004]. Also, the interconnections between abductive and inductive logic are well-explored by [Flach and Kakas, 2000].

solve the original problem. X's problem is that his T is attainable only on the basis of what he now knows (K) or can readily get to know (K^*). His situation *now* is that T cannot be attained either way. If he selects an H such that the truth of K revised by H *would* hit T, then *conjecturing* H does not produce K^*. In particular, K together with H (hereafter $K(H)$) is not a knowledge-set for X. (It does not solve X's ignorance problem).

This highlights the second irreducible element of conjecture that option (3) embeds. $K(H)$ doesn't hit T, but we may say that it hits it *presumptively*. Accordingly, option (3) offers X not a solution of his ignorance-problem, but rather attainment *faute de mieux* of a lesser target. Instead of a target that admits of only *epistemic* attainment, it proposes a conjectural variant of it that provides *presumptive* attainment. This is deeply consequential.

Proposition 3.3 (Ignorance-preservation) *Whereas deduction is truth-preserving and induction is probability-enhancing, abduction is ignorance-preserving.*

Proposition 3.3 sets forth what we will call the *ignorance condition* on abduction.

Option (3), as we see, is not a solution of an IP; it is a *transformation* of an IP into a problem that conjecture can solve. It is a response to an IP that requires X to lower his sights with regard to T. It turns on X's disposition to satisfice rather than maximize.

Here, too, it is prudent to re-emphasize the dynamic character of IPs and the responses that they induce. A cognitive agent might try and fail with option (1), and then move to option (3). If it also failed him, option (2) might recommend itself. If option (3) succeeded, X might persist with it until, so to speak, he came to know better; in which case he might move to option (1). So we have it that, in the beginning, X might try to overcome his ignorance, and, failing that, might try to conjecture to a lesser target. If this fails, he might acknowledge that his ignorance has got the better of him. Yet even if he succeeded conjecturally, he might later chance upon the means to abandon conjecture for fact, and so solve, with new knowledge, the problem that started it all. Accordingly, we say that

Proposition 3.4 (IP-relativities) *IPs arise in relation to targets in play at a time and resources then available. Responses to IPs retain those targets and proceed in ways permitted by subsequent resources.*

Peirce and others have emphasized that it is a condition on the *scientific* admissibility of an abductive conjecture H that it be testable, at least in principle. By these lights, a solution to an abduction problem is also a step in a process that might eventually solve the originating ignorance problem. So, for the class of cases that Peirce has in mind,

Proposition 3.5 (Ignorance-mitigation) *Although a solution to an abduction problem preserves the ignorance that gave rise to it, it may also contribute to the solution of the originating problem by identifying candidates for the status of new knowledge.*

In some contexts, abductive conjectures are not scientifically testable. For example, various forms of philosophical skepticism attract inference-to-the-best-explanation abductions. It may be that the best explanation of our external world experiences is that there is an external world that produces them. But to require that the external world hypothesis be testable is to beg the question against the skeptic, which in turn, ruins the anti-skeptic's refutation. Accordingly,

Proposition 3.6 (Testability) *Testability is not intrinsic to the making of successful abductive hypotheses.*

3.4 Frames

The dynamism of the IP-problematic also bears on the structure of options (1) and (2). Each turns on the availability of K^*. K^* is some future state of X's knowledge at a given time t. t is the time at which X recognizes that he has an IP, and his knowledge at that time is K. K^* is what X knows later, not anytime later, but later relative to what we might call the *frame* of his IP. It is impossible to be perfectly precise about this, save by stipulation. But intuitively the idea is sound, and clear enough to be getting on with. Consider an example. Harry wants to know whether Sarah will come to the picnic. He doesn't know. He phones her apartment; no answer. He phones her best friend; she doesn't know. There is presently no K * for Harry that solves this problem. He has no idea, and so waits until tomorrow to see for himself. He goes to the picnic and finds that Sarah isn't there. Today Harry acquiesces in option (2); but tomorrow he is able to deploy option (1). Doing so solves his ignorance-problem. But suppose instead that Harry fell ill and wasn't able to attend the picnic. Suppose that he never acquired a shred of additional information about Sarah's whereabouts on that day. Now, sixty years later, Harry is on his death-bed. Sarah appears. "Oh, Harry", she says, "how I wanted to attend that picnic all those many years ago!". Harry now knows that Sarah hadn't been there. But he hasn't resolved his IP problem. His new knowledge is outside its frame. This suggests that

Proposition 3.7 (IP-duration) *Typically an IP has a tacit "sell-by" date, after which it expires.*

We now have the means to define *abduction problems* AB. With K and T set as before,

Definition 3.8 (Abduction problems) *X has an AP with respect to K,T iff he has an IP with respect to K,T in response to which he is disposed to exercise option (3).*

3.5 Generalizing IPs

APs are not natural kinds. An AP is an IP to which X responds in a particular way. X substitutes conjecture for knowledge. It is the received view that all abduction problems are transformations of ignorance problems. This is a mistake. It is easy to see that the structure of abduction problems is wholly preserved if we substitute for K any cognitive state in comparison with which presumption is epistemically junior (belief is the obvious example). Accordingly, given that an ignorance problem represents an epistemic shortfall, a variant of it would represent a doxastic shortfall, or in some cases a plausibility shortfall. In each case, the conjecture deployed by the abducer's solution would have to meet two strong conditions.

Proposition 3.9 (Cognitive juniority) *If H is a solution of an AP, H has a lesser cognitive status than the cognitive standard against which the original problem arose.*

Proposition 3.10 (Effective juniority) *If H is a solution of an AP, then although there is a cognitive disparity between it and the cognitive standard against which the AP arose, H's cognitive juniority must comport with the requirement that it produce a presumptive solution of AP.*

Proposition 3.9 generalizes on the ignorance-preserving character of abductive solutions to IPs. It provides that in its fully general form, abductive solutions are cognitive deficit-preserving. Proposition 3.10 offers the helpful admonition, that for all their cognitive limitations comparatively speaking, successful Hs must have the wherewithal to produce rationally adequate, though cognitively subpar, solutions of their APs. Proposition 3.9 gives us occasion to broaden the ignorance-condition. As now we see, in its more general form, the condition requires that abductive theories honour the *cognitive-deficit condition.* Henceforth we shall read the ignorance-condition in this more general way, in the absence of indications to the contrary.

3.6 Avoiding a Confusion

When a resoning agent conjectures an H that bears the resumtive attainment relation to his cognitive target T, he is operating at an *epistemic* disadvantage. If he cannot attain T on the basis K of what he now knows, he may conjecture a proposition H that he doesn't know but which, if it were true, would, in apposition to what he does know, attain T. Or, in a variation, if T cannot be attained on the basis of what a reasoner *strongly believes* or what he *takes to be highly probable*, his hypothesized H must be a proposition that he neither (that) strongly believes nor takes to be (that) highly probable. As we see, the epistemic juniority of H is relative to the epistemic standing of the K in relation to which the ignorance-problem arose initially. So it bears repeating that the agent's recourse to H is from a position of *relative* epistemic juniority, and that this aspect of juniority is expressly recognized in the fact that in selecting it, the agent is proceeding conjecturally.

Note, however, that the content of the agent's conjecture of H is that H is *true*. This is as it should be, given that th conjecture of H turns on the fact (or what the abducer takes to be a fact) that if H *were* true, then H in apposition to K would attain the cognitive target T. Philosophers often characterize truth as an *alethic* property of propositions (or theories). Given that 'alethic' derives from the Greek word for 'true', the appellation has a certain redundancy about it, but not one that occasions any real harm. In fact, it is a baptism that affords us an essentially important distinction for the logic of abduction. Accordingly,

Proposition 3.11 (Epistemic v alethic factors) *While it is essential that a successfully abduced H possess the requisite* epistemic *juniority, it is neither necessary nor desirable that it be* alethically *subpar.*[7]

Corollary 3.11(a) *If we put it that abducing a H is always a kind of guessing, it is easy to see that what the abducer hopes for is that his guess will* turn out to be true. *Abducers deliberately set their task as one of guessing, but they do* not *aspire to guess what is false.*

The same lesson applies to K-parameters of strong belief or propositions held as highly probable. In conjecturing H, one's epistemic hold on it must be of a lesser grade than that of strong belief or propositions held as highly probable. But nothing precludes the abduced hypothesis hitting the alethic standard of truth. On the contrary.

3.7 Locating Abduction on the Logical Map

From its inception, logic has served two masters, *enquiry* and *inference*. In a rough and ready way, enquiry deals with premises-searches, and inference deals with

[7] In classical approaches to truth, any proposition that is alethically subpar is false. In many-valued approaches, an alethically subpar proposition has a less truth-like value than the proposition to which it is subpar. In truth-approximation approaches, one proposition is alethically subpar to a second when the former is less approximately true than the latter. This proceeds not only from the abductive character of verdicts but also from the admissibility of testimony.

premiss-projections. Throughout the history of logic, inference has been dominantly represented as the drawing of subsets of consequences from sets of priors. Enquiry has had a less firm grip on the evolution of logic. Aristotle makes fragmentary provision for it in *Topics* and *On Sophistical Refutations*, but in various subsequent periods, enquiry (or what also could be called "discovery") was excluded from the province of logic. In the present day, discovery has not found a place in the metropolitan centre of logic, but it thrives in the prosperous suburbs of dialogue logic and interrogative logic; and perhaps fledgingly in the logic of abduction. From the very beginning logic has had a decidedly easier time of it with its consequentialist approach to inference. Should we expect the same of a logic of abduction?

These and other issues take on a measure of clarity when considered against the backdrop of a basic schema for abduction, to an expanded description of which we now turn.

3.8 Abductive Schematics

Although ignorance abduction is but a case of cognitive-deficit abduction, we will here confine ourselves to the former as an expository convenience.

Let $T!$ express that T is a (contextually indicated) agent's target. Let R be the attainment relation on T, R^{pres} the presumptive attainment relation on T, H a hypothesis, $K(H)$, a knowledge-base revised by $H, C(H)$ a conjecture that H, and H^c a discharge of H. Then the schema for abduction begins to fall out.

1. $T!$ [declaration of T]
2. $\neg(R(K, T))$ [fact]
3. $\neg(R(K^*, T))$ [fact]
4. $R^{pres}(K(H), T)$ [fact]
5. H meets further conditions $S_1, \ldots, S_n$ [fact]
6. Therefore, $C(H)$ [conclusion]
7. Therefore, H^c [conclusion]

Remarks. $C(H)$ is read "It is justified (or reasonable) to conjecture that H". H^c denotes the discharge of H. H is discharged when it is forwarded assertively and labelled in ways that reflect its conjectural origins. (Here the label is 'c' in superscript position).

3.8.1 Consequentialist Abduction

The decision to treat an IP abductively is, as we have said, a decision to satisfice, to make do with less than was originally hoped for. It is a process in which the cognitive target T wears the trousers; for it is what T *calls* for that determines whether K can supply it. And, in its turn, it is the shortfall between K and T that influences our determination of what, if it existed, would close the gap. Abduction, accordingly, is conjectural provision of this gap-closure, whatever it is, a provision rooted in the counterfactual happenstance that *were* it part of what the agent now knows, his IP would not have arisen. We have it, then, that an abducer's choice of H is constrained not only by the nature of T and K, but also in what the IP in question, given T and K requires the attainment relation R to be; and, finally, by what the gap between T and K requires the presumptive attainment relation R^{pres} to be. If, for example, T calls for an explanation of some event E, it is clear that R is required to produce an explanation of E from K and, failing that, that R^{pres} is required to deliver a presumptive explanation of E from $K(H)$. In such a case, it might well be the case that a H exists such that $K(H)$, if true, would simplify a given account of E or unify certain of the laws that enter into that account. But if the simplification and the unification didn't constitute an *explanation* of T, they would be of no avail to the would-be abducer. The abducer's target wears the trousers. Similarly, if the agent's target is to repair a mistake, M (as in the *ex falso* example) and it is not currently known what occasions the mistake (i.e., what part of the agent's K has to be jettisoned), the abducer's task is to find a part of K, which if indeed it were removed would correct M (or materially assist in its correction), then it matters not what other cognitive objectives the attenuated K would now hit; it would not attain the abducer's actual target unless it *corrected* M. Again, it is T that ultimately calls the shots for abduction. It lies in what T calls for as to what, case by case, R^{pres} must consist in.

The literature on abduction is a very substantial one. Even a cursory exploration of it discloses a deeply embedded thesis about conditions that bear on the R^{pres} relation. On this view, $K(H)$ bears R^{pres} to T only if there exists a proposition V and a consequence relation $\looparrowright$ such that $K(H) \looparrowright V$. This is what we call the *consequentialist thesis* about abduction. It is a thesis that makes it intrinsic to the realization of R^{pres} that its antecedent stand in a further consequence relation to a further proposition V. Given the centrality of its place in this scheme, we shall say that according to the consequentialist thesis, V is a *payoff proposition* for T. By these lights, $K(H)$ bears R^{pres} to T only if it also bears $\looparrowright$ to V and V is a payoff for T. Accordingly, abduction is seen as the inference of H when the conditions embodied in what we shall call the *AKM* schema are met, as follows:

The AKM schema unfolds as follows:

1. E
2. $K \not\looparrowright E$
3. $H \not\looparrowright E$
4. $K(H)$ is consistent

5. $K(H)$ is minimal
6. $K(H) \not\rightsquigarrow E$
7. Therefore, H.

Although the present structure is, hands-down, the dominant schematic representation of abduction in the current literature, it may be helpful to associate it with the names of some of its more visible proponents; hence the honorific "*AKM*", in contrast with our own (the GW-schema, if the egoism may be forgiven).[8]

The dominance of consequentialism is closely allied to a further dominance in the abduction literature. Beginning with Peirce, himself, a very considerable percentage of the work on abduction, especially by philosophers, has assumed an *explanationist* call for T. Since T wears the trousers for abduction, abductions responsive to such Ts must embody explanations. When this happens, we will speak of *explanationist abduction.* So, the two dominant factors that constitute this concurrence are (in order of importance),

1. *explanationism*

2. *consequentialism*

It is no stretch at all to appreciate explanationionism's affinity to consequentialism. The philosophical literature is replete with consequentialist interpretations of explanations, of accounts in which an explanation is constituted by an explican and explicandum bound by a consequence relation. This is important. If we let E be a given explanandum and E' an explanans of it, there are philosophical understandings of explanation in which it is true that

(i) E' explains E

Only if (in some versions if and only if)

(ii) E' implies E.

On any such view of explanation, explanation is consequentialist; and any view of abduction that is both explanationist and consequentialist about explanation will be consequentialist about abduction. It is all rather iffy of course. We have already seen that abduction is not intrinsically explanationist. Even if it were, it is not at all clear that, in the putative implication of (ii) by (i), it is (ii) that wears the trousers for (i). What this means is that whereas it may be true that whenever

[8] Thus, for 'A' we have [Aliseda-LLera, 1997], and [Aliseda, forthcoming]; for 'K' we have [Kowalski, 1979; Kuipers, 1999]. And [Kakas *et al.*, 1995]; and for 'M' there is [Magnani, 2001a] and [Meheus *et al.*, forthcoming]. It is a small but representative sample.

a statement in the form of (i) is true, a statement in the form of (ii) is also true. Since the reverse certainly does not obtain, it is puzzling that in the *AKM*-schema a statement in the form of (ii) is allowed to stand in for a statement in the form of (i). The puzzlement abates once it is made clear, first, that explanationist accounts often assume without notice (certainly without fanfare) a DN-interpretation of explanation. When this is so, it is clear that the explanation at hand must embed a consequence relation. It is a *vital* embedment, since DN-explanation just is a consequence relation together with various constraints on its relata. This should give us pause. There is reason to think (see chapter 4) that not all explanation conforms to the DN-model.[9] To the extent that this is so, it may be doubted that an embedded consequence relation is a *vital* requirement of all explanationist abduction. This is to say that, even where it holds that (i) implies (ii), there may be modes of explanation for which (ii) *understates* the explanationist factor in any correlative explanationist abduction. Accordingly,

Proposition 3.12 (Non-explanationist abduction) *Abduction is not intrinsically consequentialist.*

Proposition 3.13 (Explanationism and consequentialism) *Not all modes of abduction embody consequentialism in the deep way that DN-explanation does.*

We are now well positioned to see that the *AKM*-schema is purpose-built for explanationist abduction, especially those modes of it in which the consequence relation is a deep and vital condition, as with DN-explanation. In light of propositions 3.12 and 3.13, we have it now that the *AKM*-schema, while it captures what abductions sometimes is, does nothing to capture the intrinsic structure of abduction.

While a dominant influence, the *AKM* model might appear *not* to be the sole model even among those who clearly are drawn to it. A case in point is what Aliseda calls anomalous triggers [Aliseda-LLera, 1997, p. 28]. Let it be the case that for some K and some true proposition P, $K \not\looparrowright P$ but $K \looparrowright \neg P$. Intuitively, this is a situation in which what one knows (up to now) is contradicted by some new fact. In its explanatory version K fails to explain P but succeeds in explaining its negation.

It is easy to see that in the first instance, what an anomaly triggers is not an abduction problem but rather a consistency-restoration problem (or, in its explanatory variation, an explanatory coherence problem.) With all due recognition for what holism allows for *in principle*, this first task requires us to cancel $K \looparrowright \neg P$. This in turn requires the restorer to make some deletion from K so that $K \looparrowright \neg P$ no longer holds. Since in the case before us P is not itself in K, the option of deleting P does not present itself.

[9] See also Thagard's explanatory coherence approach discussed in Chapter 6.

Even so, this is not yet an abduction problem. To be an abduction problem, it would have to be the case that the would-be restorer has no knowledge of how to proceed. When this is so, the Abductive option can be considered. In such a case, the agent's target now becomes the presumptive restoration of consistency (or explanatory coherence) by the conjecturing of an H such that $K(H) \looparrowright P$ and $K(H) \not\looparrowright \neg P$. But contrary to the appearance initially presented by the case of anomaly triggers, there is nothing in the present abduction that makes it unmodellable by the AKM-structure when $\looparrowright$ is read as "explanatory consequence".

We now turn briefly to some odds and ends. There are occasions on which the *AKM*-model tells (an essential part of) the story about a certain class of abduction problems. When this is so, the *AKM*-schema provides guidance on what constitutes the R^{pres} relation as regards a given explanatory target. We may put this as follows. Given

1. $T!$ [declaration of the target of explaining E]

2. $R(K, T)$ [fact]

3. $R(K^*, T)$ [fact]

what is desired is that we find an H such that

4. $R^{pres}(K(H), T)$

On the present assumptions, (4) is satisfied just in case it holds that

4* $K(H) \looparrowright E$

where K is as before and $K(H)$ is E' in the remarks above. What we have been trying to establish is that whereas (4*) is one way of producing R^{pres}, it is not a necessary condition on doing so in general. More generally, sometimes T itself specifies, or intimates, its own attainment conditions. In the case at hand, the target is to have as explanation of E. This makes it easy to identify T's payoff proposition, viz., that self-same E; and it facilitates the search for the further requisites of attainment (and presumptive attainment), viz., whatever it takes for

(i) E' explains E

to hold true. It merits emphasis that not all Ts either *specify* or *have* payoff propositions. If, for example, the target is to justify a recondite principle of logic L, it may suffice to produce a derivation of some obvious proposition of arithmetic A in which that logical principle occurs non-redundantly as premisses. Following Russell (in section 5. 6 below) we might well take this as grounds on which to hypothesize that the recondite principle L is indeed justified. But it is well to note that nowhere in this scenario is there any question that the abduction requires (or permits) that L itself is a payoff proposition for T or that L be in the counterdomain of any consequence relation on display in the abduction. Let the proof that doesn't deliver the goods for A be schematized as

$$\begin{array}{c} P_1 \\ \vdots \\ P_n \\ \hline A \end{array}$$

Assume now that if L is added as premisses, the proof goes through. In other words, whereas $\{P_1, \ldots, P_n, L\}$ doesn't suffice for A, $\{P_1, \ldots, P_n, L\}$ does. For this to be so, there must be a consequence relation on $\langle\{P_1, \ldots, P_n, L\}, A\rangle$. But A is not the payoff for T. Rather $\langle\{P_1, \ldots, P_n, K\}, A\rangle$ is. And this is itself neither a proposition nor the consequent of any consequence relation of which the abduction must take note. We repeat: sometimes T has a payoff proposition; sometimes it specifies this proposition; and sometimes this proposition is required to be in the counterdomain of a consequence relation the abduction must take note of. When these facts obtain, it is essential that the abductive enterprise take them into account. When they do not obtain, there is nothing to take into account; and no schema should posit them unduly. Accordingly,

Proposition 3.14 (GW and AKM) *The* GW-*model cannot be thought of as a generalization of the* AKM-*model.*

As is already apparent, a schema for abduction is open to two sorts of critical assessment. One examines whether it provides a suitably comprehensive number of parameters. The other examines whether those parameters have been adequately conceptualized. What we have been saying so far about the contrast between the GW-schema and the AKM-schema instantiates both kinds of critical approach. We are suggesting, both that the AKM-parameters are too few and that their interpretation is too narrow to afford a suitably comprehensive representation of the structure of abductive inference. Accordingly,

Proposition 3.15 (Under-representation) *The AKM-schema under-represents the logical structure of abduction.*

A related difficulty presents itself at the conclusion of the revised AKM-schema. H is detached without due regard to its intrinsically conjectural character. Against this it might be said that the "therefore" of the last line is qualification enough, since it is obvious that it denotes a weak conclusional link, something along the lines of "it is plausible to conclude that H". But this is wrong. What (7) requires is something like "it is plausible to conclude that H is *a justified conjecture*". Again, it may very well be that this too is assumed, and left schematically implicit in the interest of clutter avoidance. Even so, how much of abduction to try to capture schematically is an important question. Omissions of abductively salient factors need to be justified.

3.8.2 The Good that AKM Does

At a certain level of abstraction, the AKM-schema does valuable work for a certain range of cases — viz., the consequentionalist ones. It scores well on the following points.

1. The AKM-schema acknowledges (tacitly) the cognitive-deficit character of abduction problems.
2. It highlights three subtasks for abductive logicians.
   a. They must give an account of $\looparrowright$, when applicable
   b. They must give an account of H
   c. They must give an account of the therefore-operator.

Before leaving this matter, let us attend to a slightly different example. Suppose, again, that the agent's target is to have a proof of P. Let it be that neither K nor H entails P, but that $K(H)$ does. If the abducer is satisfied with this, he is downgrading his solution in a quite crucial way. He started out questing for a proof of P, but he settles for a conditional of proof. In other words, he satisfices — a fact that is unrepresented in the AKM-model.[10]

The moral we draw from this brief discussion is that we won't get the logic of abduction right (or anyhow deeply or comprehensively right) unless we let it loose on structures that reflect all the essential peculiarities of abduction on the hoof. The GW-model is offered with this imperative in mind. It retains the programmatic virtues of the AKM-approach, but ventures beyond.

Accounts of abduction that flesh out structures such as the AKM-schema or the GW-schema are sometimes called *models of abduction*. Modelling a concept or set of concepts is a methodological commonplace for logicians. It is commonplace that counsels a considerable circumspection in attributing under-representation to a model. This is because all models, to some extent or other, are under-representations of their explicanda. How, then, can the criticism of the AKM-model embodied by our Proposition 3.15 be justified, given that the same Proposition is likewise true of the GW-model? Isn't the GW-model a standing invitation to a charge of *tu quoque* from supporters of the AKM-schema?

[10] We note in passing that if an abducer is sufficiently at lease with this presumptive proof of P to detach H conjecturally, he may subsequently take "the next logical step". He may declare it an axiom! Since axioms are (save for auto-demonstration) insusceptible of proof in any system they govern, to choose an axiom is to stipulate it. But what is stipulation but conjecture with a certain swagger?

In any model of abduction (or anything else the logician turns this technique upon), some facts about real-life abduction will be suppressed or ignored. Others will be retained but also idealized. These suppressions and idealizations the modeler typically justifies sometimes on grounds of comparative unimportance or low salience, and sometimes on grounds that doing so enables the model to demonstrate interconnections or systematicities that enrich the model's clarity or explicational heft.

Still, it is plain that some models do better than others on the score of clarity and explicational success, and that sometimes this betterness pivots on comparative numbers of parameters and comparative scope of interpretation. So the criticism expressed by Proposition 3.15 needs to be re-phrased.

Proposition 3.16 (Under-representation again) *Considered in relation to a comprehensive logic of abduction, the AKM-model is under-representative to the point of distortion.*

3.8.3 The Reach of Abduction

We can say that a logic of abduction will have at least two sublogics. In the caseof consequentialist abduction, one of the sublogics gives an account of the requisite consequence relation that the abductive schema reflects (subtask (a)). The other gives an account of its conclusional operator "therefore" (subtask (c)). A contentious question in relation to subtask (b) is whether a sublogic for H exists, and, if so, how it would go. The H-factor presents the abduction theorist with at least two questions.

1. What are the conditions under which hypotheses are thought up?
2. What are the conditions under which hypotheses are deployed?

It is easy to see that part of the answer to (2) is that deployed H's should honour the abductive schema. In the AKM-model, H is required not to bear $\looparrowright$ to E, and not to be inconsistent with K. It is also required that $K(H)$ be minimal. In the GW model, the conditions on H are less specific. The reason for this is that we are unsure about the AKM-constraints. Let us take these in order.

a. $H \not\looparrowright E$ (*H's deductive independence*): The AK-model allows for $\looparrowright$ to be a deductive consequence. There are lots of cases in which a solo P bears $\looparrowright$ to a payoff V. Why rule it out that such a P might be a candidate for H? The answer appears to be that allowing it would preclude this fact from constituting a DN-explanation of V. So it would. But not all explanationist abduction is DN, and not all abduction is explanationist. So we find the constraint to be over-narrow.

Another reason for the deductive independence of H is to discourage trivial abductions in which $H \looparrowright V$. But again, independence over-determines the objective. Its more realistic accommodation is by way of abductively motivated constraints on the $\looparrowright$ relation itself, in light of the cognitive aim represented by T.

b. *H's consistency with* K[11]: There are cases in which the abducer is required to reason from data-bases that contain unresolved inconsistencies. Juries, for example, must determine the guilt or innocence of accused persons from evidence-bases that are routinely inconsistent. Verdicts are based on the acceptance or rejection of what lawyers call "theories of the evidence" or "theories of the case". A theory of the evidence is an abduction that generates a verdict on the strength of what best explains the evidence, inconsistency and all.[12] Here, too, we find the constraint excessive. [13]

c. *K(H)'s minimality*: An ambiguity lurks. Does the condition require that H be the least modification of K that delivers the intended goods? Or, does it require that H modify the least class of K that delivers the goods? Or does it mean both? What we have here, in all three cases, is a contingency elevated to the status of a logically necessary condition. It is true that abduction problems don't require for their solution everything whatever the agent may know at the time. It is also true that winning hypotheses aren't wantonly redundant. In actual practice, abductive reasoning is from subsets of K augmented by not overly redundant hypotheses. This is a fact for our schematic models to take note of. But minimization achieves this end over-aggressively.

Minimality is also a way of averting the useless proliferation of abductions by deductive closure on winning Hs. So, if H is a winning hypothesis, we don't want it to be the case in general that $H \vee Q$, for arbitrary Q, is also a winning hypothesis. But, here too, the more natural mode of discouragement is not the banishment of all redundancy, but rather constraints on the consequence relation in light of the cognitive content of T.

[11] See here [Boutilier and Becher, 1995]

[12] This may appear to generate a very bad problem for criminal jurisprudence. If the standard in criminal trials is *proof* beyond a reasonable doubt, how can it be envisaged than an abductive *conjecture*, however confidently made, could rise to it? This is discussed in greater detail in chapter 8 below.

[13] We return to the inconsistency issue in chapter 5 below.

3.8.4 Simplicity

Minimality is sometimes thought to recommend itself on grounds of simplicity. There is no doubt that simplicity has its attractions. But we join with those who find, in Kuipers' words, that "simplicity may well retard empirical progress" [Kuipers, 1999; McAllister, 1996, p. 322] and [Rescher, 1996][14]. Presumably Kuipers is drawn to minimality for reasons other than simplicity. We suppose that it is an entirely natural desire to purge abductive inference from excessive redundancy.

A further reason to distrust a simplicity requirement for inference is that the most simple assumption is not always self-announcing. Here is an example drawn from [Goddu, 2002, p. 15]. Consider the argument

1. All monkeys are primates.
2. So, with certainty, all monkeys are mammals.

It may strike us, as it does Goddu, that the simplest implicit premiss that will make this argument valid is

(1*) All primates are mammals.

If simplicity here is weakness, the present claim is false. As Hitchcock [2002, p. 158] rightly points out, the weakest missing premiss is, in fact,

(1**) Either not all monkeys are primates or all monkeys are mammals.

That (1**) is simpler than (1*) is shown by the fact that whereas (1*) entails (1**), it is not the case that (1**) entails (1*). Yet no one in his right mind would require that an actual reasoner conform his reasoning to the requirement that the assumption of (1*) be rejected in favour of the assumption of (1**).

3.9 The Cut Down Problem

Perhaps the greatest problem posed by the thinking up of hypotheses is that, on any given occasion, a candidate for selection occupies an up to arbitrarily large space of possibilities or (candidate space). Whatever the details, it appears that abductive agents manage to solve what might be called a cut down problem. One of the attractions of Atocha Aliseda's semantic tableau approach [1997] is that it reveals

[14] Rescher: "Simplicity — is it not comitted to the idea that nature proceeds in fundamentally simple ways? By no means! We have no ground whatever for supposing the "simplicity" of nature. The so-called Principle of Simplicity is really a principle of *complexity management*"[Rescher, 1996, p. 26] Emphasis added

the structure of cutdown for certain ranges of cases. But these are rather narrow ranges, as we shall soon see, made so by the technical constraints that semantic tableaux impose. In a more general sense, it would appear that the hypotheses that an abducer actually entertains are relevant and plausible subsets of large candidate spaces. (We note in passing that the idea that the minimality condition seeks to honour is handled here non-quantitatively by relevance and plausibility filters.) It is doubtful that the full story of the dynamics of cut down can be told in any logic, no matter how capacious; but part of it, certainly, requires the logician's touch. Accordingly,

Proposition 3.17 (Relevance and plausibility) *In giving an account ofH (subtask (b)), an abductive logician should deploy the resources of the appropriate logics of relevance and plausibility.*[15]

It would seem that plausibility also bears in a central way on the question of hypothesis *selection.* It is implicated in a further step of the cut down process. It cuts down the set of *entertained* hypotheses to subsets (ideally a unit set) of the most plausible.

Abductive reasoning is shot through with considerations of plausibility and presumption. In the GW-model it is explicit that presumption plays a role. It plays it in two connected ways. If we have a successful H, then $K(H)$ will hit the abducer's target presumptively. Correspondingly, it may plausibly be inferred that the conjecture of H is justified; that is to say, that the presumption of H is reasonable. Most of the work to date on the logic of presumption has been done by default logicians in the computer science and AI communities. As we have them now, such logics haven't adapted well to the particular requirements of abduction. There is work still to be done.

Proposition 3.18 (Presumption) *The logic of the conclusional operator "therefore" (subtask (c)) should subsume an appropriate logic of presumption.*

Finally, if the GW-model is taken as accurate, or something close to it, it is advisable to pay sharp attention to the differences between, on the one hand, R, the attainment relation on T and R^{pres}, the presumptive attainment relation on T, and, on the other, (when applicable) to $\looparrowright$, the consequence relation on $\{K(H), V\}$, where V is a payoff for T. At the same time, the conditions on what qualifies a V as a payoff for a T require fleshing out.

[15] For relevance, see [Gabbay and Woods, 2002]; for plausibility, see [Rescher, 1976a] and see chapter 7 below

3.9.1 Abduction as Practical

We have said that we would expect a theory of abduction to do well if embedded in a suitably general practical logic of cognitive systems. This would seem to suggest that abduction is intrinsically, or at least dominantly, a form of reasoning attended by comparatively scant resources in quest of comparatively modest targets. We imagine that not every reader will see things in quite this way. Perhaps they would be minded to ask whether we are prepared to declare abduction off-limits for institutional reasoners. Our reply is that the practicality of abduction is a matter of degree. The ignorance-condition decrees that no matter how lofty an abducer's goal, it must be a lesser thing epistemically than that cognitive level of his knowledge-base then and there. Thus it is intrinsic to abduction that abductive reasoning is more practical than the reasoning that extends from (and in this case fails) his knowledge-set then and there. It also bears on this matter that abduction is a substitute for exploration. This is a welcome economy in as much as conjecture is often cheaper than the acquisition of relevant new knowledge. It is true that computers enable us to make exhaustive searches of enormous possibility-spaces. It is also true that some programs achieve very drastic cut downs of such spaces very quickly. To the extent that abduction involves the picking out of an H from up to arbitrarily large candidate spaces, perhaps we should say that some abduction problems are well-matched to the resources that typify theoretical agency. It won't work. Abduction problems, no matter whom they are solved by, involve the timely selection of an H from sometimes huge spaces. Whether Harry performs the abduction or HAL does, H is selected in a timely way. The difference here is that it is not intrinsic to Harry's abductive success that he exhaustively search through the huge possibility spaces of which his winning $H$is a member. If HAL makes such a search, it is doing something not typical of abduction. So abduction does indeed retain the practical cast that we have claimed for it.

3.9.2 Proof-theoretic Abduction

Care should be taken not to leave the impression that explanationism is all there is to the consequentialist approach to abduction. Explanationism may well be the favorite model among philosophers of science, but among computer scientists and formal logicians, there is considerable enthusiasm for what might be called a proof-theoretic orientation. In so saying, the notion of proof enters the model in a rather general way and at a certain level of abstraction. In its most basic sense, a trigger for this conception of abduction is a *database* K that doesn't prove some desired formula or unit V. (Schematically, $\langle K \looparrowright V))$. The object of the exercise is to abduce a H such that it together with K now succeeds in proving V. (Schematically, $K(H) \looparrowright V$). In this set-up $\looparrowright$ is any consequence relation that bears the intuitive interpretation of 'proves', and V can be considered the payoff proposition in relation the target calling for means sufficient for proving it. In a great many treatments, little attention is paid to the fact that if such inferences are abductive, H cannot be on the same epistemic level as K. Relatedly, little attention is paid to the extent, if any, that this fact requires that the consequence relation be attenuated in order to deliver the goods expressed by the fact that $K(H) \overset{\circ}{\mapsto} V$. For all these reasons, and more, it is clear that this proof-theoretic orientation comports well with the AKM-model (which is another good reason not to give the model short shrift). In its proof-theoretic mode, both the *AKM* and *GW*-models are normally associated with what are (misleadingly) called inferentialist approaches to abduction. This would be a good point at which to take note of different approaches from (narrow) inferentialism. One is prominent in AI, beginning with Pople's influential paper of 1973 [1973], and developed in a number of subsequent works by Pople and other investigators, especially those working in logic programming [Kakas *et al.*, 1995], knowledge assimilation [Kakas and Mancarella, 1990] and diagnostics and other forms of medical reasoning [van den Bosch, 2001], [Poole *et al.*, 1987], [Peng and Reggia, 1990; Josephson and Josephson, 1994] and [Ramoni *et al.*, 1992]. Abduction is also dealt with in Bayesian networks and connectionist logics [Josephson and Josephson, 1994; Konolege, 1996; Paul, 1993; Flach and Kakas, 2000].

Two further contexts for abduction should also be mentioned. One is linguistics [Hobbs *et al.*, 1990; Chomsky, 1972; Heim, 1991; Gervás, 1995]. The other is mathematics [Polya, 1945; Polya, 1954; Polya, 1962; Russell, 1907][Gödel, 1944; Gödel, 1990a; Gödel, 1990b]. We take up the issue of interpretative abduction in chapter 9. Aspects of mathematical abduction will occupy us in chapter 5. Inferentialist approaches have a dominantly semantic orientation, concentrating on the specification of truth conditions on the implementation of the abductive inference schema. AI developments emphasize the role of algorithms in constructing abductions. In chapter 6 we examine diagnostic-abduction, [Paul, 1993, pp. 109–152], and in that same chapter, as well as chapter 9, we revisit connectionist abduction. Here we shall give a quick sketch of the basic structure of the logic programming approach and of the rudiments of how abduction fares in theories of knowledge assimilation.

Logic programming arises from pioneering work by Robert Kowalski and Alan Colmeraner in 1974, [Kowalski, 1979; Lloyd, 1987]. One of its principal implementations is Prolog. Its underlying logic is first order. Prolog comprises a program P, queries q and a problem-solving device R, called resolution. We consider an elementary example.

Program P

lawn-wet← rain.

lawn-wet←sprinklers-on

(These are Horn-clauses in which the contained terms are literals).

Query q: lawn-wet

We say that a query *succeeds* when it is derivable from the program. In the present example q does not succeed. At this juncture, Prolog moves into an abductive mode. As can easily be seen, q would succeed if one or other of certain possibilities were added to P as hypothesis. These possibilities are: *rain*, *sprinklers-on* and *lawn-wet*. Abduction is here understood on the process evincing these possibilities. In its non-abductive mode, the failure of these to be listed in P as facts would trigger failure. In its abductive mode, the resolution mechanisms introduces these "non-facts" as hypotheses. The query now succeeds.

It is important to emphasize that the resolution device is constrained in what it can select as hypotheses. It is required to select only from sub-goal literals that fail under the backtracking operation. Thus not everything that would be reasonable to forward as a conjecture is allowed in this approach. A further limitation of the logic programming orientation — one that it shares with standard systems of diagnostics — is that hypothesis-selection must be made from a pre-determined set of abducibles [Kakas *et al.*, 1995], which in turn are required to be validated by further conditions, called integrity constraints, introduced so as to mitigate the problem of computational intractability.

Many inferentialist and most computer-based approaches to abduction pivot on the fact that for some background K and a payoff V, neither $K \not\rightsquigarrow V$ nor $K \not\rightsquigarrow \neg V$ holds. This is a substantial constraint, excluding from consideration abductions arising from new facts that contradict what would otherwise have been expected from K. (See the discussion of the case of *The Open Door* in chapter 7.) Knowledge-assimilation approaches are organized to take into account these excluded cases. They are theories of belief revision prompted by such phenomena. Typical settings for this kind of abductive trigger are diagnostics [Peng and Reggia, 1990]; (see below, chapter 6), belief-revision in databases [Aravurdan and Dung, 1994] and theory tweaking in machine learning [Ginsberg, 1988]. Ensuing from Gärdenfors theory of belief change [Gärdenfors, 1988], the principal means of incorporating new information into a database, a scientific theory or domain of common sense beliefs are the operations of *expansion*, *contraction* and *revision*, none of which is intrinsically abductive, contrary to what is rather widely supposed. However, certain of these operations is adaptable to the conjectural requirements of abduction. Expansion is simply a matter of adding a new fact P to K. Doing so enlarges K to $K(P) = K \cup \{P\}$. But this is not abduction. P is not conjectural, and $K(P) \not\rightsquigarrow P$, where $K(P)$ is itself a knowledge-set at epistemic par with the original K. Contraction is different. It requires the deletion of a subset of K, with which the new fact P is inconsistent. Except in those rare cases in which the structures of K and P permit unique determination of the candidate for deletion, there is room in reaching these judgements for conjecture. The same is true of the revision operation, which is a composite of expansion and contraction. In standard systems of knowledge assimilations, additional constraints are imposed. Two of the most prominent involve closure under belief-change operations and consistency of outcome. As these tend to be classical constraints, they alienate such systems from the give-and-take of real-life belief changes in real time. In the coming pages, the schema we eventually adopt for abduction will meld various of the features of inferential, computational and knowledge assimilation approaches. In a general sense, then, all the standard approaches to abduction are inferential.

The challenge presented by abductive logic is to remove the cognitive irritations to which they give rise. The general form of this removal is *conjecture* and *discharge*.

By any standard, providing a full logic of abduction will prove to be a formidable task. We do not have it in mind to perform it fully. Instead we shall content ourselves with the presentation of what we take to be key results in various areas that an abductive logic encompasses. Beyond that we will offer suggestions and promissory notes. If we wanted to be old-fashioned about it, we could say that *The Reach of Abduction* is our prologomenon to the logic of abduction. Much in this spirit, rather a lot of what we will propose here will be prescriptive rather than executional. We shall be proposing what a full logic of abduction should look like and what its principal tasks would be. Our reserve in this regard is not born of undue modesty. There are some parts of the programme of a logic for abduction that we don't yet know how to execute. In other cases the opposite is true. We know (or think we know) how those things go. *The Reach of Abduction* is as much a call to arms to the research community at large as it is a set of settled skirmishes.

We have remarked abduction's connection to what Reichenbach called the 'context of discovery', which he contrasted with the 'context of justification' [1938]. Abductions in this sense are said to be the business of the logic of discovery. Reichenbach was not alone in thinking it possible to have a logic of scientific justification, which, as he supposed, is precisely what an inductive logic is designed to be. But a logic of discovery, or an abductive logic, Reichenbach regarded as a mistake in principle, because it confuses psychological considerations with logical considerations. Reichenbach's skepticism was shared by most logical positivists, but as early as 1958, Hanson [1958] worked up a contrast between reasons for accepting a given hypothesis and reasons that suggest that the correct hypothesis will be one of a particular kind or description.[16] A theory which analyzes reasons of this second kind Hanson called a logic of retroductive reasoning. In Hanson's account, a logic of retroduction resembles a logic of analogical reasoning. Hanson's efforts were criticized — even pilloried. This had as much to do with their novelty

[16] More recent discussions of the logic of discovery include[Laudan, 1980; Nickles, 1980; Musgrave, 1989; Kelly, 1989; Savary, 1995; Kuipers, 2000].

as with their deficiencies. Even so, the very idea of a logic of discovery trails some important questions in its wash. One deals with the extent to which the supposed contrast between contexts of justification and contexts of discovery catches a hard and fast distinction. Another question is whether the contrast between a psychological account of hypothesis formation and a logical theory of the same thing will hold up. A related issue is the extent to which a heuristics for a set of reasoning problems can be distinguished in a principled way from a logic for such problems. A further matter — also closely connected with the others — is whether a theory of hypothesis formation is able to retain a sharp distinction between descriptive adequacy and normative soundness.

3.10 The Adaptive and the Epistemically Subpar

It bears repeating that in its most stripped-down sense, abduction is a procedure in which something that lacks epistemic virtue is accepted because it has virtue of another kind. The fundamental structure of an abduction problem is that a target we desire to hit cannot be hit with anything that we presently know. This being so, the requisite ignorance must be invariant throughout the process of abduction. It is quite true that, once deployed, hypotheses are often made the objects of attempts to confirm them (or corroborate, for those of Popperian bent). Sometimes these attempts are met with success. But this is not abductive success, but rather post-abductive. Accordingly, we have

Proposition 3.19 (Hypotheses and epistemic virtue) *If H is an hypothesis entertained or engaged in an abduction exercise, it is essential that H lacks some degree of epistemic virtue.*

In standard approaches to abduction, the epistemic-deficit condition is recognized only tacitly if at al. An exception is the adaptive logic orientation of Meheus and her colleagues [Meheus *et al.*, forthcoming]. In this manner of proceeding, abductive inferences are modelled as proofs in a modalized adaptive logic. In such proofs, priors carry the necessity operator $\Box$, whereas abductive conclusions carry the weaker modality of possibility, $\Diamond$. There is an apparent *structural* congruence between the $\Box$-statements and the $\Diamond$-statements of an adaptive abductive logic and the asserted priors and conjecturized conclusion of abductions according to the GW-schema. However, as might be expected the two distinctions reflect significant conceptual dissimilarities. Even so, we have here structural acknowledgement of the ignorance condition.

3.11 Knowledge-Sets

The idea of knowledge-*sets* is something of an understatement. We may take it that the totality of what a human knower knows at a given time is a comparatively fuzzy assemblage of various modules. This modularity has something to do with the variability of our epistemic capacities and circumstances. A proposition that lacks a proof may be thought not to qualify as mathematical knowledge; a proposition that fails to negotiate the rigours of scientific method may not be thought of as scientific knowledge; a proposition that failed the requisite standard of forensic rectitude may not qualify as legal knowledge; and so on. On the other hand, propositions of this sort might well qualify as common knowledge.

There follows from this a point of considerable importance for a theory of abduction. When an abductive trigger presents itself to an agent, a target is created which cannot be hit with the agent's present epistemic wherewithal. Whenever this is so, the failure to hit the target is always a failure to hit it in accordance with the requisite epistemic standards. If T is the target of explaining some event scientifically, then the knowledge that the abducer lacks is not all knowledge that might bear on T, but rather the *scientific* knowledge by which T could be explained. This same relativity is present in the case of hypothesis selection. In finding an H that would enable the abducer to produce the desired scientific explanation, the abducer must hypothesize that H is a serious candidate for scientific knowledge, never mind that it does not presently so qualify. It might be thought that the ignorance condition obliges the abducer to refrain from selecting H from anywhere in his K-set. This is a serious misconception. The prohibition extends only to the module that harbours (in the present case) his *scientific* knowledge, a prohibition that is trivially satisfied anyway by the very structure of an abductive trigger. The abducer is free to select a candidate from any other K-module, provided he is prepared to "bet" — apart from its requisite abductive fit — that, once discharged, it has a chance of performing in ways that would elevate it to membership in the scientific K-module. Subject to this key limitation, an abductive agent is free to search his own K-set for possible hypotheses. An even greater latitude is open to him with regard to his *belief*-sets and his *plausibility*-sets.

It is here worth repeating that not every abduction problem is an epistemic abduction problem. A *doxastic* abduction problem as the triggering of a target that can't be hit with what the abducer currently *believes* (and let us assume the modularity of belief in the same general kind of way, and to the same kind of end, as we have just done with knowledge). By the same token, it should also be possible to speak of *probability* and *plausibility* abduction problems, in which a target cannot be hit with what the abducer takes to be probable or, as the case may be, plausible. Here, too, given that probability and plausibility comes in degrees, the problem is that the target in question cannot be hit with anything he takes to be probable or plausible to a certain degree or higher. Lower probabilities and plausibilities are free to be mined with a view to their possible subsequent upgrading.

Proposition 3.20 (Multiple relativity) *Abduction problems are definable in relation to knowledge, belief, the probable and the plausible. In each case, the problem is relativized to the requisite module or modules.*

As earlier remarked, unless we indicate the contrary, our discussion of K-abduction will stand in for all these varieties.

The relativity of abduction in relation to the variability of its K-modules carries direct consequences for abduction's *conjectural* component. It portends a distinction between two sorts of conjecture within which the factor of variability recurs. The distinction is one between what we might call *cold-start* conjectures and *upgrade*-conjectures. Cold-start conjectures answer well to the Peircean element of abductive surprise (even more emphatically expressed by N.R. Hanson as astonishment). Cold-start conjecture is required not only when there is nothing in the reasoner's K-set that hits the desired target T, but also when nothing in K-sets of lesser epistemic stripe or in belief sets, probability sets or plausibility sets appears to do the job either. *In extremis*, this requires the conjecturer to do some *originary* thinking, as Peirce calls it; that is, to think outside the box. In conditions of such austerity, prior belief must act at arm's length. There are two principal ways in which this happens. Not having any beliefs that strike him as adequate for the hitting of his abductive target, the conjecturer is free to take a novel step and reflect upon what may strike him as *possible* candidates. He may also review his prior beliefs in hopes of finding there occasion for *analogical extension*. Thus,

Proposition 3.21 (Possibility and similarity) *Two of the primary operations of cold-start conjecture are* modalization, *i.e., the recognition of possibilities, and* analogy, *i.e., the recognition of similarities* in difference.

Further,

Proposition 3.22 (Creativity and ignorance) *In solving abduction problems, the demands of originary (or creative) thinking are proportional to the depth of the abducer's ignorance.*

Upgrade-conjecture operates in a less arms-length way. It allows deployment even of prior beliefs the abducer holds with (up to) substantial conviction and on the basis of (up to) substantial evidence. What cannot be allowed is that these beliefs, evidenced in these ways, are such as to hit the epistemic standards of the K-set with respect to which the abducer's problem arose in the first place. It is this feature that motivates the upgrading character of conjecture. For if a candidate proposition H is already firmly believed on good evidence, there is no occasion to conjecture *that* H. Rather the conjecture can only be that H hits the epistemic standards of K, or higher. In contrast, in cold-start conjecture, there is room for the suitably modest conjecture *that* H, together with the conjecture that H *meets the abductive problem's requisite epistemic standard.*

Accordingly, we propose a further adequacy condition. A conjecture is a kind of proposal. In strictness, this imposes restrictions on how to interpret a conjecture's fit with the that-clauses "that H" and "that H hits epistemic standard k". Thus, where it makes good sense to speak of someone's proposing, say, that Harry leaves the room at once, it is a stretch, at best, to find this equivalent to proposing that "Harry leaves the room" is true (or probable, or plausible). Proposing is a performative, and conjecture, being proposalistic, absorbs this same feature. It may be that this is what explains Peirce's understanding of the conjecture that H (or that P meets standard k) as the *assertion* that H is a suitable candidate for testing with respect to its claim on truth or meeting epistemic standard k. Although we ourselves think that Peirce wrongly confuses the conjecture that H with the decision to send P to trial, even so,

To return to an earlier example — a variation of anomaly: triggered abduction — let K be a proof of a wff $A \wedge \neg A$, the proof rules of simplification, addition, disjunctive syllogism, and the wffs A, $A \vee B$, $\neg A$, B, which follow from predecessors in accordance with those rules. Then the $\{K, \vdash\}$-situation S that an abducer is faced with is that a contradiction $A \wedge \neg A$ proves arbitrary B. If the abducer is not minded to accept this result, then his abduction problem may be triggered by that fact. The problem is solved if, by manipulation of K, the undesired consequence no longer obtains. Let us suppose that, upon reflection, the abducer deletes the disjunctive syllogism rule from K, declaring it now to be admissible at best, but not valid.

Our example is instructive in another way. As we set it up, our would-be deducer knows that disjunctive syllogism is a valid rule. However, at the conclusion of his abductive reasoning, he hypothesizes that disjunctive syllogism is *not* a valid rule. By his own conjectural lights, therefore, the abducer must now dispute the accuracy of having placed this rule in his pre-abductive K-set.

We repeat the point that propositional contents of adjustments to an original database in an abduction must be introduced as hypotheses, rather than as premisses expressing known facts, or purported facts. So, in our second example, the removal of disjunctive syllogism counts as abductive only if it is done *conjecturally* and subject to the condition of discharge.

Historical note. In their early note on a simple treatment of propositional logic, Anderson and Belnap [1959], produced the system and the noted *en passant* that it couldn't consistently permit the validity of disjunctive syllogism. Instead of faulting their formulation, they performed an abduction. They conjectured that the best explanation of the fact that their formulation excluded disjunctive syllogism was that disjunctive syllogism was an invalid proof rule. (Afterwards, they asserted it as if it were a fact that spoke for itself. But that was afterwards.)

3.12 Filtration Structures

Let H be a winning hypothesis for a given AP. It is plausible to suppose that H exists in an arbitrarily large space $\mathbb{S}$ of possible hypotheses for AP. We can take it to be well beyond the agent's computational capacity to make an *exhaustive* search of $\mathbb{S}$ How then does he "find" H in $\mathbb{S}$? It would appear that beings like us are capable of economically motivated partial searches.

Little is known of how beings like us search large possibility-spaces. Even so, there are certain conjectures that we might make. One, as we have already mentioned, is that in the early stages, the search for a hypothesis is a *memory* search. There is a particular reason for this. There is no immediately discernible phenomenological difference between a cognitive irritant that can be soothed by calling to mind the requisite piece of knowledge, and a cognitive irritant that can be soothed only by putting into play something not known. In cases of actual irritation, a reasoner often can't tell which of the two situations he is in. When this is so, it is natural for the agent in question to conceive his quest as a search for what he already knows. Naturalness aside, there is a certain economy in so proceeding, that is, in having a search for conjectural possibilities start out as a memory search. Doing so takes full advantage of the relativity of what is sought for. The ignorance-condition on abduction precludes the removal of a cognitive irritant by means of anything having a certain high enough degree of epistemic virtue. But this leaves it entirely open that the solving proposition will nevertheless already be in the abducer's memory store, as either a proposition of lesser epistemic virtue or as something he believes or finds probable or plausible. Thus the economic advantage is that the winning hypothesis may already be in the agent's mind.[17]

The point at hand also requires an admonitory word about the role of conjecture and hypothesis. Suppose that, for a given case, the winning hypothesis is indeed already in the agent's mind. Suppose that it is something he already believes and that, given the epistemic standard evinced by the abducer's K-set it is, as required by that standard an epistemically lesser thing. Even so, the proposition in question is not a hypothesis for the agent; it is something he already believes. How, then, can this belief be a candidate for abductive conjecture? The answer is that the conjecture is not that the proposition has the degree of epistemic virtue that it

[17] The *scope* of this advantage has been a matter of speculation since antiquity. See[Magnani, 2001a] and [Paavola and Hakkarainen, forthcoming] for a discussion of the link between the paradox of knowledge in the *Meno* and the structure of abduction.

already is taken to have as an object of the abducer's belief, rather that it has the degree of epistemic virtue that qualifies it for membership in a K-set.

Our conjecture, then, is that large candidate-spaces are filtered by subsets of objects of the abducer's memory. But memory searches are themselves searches of considerable magnitude. It is necessary to ponder additional filters.

The space $\mathbb{S}$ of conjectural possibilities can be seen, on the present assumption, as input to the next step in the abduction process, as input to the *engagement*-sublogic, as we might say. Engagement is a way of activating (some of) the possibilities in $\mathbb{S}$. Here we might suppose that the activation condition is one of *relevance*. Relevance we can take as a filter that takes $\mathbb{S}$ into a proper subset $\mathbb{R}$ of relevant possibilities. Thus the relevance filter would appear to cut down the space $\mathbb{S}$ to the smaller space $\mathbb{R}$.

The relevance filter plays a role similar to Harman's Clutter Avoidance Principle, which bids the reasoner not to clutter up his mind with trivialities [Harman, 1986, p. 12]. In its role here, the relevance filter enables the abducer to concentrate on possibilities that are fewer than those found in $\mathbb{S}$, and make a better claim for activation. $\mathbb{R}$, the space of relevant possibilities, advances subsets of $\mathbb{S}$ closer to the point of activation.

A second task is to bring hypothesis-activation off, to replace possible hypotheses that are conjectural possibilities with hypotheses that the abducer actually deploys by *actual* conjecture, and thereafter releases for service as premisses in further inferences. This, too, has the look of passing larger sets through a contraction filter. The relevance filter took possibilities into relevant possibilities. What is wanted now is a finer filter. We have already mentioned that the abducer is free to filter such possibilities through the screen of what he knows, or believes, or takes as probable or possible, provided that no such proposition achieves the epistemic standards achieved by (the module) of the knowledge-base from which the abductive target cannot be hit (a cut down of $\mathbb{R}$ to $\mathbb{P}$).

Earlier we remarked on the distinction between propositions whose contents are plausible and propositions it would be plausible to conjecture. The two notions can be said to intersect, but it is necessary to say that they do not coincide. Newton thought that the notion of action at a distance had no plausibility as regards to its content, but he also thought that this most implausible of ideas was a plausible candidate for conjecture. Accordingly, we may postulate two kinds of plausibility filter or screen. In the pages just above, plausibility screens are those furnished by content. But it is also necessary to emphasize the tactical importance of plausible conjecture; for it is conjecture, after all, that abductive solutions turn on. In what follows, context will determine which sense of "plausible" is in play.

The plausibility screen accordingly shrinks $\mathbb{R}$ to subset $\mathbb{P}$, $\mathbb{P}$ presents the abducer with three activation options. Assuming $\mathbb{P}$ to be nonempty,

1. If $\mathbb{P}$ is a unit set of $\mathbb{R}$, then engage the hypothesis in $\mathbb{P}$.
2. If $\mathbb{P}$ is a pair set of $\mathbb{R}$ or larger, then engage all the hypotheses in $\mathbb{P}$.
3. If $\mathbb{P}$ is a pair set of $\mathbb{R}$ or larger, then engage the most plausible hypothesis in $\mathbb{P}$.

As an aid to exposition we can now use the letters $\mathbb{P}$, $\mathbb{R}$ both as names of the presumed filters and as the names of the resultant sets. The same is true of $\mathbb{M}$. It names either the order on $\mathbb{P}$ or the resultant set. The triple $\langle \mathbb{S}, \mathbb{R}, \mathbb{P} \rangle$ represents a filtration structure on an initial space of possibilities in which the succeeding spaces are cutdowns on their predecessors. It is possible that there also exists a further filter $\mathbb{M}$ on the space of plausibilities to a unit subset, of (intuitively) the most plausible candidate for engagement/discharge. It is important to emphasize that $\langle \mathbb{S}, \mathbb{R}, \mathbb{P} \rangle$ and, $\langle \mathbb{S}, \mathbb{R}, \mathbb{P}, \{\mathbb{M}\} \rangle$ exist independently of whether any abductive agent has ever thought of it. The filters that cut sets down to smaller sets at each stage do so independently of whether anyone has actually tried to deploy those filters. In the speculation that we have been entertaining these past few pages we have put it that in reaching his desired H, the abductive reasoner proceeds by solving the cut down problem by constructing (or reconstructing) the requisite filtration structure. In so supposing we have it that the would-be engager of H proceeds in a top-down fashion, first by entertaining mere possibilities, then cutting to relevant possibilities and finally to the most plausible of these. But there is reason to doubt this supposition.

Proposition 3.23 (Individuals and filtration structures) *Everything that is empirically known of actual human cognitive agency* on the ground *suggests that constructing filtration structures is* not *the* modus operandi *of the individual abducer's reasoning; for one thing, it would involve the searches of spaces that are computationally beyond the reach of such agents.*

Corollary 3.23(a) *Even if actual reasoners did on occasion construct filtration structures, there is reason to think that the order of the application of its filters would admit of some variation. For example, considerations of what is relevant and plausible might bear on an agent's determination of what he finds* possible.

This can hardly be a surprising finding, given what is known of the fast and frugal character of practical reasoning. So, while it is perfectly reasonable to suppose that hypothesis-engagement involves considerations of relevant plausibility, *how* it does is not all that clear.

Proposition 3.23 throws some light on the so-called process-product distinction. In fact, it straddles it by honouring the joint fact that *if* a real-life abducer constructed a filtration-structure, the winning hypothesis, if there were one, would show up in it in a determinate way; yet, there is little evidence that in the give-and-take of abduction resolution as actually played out, the process of finding a winning hypothesis seems not to involve the actual construction of this product we are calling filtration structures.

Let us be clear. There is reason to think that, if the H for which the abducer seeks exists, then there exists a filtration structure in which H exists and in which it has a determinate place. This elucidates an important aspect of *product*-abduction. On the other hand, there is no empirical evidence that in finding H abducers construct a filtration structure and determine the place in it in which H resides. This elucidates an important aspect of *process*-abduction. This is not to say that relevance and plausibility play no role in a good account of abduction. It is only to say that they appear not to play the top-down roles that they play in the construction of filtration structures. (We return to the factors of relevance and plausibility in later chapters.)

3.13 Hypothesis-Engagement

Embedded in the notion of hypothesis-engagement is the notion of hypothesis-discharge. The abductive schema's penultimate conclustion is $C(H)$. "$C(H)$" schematizes the claim that it would be *justified* to conjecture that H. "$C(H)$" expresses an assertion in which 'C' functions as a predicate or a sentential deontic operator. ("It is permitted to conjecture that H".) But might we not, with equal reason, represent the conclusion that is drawn at this juncture of the abductive schema as the speech-act of conjecture, rather than an assertion to the effect that such a speech act would be justified (in which case 'C' would function as a speech act identifier)? Some readers might think that there is nothing that favours the one interpretation over the other. This is not our own view. In our approach to abduction, it is not essential that an abducer actually conjecture H. What is essential is that the abducer perform a two part task. He must note that H is worthy of conjecture and he must discharge it, that is to say, release it for possible use as a premiss in future reasonings within or from $K(H)$. And he must do this latter in ways that ground his release-decision in his judgement of the conjecturability of H. An utterance or inscription of "$C(H)$" accomplishes the first task. An utterance or inscription of "H^c" accomplishes the second task. Accordingly

Proposition 3.24 (Conjecture in abduction) *In selecting an H that solves his abductive problem it is not strictly necessary that the agent actually conjecture that H. But it is necessary that he assert or hold that H is worthy of conjecture.*

We should be clear that hypothesis-engagement and hypothesis-discharge are not discrete, separably performable undertakings. Rather hypothesis-discharge is part of hypothesis-engagement. It is true that hypothesis-discharge presupposes the performance of a prior task. But it is not the task of hypothesis-engagement. It is the task performed by the assertion that H is worthy of conjecture.

Proposition 3.25 (Engagement and discharge) *Hypothesis-discharge is a constituent of hypothesis-engagement.*

Corollary 3.25(a) *Any process of abductive reasoning that terminated with the conclusion "$C(H)$" would express a* partial *abduction.*

Hypotheses are engaged when they are given the status of premisses in subsequent inferences in the domain $K(H)$. Some might see such inferences on the model of conditional proofs. A better analogy is that a labelled deduction, in which H's subsequent premissory occurrences bear some mark of H's conjectural history. Accordingly,

Proposition 3.26 *Full hypothesis-engagement is represented by the* pair *of conclusions $\{C(H), H^c\}$, in which 'C' is the justified conjecture modality, and 'c' labels H as having been in the scope of 'C'.*

See here [Gabbay, 2000] for a somewhat different labelled deductive approach to abduction.

3.14 Grounds of Action

In a standard situation an ignorance-problem presents an agent with two choices. One is to acquire the knowledge that solves the problem and then to *act* on it in ways that may conduce to the agent's further interests. The other is (perhaps temporarily) to admit defeat and to postpone any action that would be suitably occasioned by a solution to the problem if it existed.

As we have seen, there is also a third option. Perhaps its principal attraction is that it is an alternative to the passivity of giving up on one's IP. It is, of course, a qualified alternative, since it does not solve the IP but rather solves it presumptively.

Notwithstanding this essential qualification, an abductive solution bears on the question of *action* in two important ways. In the one case, the abducer's embrace of H^c constitutes the *cognitive act* of releasing H for generally unfettered inferential work in the domain of enquiry within which the abducer's IP arose in the first place. In the other case, it is open to the agent to take whatever *further actions* as may comport with his other interests, on the basis of conclusions in the descendent class of inferences dependent upon H. This is far from saying that H's conjectural origins are overlooked in such cases. It means only that the actions are taken so with requisite regard to the higher risk than that that would attach to actions occasioned by what the agent does really know. Accordingly, it is a deep fact about abduction that

Proposition 3.27 (Abduction as a spring of action) *Abduced hypotheses H give agents a basis for consideration of subsequent actions involving degrees of risk concomitant with the strength of H's conjecture.*

3.15 Tasks for an Abductive Logic

If we were to take the *GW-schema* as our guide, a theory of abduction would attempt to account for all the schema's parameters — $T!, T, K, K^*, H, K(H)$, $S_1, \ldots, S_n, \therefore, C(H), H^c$ and, where applicable, $\looparrowright$ and V. It is a fair question as to how any of these fall within the *logician's* ambit. If one takes the straight mainstream approach to logic (set theory, model theory, proof theory and recursion), there is not much that logic can do to elucidate these parameters. But if one takes a more laws-of-thought orientation in the manner of chapter 2, there is substantially greater prospect for logical engagement. This is especially true for our operational test of what counts as a logician's work. By that test, a logician's proper work is what he's interested in in conjunction with what he's good at. Accordingly, the challenge posed by our abductive parameters to the logician is, in effect, "Give it your best shot and then we'll see". We bring this chapter to a close with a brief consideration of how that challenge might be met. In so doing, we want to remind the reader of two important qualifications. One is that not even the most latitudinarian of logicians will assert exclusive domain over these issues. The other admonition that bears repeating is that working out just the logical aspects of the abductive parameters is a daunting job, made so both by its sheer size and the difficulty of some of its questions. For this reason, much of what we will have to say for ourselves here and in the book's succeeding chapters will be fragmentary, tentative programmatic and promissory. As someone said at the Cognitive Science Meetings in Chicago, in August 2004, "This will give us gainful employment for at least a generation!"

$T!$ expresses the desire that some target be hit. It is a form of expression that optative logicians have taken note of. *Optative logic* enjoyed a bit of a flurry in the 1950s, but seems not to have been an active research programme more recently. Nevertheless, the optative slack has been vigorously taken up by various kinds of *goal-directed* logics; and we may expect them to play a role in a final theory if ever it is produced. Abductive theories recognize that abduction arises from a disappointed hope, and that a successful abduction always manages to answer to some variation on that hope. In other words, T is an optative infinitive, with regard to which all known approaches to abduction attempt to specify realization conditions. This motivates the role of V as a *payoff* for T, in those cases in whcih abduction is consequentialist. Whatever else we may of it, V is an optative realization (or goal-directive) condition.

The *consequence relation* also plays an essential role in this process of optative closure. $\looparrowright$ is required to bear an interpretation such that the truth of $K(H) \looparrowright V$ constitutes presumptive attainment (R^{pres}) of T. Although the GW-schema mandates a greater variability in such interpretations, consequence is a bread and butter issue for logicians.

K is the abducer's present knowledge-base. K^* is his knowledge base sometime later and within the frame of the problem he encountered with K. Like Peirce, we are *fallibilists* about K. K at a time is what is taken for knowledge at that time. Later we might come to know better. So the fallibilism at hand is also a dynamical factor. Accordingly, the general setting for abductive processes can be expected to import the general structure of *dynamic logic.*

Our fallibilism complicates the structure of knowledge-succession. In particular, K^* need not contain K as a proper subset, never mind that K^* is always an adaptation K. Such problems are well-explored by epistemic logics and theories of belief-revision and belief-update. $K(H)$ also poses a belief dynamics problem.

H is a hypothesis. There exist logics of hypothetical reasoning that can be expected to play a role here. All abductive theorists recognize the necessity of subjecting H to various constraints $(S_1, \ldots, S_n)$. If, as in our approach, we do not require that $K(H)$ be consistent, then the underlying logic must be *paraconsistent.* The base logic must also be *non-monotonic.* If, as we think, minimality is not a condition on H or on $K(H)$, some attention must nevertheless be paid to the requirement that H (and $K(H)$) contain no more *redundancy* that abets smoothness of communication. Aristotle was the first logician to impose an irredundancy condition on deductions. We see no reason to withhold the factor of redundancy from the attention of the modern logician.

Other conditions are often imposed on H. One is that it be *relevant.* Another is that it be *plausible.* Logicians have produced a huge literature on relevance; and some of the basic groundwork has been done for *plausibility logics.* And if, as we believe, it is also necessary that a winning H be cognitively junior to K, this is something for the *epistemic logician* to take note of. For, again, how can it be rational to engage in a form of reasoning that is guaranteed to preserve one's original ignorance?

$\therefore$ is the abducer's conclusion operator. It faces the abductive logician with the task of specifying the inferential force of an abductive conclusion. If, as we think, its force must always be weak enough for H to honour the ignorance condition on abduction, then a logic of *plausible* (or *presumptive*, or *defeasible*) *reasoning* would have a natural place in any such specification.

$C(H)$ we have already spoken about. It is a modal sentence, with 'C' a deontic operator for permitted conjecturability. H^c denotes the release of H together with labelled recognition of its conjectural origins. So, again, $C(H)$ has a place in deontic logic, and H^c might respond well to the labelled approach to inference.

Perhaps what is most striking about abductive parameters is not their resistance to the logician's probes, but rather their collective call upon so hefty a logical pluralism. The great task of an abductive logic is to aggregate this pluralism in a systematic way.

This Page is Intentionally Left Blank

Chapter 4

Explanationist Abduction

> There is no necessity for supposing that the true explanation must be one which, with only our present experience, *we could* imagine.
>
> Lord Baden Powell

This chapter and the next develop certain features of the sublogic of the ↬-relation. Our approach is somewhat tentative and without any pretensions to completeness. Compared to developments in other branches of logic, it is still early days for the logic of abduction.

4.1 Peirce

The present chapter develops aspects of the sublogic of the ↬-relation by focusing on explanationist approaches to abduction. We begin with explanationism's most important advocate to date, Charles Peirce. The history of logic has not accorded Peirce the attention that his work demands, although there are signs that repair of this omission is now underway. (See, for example, [Peirce, 1992], [Houser *et al.*, 1997] and [Hilpinen, 2004, p. 611–658].) Writing in the tradition of Boole, Peirce extended the algebraic approach in the direction of what would become modern proof theory. His highly original logic of relatives would also play an influential role in the efforts of Schröder, Lowënheim, Skolem and Tarski to bring forth a mature model theory.

In 1898 Peirce gave a series of eight lectures at Harvard.[1] Delivered 'off-campus', these powerful and original pieces were presented at William James' behest, in yet another gesture of generosity towards James' beleaguered friend.

[1] Reprinted as Peirce [1992].

Peirce's range was striking. He had novel and prescient things to say about probability, likelihood and randomness; about quantification (of which, with Frege, he was independent co-developer); about causality, space and time, and cosmology. Of particular note is Peirce's emphasis on abduction (or retroduction, as he would also call it). Peirce had been writing on abduction since the 1860s. The Harvard lectures reflect what [Kruijff, 1998] calls the 'late theory of abduction'. (See also [Burks, 1946; Fann, 1970] and [Thagard, 1988, p. 53].)[2] Peirce came to believe that abduction is the dominant method in science, in fact, the only purely scientific method beyond brute observation. For Peirce, scientific reasoning stands in sharp contrast to practical reasoning or, as he also says, to reasoning about 'matters of Vital Importance' [Peirce, 1992, p. 110]. In this contrast he was hardly alone, although it may be mentioned in passing that it was a matter on which he disagreed fundamentally with Mill, for whom 'a complete logic of the sciences would also be a complete logic of practical business and common life' [Mill, 1974, Bk. III, Ch. i, Sec. 1].

Peirce's writings on abduction are important in a number of ways, but especially for its emphatic rejection of the possibility of a practical logic, hence of a logic of abductive reasoning in practical contexts. As should by now be clear, this is not a view which the present authors are disposed to share.

An important distinction in Peirce's logical writings is that between *corollarial* and *theoremic* reasoning. Corollarial reasoning simply draws deductive conclusions from sets of premisses, whereas theoremic reasoning analyzes some 'foreign idea' which, although it may be discharged in the final conclusion, is needed for the conclusion to be derived. Appropriation of certain set theoretic ideas in mathematical reasoning is an example of theoremic reasoning. Indirect proof is a second example. Peirce also likened abductive reasoning to theoremic reasoning.[3]

Peirce famously sees abduction as an inference in the form,

> The surprising fact C is observed
> But if A were true, C would be a matter of course.
> Hence there is reason to suspect that A is true [Peirce, 1931–1958, p. 5.189].[4]

Peirce's formulation has exercised a considerable influence on subsequent approaches to abduction. It resonates in AI investigations [Flach and Kakas, 2000], in logic programming [Kakas *et al.*, 1995], studies in knowledge acquisition [Kakas

[2] Fann identifies 1865–1875 as the early period, followed by a transition period and then, in 1890–1914, the late period.

[3] Reasoning of this form is commonplace in theoretical computer science. There is also a connection with Interpolation Theorems.

[4] The factor of surprise persists in subsequent accounts. See [Hanson, 1958] and [Thagard, 1988, p. 52].

and Mancarella, 1994] and natural language processing [Hobbs *et al.*, 1990]. Abduction, says Peirce,

> although it is little hampered by logical rules, nevertheless is logical inference, asserting its conclusion only problematically or conjecturally, but nevertheless having a perfectly definite logical form [Peirce, 1931–1958, p. 5.188].

Abduction is, or subsumes, a process of inference. This is what Kapitan calls the *Inferential Thesis*.[5] It is also a strategy for making refutable conjectures. This is what Kapitan calls the *Thesis of Purpose*. The purpose of scientific abduction is twofold. It entertains possible hypotheses and it chooses hypotheses for further scrutiny [1931–1958, p. 6.525]. The purpose of abduction is therefore to 'recommend a course of action' (p. MS637:5).[6] Abducted conclusions are not matters for belief or for probability. In abduction 'not only is there no definite probability to the conclusion, but no definite probability attaches even to the mode of inference. We can only say that ... we should at a given stage of our inquiry try a given hypothesis, and we are to hold to it provisionally as long as the facts will permit. There is no probability about it. It is a mere suggestion which we tentatively adopt'.[7]

According to the *Comprehension Thesis*, scientific abduction subsumes all procedures and devices whereby theories are produced [Kapitan, 1997]; *cf.* [Peirce, 1992, p. 5.146]. Yet 'a retroductive inference is not a matter for belief at all,. ..[an] example is that of the translations of the cuneiform inscriptions which began in mere guesses, in which their authors could have had no real confidence' [Peirce, 1992, p. 176]. Practice involves belief ineradicably, 'for belief is the willingness to risk a great deal upon a proposition. But this belief is no concern of science ... .' Belief contrasts with acceptance, as can be gathered from this passage, which might well have been penned nearly a century later by philosophers and inductive logicians in the manner of L.J. Cohen:[8]

[5] Kapitan [1997, pp. 477–478], *cf.* [Peirce, 1931–1958, pp. 5.188–189, 7.202].

[6] Kapitan [1997, pp. 477–478]. MS citations are to the Peirce MSS, in care of the Peirce Edition Project. 'MS 637' denotes manuscript 637; (5) denotes page 5.

[7] [Kapitan, 1997, p. 142] Peirce distinguishes (1) *induction*, or the establishment of frequencies in populations by sampling and (2) *probable inference*, or the inference from a known frequency in a randomly selected sample. He anticipates Neymann-Pearson statistical sampling theory with its notion of *likelihood.* By requiring that inductions have a premiss to the effect that sampling is *random*, Peirce thought that all inductions turn on prior discernment of lawlike statements. That a method of sampling is random requires recognition of the equality of certain frequencies, and so is a kind of lawlike knowledge; that is, knowledge of generals. Of course, Peirce didn't get randomness right. No one did or could until the development of recursion theory well into the last century. (Putnam is good here. See his 'Comments on the Lectures' in Peirce [1992, pp. 61 and 68. Lecture Two, pp. 123–142].)

[8] cf. Cohen:'...to accept that p is to have or adopt a policy of deeming, positing, or postulating that p—i.e. of including that proposition or rule among one's premisses for deciding what to do or think in a particular context, whether or not one feels it to be true that p'[Cohen, 1992, p. 4].

> ...whether the word truth has two meanings or not, I certainly do think that *holding for true* is of two kinds; the one is that *practical holding for true* which alone is entitled to the name of Belief, while the other is that acceptance of a proposition which in the intention of pure science remains always provisional [Peirce, 1992, p. 178].
>
> Hence, I hold that what is properly and usually called *belief*, that is, the adoption of a proposition as [a possession for all time]... has no place in science at all. We *believe* the proposition we are ready to act upon. *Full belief* is willingness to act upon the proposition in vital crises, *opinion* is willingness to act upon it in relatively insignificant affairs. But pure science has nothing at all to do with *action*. The proposition it accepts, it merely writes in the list of premises it proposes to use.

And since

> [n]othing is vital for science [and] nothing can be ..., [t]here is ... no proposition at all in science which answers to the conception of belief [Peirce, 1992, p. 178].

Accordingly,

> ... the two masters, theory and practice, you cannot serve [Peirce, 1992, p. 178].

This is a useful distinction. It elucidates the logic of hypothesis-discharge. When an abducer has adequate grounds to discharge H, he has grounds to accept it and to put it to assertive premissory use. But he does not have grounds to believe it, which is precisely what the 'c' in superscript position in the abductive conclusion H^c signals. It is plain from Peirce's remarks that

Proposition 4.1 (Peirce and ignorance-preservation) *In his analysis of abduction, Peirce subscribes to the ignorance-condition.*

Peirce's conception of logic also anticipates that of Quine:

> My proposition is that logic, in the strict sense of the term, has nothing to do with how you think .... Logic in the narrower sense is that science which concerns itself primarily with distinguishing reasonings into good and bad reasonings, and with distinguishing probable reasonings into strong and weak reasonings. Secondarily, logic concerns itself with all that it must study in order to draw those distinctions about reasoning, and with nothing else [Peirce, 1992, p. 143].

About these things, 'it is plain, that the question of whether any deductive argument, be it necessary or probable, is sound is simply a question of the mathematical relation between ... one hypothesis and ... another hypothesis' [Peirce, 1992, p. 144].

Concerned as it is with the presence or absence of that mathematical relation between propositions,

> [i]t is true that propositions must be expressed somehow; and for this reason formal logic, in order to disentangle itself completely from linguistic, or psychical, considerations, invents an artificial language of its own, of perfectly regular formation, and declines to consider any proposition under any other form of statement than in that artificial language. ...As for the business of translating from ordinary speech into precise forms, ...that is a matter of applied logic if you will ... [Peirce, 1992, pp. 144–145].

As Peirce sees it, applied logic stands to logic much as white chocolate stands to chocolate; for it is 'the logical part of the science of the particular language in which the expressions analyzed [or translated] occur' [Peirce, 1992, p. 145]. Peirce, like Mill, is somewhat ambivalent about the primacy of induction. 'As for ... Induction and Retroduction, I have shown that they are nothing but apogogical transformations of deduction and by that method the question of the value of any such reasoning is at once reduced to the question of the accuracy of Deduction' [Peirce, 1992, p. 168]. On the other hand, the 'marvellous self-correcting property of Reason, which Hegel made so much of, belongs to every sort of science, although it appears as essential, intrinsic, and inevitable only in the highest type of reasoning which is induction' [Peirce, 1992, p. 145].

Of Peirce's transformations we have the space to say only that he regarded induction, probabilistic inference and abduction as distinct forms of reasoning of which Aristotle's first three syllogistic figures were limiting cases. (See Lecture Two [Peirce, 1992, pp. 131–141].) The distinctness claim is challenged, however, by the high degree of entanglement, in actual examples of reasoning, of deductive, inductive and abductive patterns of reasoning, never mind what their status is considered as ideal types of reasoning. (See chapters by [Crombie, 1997] and [Kapitan, 1997] in Houser *et al.* [1997].) Kapitan goes so far as to attribute to Peirce an *Autonomy Thesis*, according to which abduction is, or subsumes, reasoning that is distinct from and irreducible to either deduction or induction (cf. [Peirce, 1992, p. 5.146]). That is to say, it is a part of empirical linguistics, the natural science of natural languages.

Logic, then, has nothing to do with how we think. Still less has it to do with how we think about vital affairs.

> Once you become inflated with [what good deducers, i.e. mathematicians, are up to], vital importance seems to be a very low kind of importance indeed.
>
> But such ideas [i.e. the procedures of deductive thinkers] are only suitable to regulate another life than this. Here we are in this workaday world, little creatures, mere cells in a social organism itself a poor little thing enough, and we must look to see what little and definite task circumstances have set before our little strength to do. The performance of that task will require us to draw upon all our powers, reason included. And in the doing of it we should chiefly depend not upon that department of the soul which is most superficial and fallible — I mean our reason — but upon that department that is deep and sure — which is *instinct* [Peirce, 1992, p. 121 (emphasis added)].

Concerning the claim of abductive innateness, Thagard remarks that it is not at all clear that we have any special faculty for guessing right when it comes, say, to theoretical physics [Thagard, 1992, p. 71]. This is a view to which the present authors are somewhat sympathetic (but see chapter 7 below.) Thagard goes on to remark that in "current subatomic physics, many theorists are investigating the properties of spaces with ten or more dimensions, and it is hard to see how their speculations might be at all constrained by biologically evolved preferences for certain kinds of hypotheses ... [T]here is no current reason to adopt Peirce's view that abduction to scientific hypotheses is innately constrained" [Thagard, 1992, p. 71].

Peirce thinks that all forms of reasoning, even deduction, require observation. What is observation? It is

> the enforced element in the history of our lives. It is that which we are constrained to be conscious of by an occult force residing in an object which we contemplate. The act of observation is the deliberate yielding of ourselves to that *force majeure*, an early surrender at discretion, due to our forseeing that we must, whatever we do, be borne down by that power, at last. Now the surrender which we make in Retroduction, is a surrender to the Insistence of an Idea. The hypothesis, as The Frenchman says, *c'est plus fort que moi*.[9]

This is a view that has attracted some experimental support. (See, e.g., [Gregory, 1970; Rock, 1983; Churchland, 1989; Churchland, 1995].)

The passage is striking and at seeming odds with Peirce's insistence that is not useful in ordinary affairs and that it has nothing to do with belief or probability.

[9] Peirce also thinks that observation is itself abductive. Perceptual claims are 'critical abductions, an extreme case of abductive inference' [Peirce, 1992, p. 5.181]. *Cf.* [Thagard, 1992, p. 170].

(After all, for Peirce the most common fallacy of abduction is Bayesianism, that is, to 'choose the most probable hypothesis' [Putnam, 1992, p. 78]. See the section on *Bayesian Inference* below.) But belief is like observation. It presses in on us and demands our surrender. And what is so insistent about the Insistence of an Idea if it does not somehow call for and eventuate in belief?

We begin to see the irreducibly practical aspect of abduction (in Peirce's own sense of "practical"), contrary to what Peirce wants to think. It shows itself in two ways, in the Insistence of an Idea and in the inductions that let loose our sentiment, configure our habit and culminate in belief. This last aspect is ascientific, an all too human acquiescence in a proposition by which our vital affairs might be guided. But the first is purely scientific, an intrinsic part of abduction itself, and a seeming contradiction of Peirce's settled and repeated views. Let us see.

In its fullest sense, abduction is two things for Peirce:

(a) surrender to an Idea, to a *force majeure*; and

(b) a method of testing its consequences.

In surrendering to an idea, say the existence of quarks, or the promise of quantum electrodynamics, it is Peirce's view that we do not come to a belief in quarks or to a belief that QED is true, but rather to a belief that quarks are a good thing to suppose and a good bet for testing, or that QED has a good predictive record or at least the promise of such. In this, abduction resembles Peirce's conception of practical reasoning (and observation, too). At the heart of this resemblance is the concept of a 'vital affair'. It seems to be Peirce's view that the vitality of an affair is *entirely* a matter of the manner in which we conduct its enquiry. There may be matters that especially conduce to such methods, but there seems to be nothing intrinsic to the methods that preclude them from application to any subject matter.

In the case of abduction, these methods are applied to the important business of making conjectures for subsequent test. Essential to this is the conviction that a hypothesis H is a good enough conjecture to warrant our testing it. That is a vital affair for scientist and laymen alike. What this shows is that belief is indispensable for abduction, but it does not expose an inconsistency in Peirce's account. For here the object of belief is not hypothesis H, but rather is some proposition of which H is the *subject*, viz., the proposition that H is a good conjectural bet; in other words, that $C(H)$.

4.1.1 Surprise

That Peirce should have made the factor of surprise a condition on abduction is perhaps itself surprising. It may be that Peirce was excessively restrictive in imposing this condition; but it cannot be denied that often what an abductive agent sets out to deal with an abduction problem, he does so precisely because it does take him by surprise.

How should surprise be understood here? Perhaps we could say that a surprise is an event or a consequence which does unexpected violence to a pre-existent belief. A surprise is doxastically *startling*. ‘A man’, writes Peirce, ‘cannot startle himself, by jumping up with an exclamation of Boo!’

Peirce also sees the postulation of A as ‘originary’ or creative. Abduction ‘is Originary in respect to being the only kind of argument which starts a new idea’ [Peirce, 1931–1958, p. 5.171]. Conjectural creativity comes in degrees, in ascending order as follows:

1. Showing for the first time that some element, however vaguely characterized, is an element that must be recognized as distinct from others.
2. to show that this or that element is not needed.
3. giving distinctness — workable, pragmatic, distinctness — to concepts already recognized
4. constructing a system which brings truth to light
5. illuminative and original criticism of the works of others [Peirce, 1931–1958].

It is possible to discern a less fine-grained taxonomy in which novelty comes in just two degrees, what J.R. Anderson calls “conceptual re-arrangement” and “concept creation”. It is interesting that this distinction is virtually identical to one drawn by Kant in his pre-critical writings and persisted with until his last published work [Kant, 1974]. Seen Kant’s way, the distinction is one between analysis and synthesis. Analysis, says Kant, is the business of making our concepts clear, whereas that of synthesis is that of making clear concepts. Here is Peirce on the former notion: ‘It is true that the different elements of the hypothesis were in our minds before; but it is the idea of putting together what we had never before dreamed of putting together which flashes the new suggestion before our contemplation’ [Peirce, 1931–1958, p. 5.181]. The second kind of novelty Anderson characterizes as the formation of a new concept. It is, as Kruijff says, ‘the creation of an idea that had never been in the mind of the reasoner before’ [1998, p. 15], as witness Planck’s innovation of the quantum or Gell-Mann’s of the quark.

4.1.2 Testability and Economics

It is arguably fundamental to the logic of discovery that abductive agents have the imaginative capacity to hit upon the correct hypothesis, or nearly so, out of very large sets of possibilities. Peirce’s approach to this central issue is itself abductive.

> It is a primary *hypothesis* underlying all abduction that the human mind is akin to the truth in the sense that in a finite number of guesses it will light upon the correct hypothesis [Peirce, 1931–1958, p. 7.220].

Peirce conjectures that the *guessing instinct* is that which explains why beings like us are able with striking frequency to venture hypotheses that are at least approximately correct, while in effect discarding an almost infinite number of unsuitable candidates [Peirce, 1931–1958, p. 5.172].[10]

The heart and soul of abductive reasoning appears to be the reduction of large sets of possibilities to small subsets of candidates for hypothetical endorsement (see chapter 3 above). In selection of a hypothesis, the abductive agent must in some manner solve what we are calling the Cut Down Problem, although, as we said at the end of the previous chapter, this need not (cannot) be done by actually *constructing* a filtration-structure $\langle \mathbb{S}, \mathbb{R}, \mathbb{P}, \{\mathbb{M}\} \rangle$. According to Peirce, the Cut Down Problem is solved by largely economic considerations.

> Proposals for hypotheses inundate us in an overwhelming flood while the process of verification to which each one must be subject before it can count as at all an item even of likely knowledge, is so very costly in time, energy, and money, that Economy here, would override every other consideration even if there were any other serious considerations. *In fact there are no others* [Peirce, 1931–1958, p. 5.602 emphasis added].

Notwithstanding that an inductive agent's conjectural economies are largely a matter of instinct, Peirce supposes that there are three 'leading principles' for the selection of hypotheses [Peirce, 1931–1958, p. 7.220]. One is that the hypothesis should be generated with a view to its *explanatory potential* for some surprising fact or state of affairs [1931–1958, p. 7.220]. Another is that the hypothesis should be *testable* [1931–1958, p. 5.599]. The third such principle is that the abducer's choice of hypothesis be made *economically* [1931–1958, p. 5.197], since

> If he examines all the foolish theories he might imagine, he never will . . . light upon the true one [1931–1958, p. 2.776].

This is important for Peirce. It explains why we don't specify $\mathbb{R}$ by searching $\mathbb{S}$ for the relevant possibilities, why we don't specify $\mathbb{P}$ by searching $\mathbb{R}$ for plausibilities, and so on. Such searches are both uneconomic and ineffective. We note in passing our differences with Peirce on all three scores. We have already explained why we think that testability and explanativeness are not intrinsic to abduction. We will have something to say about economics later on.

[10] *cf.* Kant ". . .the schematism of our understanding . . .is an art concealed in the depths of the human soul" [Kant, 1933, pp. A141–B181].

The economics of abduction is driven in turn by three common factors: the cost of testing [1931–1958, p. 1.120], the intrinsic appeal of the hypothesis, e.g., its simplicity, [1931–1958, p. 5.60 and 6.532], where simplicity seems to be a matter of naturalness [1931–1958, p. 2.740]; and the consequences that a hypothesis might have for future research, especially if the hypothesis proposed were to break down [1931–1958, p. 7.220][11].

There remain two further matters to mention. One is the caution with which Peirce proposes hypotheses should be selected. Peirce's caution is largely a matter of cardinality. In general, it is better to venture fewer hypotheses than more. 'The secret of the business lies in the caution which breaks a hypothesis up into its smallest logical components and only risks one at a time' [1931–1958, p. 7.220]. On the other hand, Peirce favours the use of broad hypotheses over narrow, since it is the business of an abductive explanation to 'embrace the manifold of observed facts in one statement, and other things being equal that the theory best fulfills its function which brings the most facts under a simple formula' [1931–1958, p. 7.410]; see also [1931–1958, p. 5.598–599]. In other words, the abducer is bidden to strive for the fewest conjectures that will cover the largest set of data. (See also [Peng and Reggia, 1990] and [1993]. As Thagard has it, preferred hypotheses explain many facts without co-hypotheses [1993, p. 56].)

It is easy to interpret Peirce's leading principles and the other criteria of economy as instructions for the free use of the abducer, as rules which the abductive agent may follow or break at will. It is difficult however, to reconcile such a view with Peirce's repeated insistence that hypothesis selection is largely instinctual. We shall return to this question. It is an issue which bears in a central way on the question whether a logic of discovery is possible.

In sum, according to Peirce an abduction problem is set off by a surprising event which can't be explained by what the abducer presently knows or accepts. The abducer's task is to find a hypothesis that generates an explanation of this event. Such hypotheses are somehow "selected" from indefinitely large sets of possibilities. In some cases, the abductive process is "originary"; it produces wholly novel conjectures. The Cut Down Problem is handled instinctually, by the abducer's innate flair for thinking of, for picking and being moved by (*force majeure*) the right hypothesis. It is all a matter of guessing, but the guesses have a remarkable track record for fruitfulness, if not strict accuracy. Even so, at the tail-end of abduction, no hypothesis ever scientifically justifies discharge, or the displacement of hypotheticality by categorical utterance. The best that a hypothesis can aspire to is quasi-discharge or (anachronistically) Popperian discharge. Notwithstanding that hypothese are forced upon us by an innate capacity for making the appropriate guesses, there are conditions which favour the exercise of the guessing instinct, each having to do with hypotheses under consideration, so to

[11] See also [Rescher, 1976a; Rescher, 1996] and [Shi, 2001]

speak. The innate guesser will favour hypotheses that have greater rather than less explanatory yield, that are testable, low-cost, few in number, and broad in their reach. Explanatory yield, in turn, turns for Peirce on factors of simplicity and relevance for further research. Quite a lot to be encompassed in the innate capacity for guessing right!

4.1.3 Insight and Trial

The two core ideas of surprise and economic determination have a considerable bearing on the subtitle of this book. The words "insight and trial" give rise to two quite natural expectations on the part of the reader. One is that

1. The authors of *The Reach of Abduction: Insight and Trial* believe that the factors of insight and trial are intrinsic to abduction.
2. The authors of *The Reach of Abduction: Insight and Trial* intend to discuss these two factors.

The second expectation is correct, but the first is not. Here is why.

There can be no doubt that sometimes an AP is triggered by an astonishing event. Equally, it is often the case that a successful explanation of such an event will require the solver to do some highly originary thinking, as Peirce calls it. We can also call it *insight*. In its most intense form, insightful thinking is highly creative and outside-the-box. The phenomenon of outside-the-box thinking presents experimental psychology with a considerablre challenge, and has occasioned some interesting work (e.g., [Blois, 1984; Myers, 2002] and [Weber and Perkins, 1992]).

Insight is not integral to abductive success. This reflects the fact that abductive triggers are not required to be astonishments. Abduction is triggered by the irritation of ignorance. If an agent's ignorance is occasion for his wanting to know whether P, not knowing whether P, here and now, need not be a convulsive deficiency, and, accordingly, need not call for a spectacular redress. "Need not" is something of an understatement. There are large classes of abductive problems for which outside-the-box effects would guarantee failure. So we must say that

Proposition 4.2 (Surprise and insight) *Although some surprises call for insight, an account of abduction should not make this an intrinsic requirement.*

Peirce's other core idea, the economic character of abduction, flows from his conviction that, whatever its other details, abduction always embodies a decision to send a hypothesis to trial. If this is right, Peirce has a plausible case to press for the indispensability to abduction of economic considerations. Our own view is that Peirce is mistaken on both counts. Apart from the fact that some perfectly reasonable abductions don't admit of trial, i.e., are untestable, it is also the case that a decision to send a proposition (or rule, or theory) to trial is independent of both concluding that $C(H)$ and concluding that H^c. A full abduction terminates with the judgement that H^c. H^c expresses the abducer's willingness to put H to premissory work. Subsequently we may have reason to think that H has performed this work fruitfully. Or evidence might suggest otherwise. The question whether H has had a successful or otherwise post-abductive premissory career is a way of asking how H has tested out.

It is quite true that for H, as for other things, "by their fruits shall we know them". But this is not the kind of test that Peirce has in mind. Peirce means *experimental trials*.

Experimental trials cost money. It is not uncommon for less plausible hypotheses to be less costly to test than more plausible hypothes. When the gap between the two plausibilities is wildly disproportional to the cost-differential of their respective trials, a decision to test the less plausible is sometimes reasonable. But since abduction is never a matter of sending a proposition (etc.) to experimental trial, it can hardly be the case that economic considerations are intrinsic to an abduction's reasonableness.

It is possible that our disavowal of the economics-condition on abduction may generate a confusion. If so, confusion will have arisen from our claim in chapter 1 that abductive agents operate in *cognitive* economies. But a cognitive economy is not a *money* economy (although they can overlap). When, as we said, a practical agent operates with scant resources, what he is short of is not money but rather information, or time, or computational capacity. While it is quite time that having some money might alleviate some of these scarcities, it is also true that, even when funds are available for this purpose, practical agents typically don't avail themselves of them, nor is this reticence any kind of standing rebuke to their rationality. Accordingly,

Proposition 4.3 (Abduction and trials) *A decision to send a proposition (etc.) to experimental trial is neither necessary nor sufficient for its abduction.*

4.2 Rationality and Diminished Epistemic Virtue

It is possible that readers are discouraged by what we might call the "false appearance" of certain abductive triggers. We have already observed that it is often the case that what is initially taken as an abductive trigger is a state of affairs that triggers not the conjecture of an hypothesis, but rather a premiss-search (concerning which, see chapter 8, below). We have rejoined that, phenomenologically speaking, there is little to differentiate trying to *hypothesize* something that will explain a state of affairs and *remembering* something that will also turn the same trick. It is wholly unsurprising that what starts out as having the feel of an abductive search actually terminates in something remembered, in regard to which the nescience condition holds no sway. So it is worth repeating the suggestion that

♡ **Proposition 4.4 (Economies achieved by memory)** *Partly for reasons of economy, searches for hypotheses begin as reminescential searches for what is already known. If the search is unsuccessful then,* faute de mieux, *conjectures may be considered.*

Accordingly,

Proposition 4.5 (Revising the notion of trigger) *Something is an abductive trigger when a* memory search *induces the agent to realize (usually tacitly) that the attendant cognitive irritation cannot be removed by what he currently knows.*

4.3 Explanationism

As is clear, there is a large class of abduction problems that call for conjectures that would explain some given fact, phenomenon or feature. An account of the abductive structure of such problems and of the solutions that they admit of requires that the concept of explanation have a load-bearing role in. This in turn requires that the idea of explanation be well-understood. Such is a job for the conceptual analysis component of a logic of abduction.

The concept of explanation enters the logic of certain varieties of abduction in two ways, one primary and the other derivative. In its primary occurrence, explanation is one of the varieties of the consequence relation $\rightsquigarrow$, as set out in the abductive schema discussed earlier. As an aid to exposition, we here repeat the most elementary fragment of that schema. An abductive inference always encompasses an inference in the form

1. T
2. $K(H) \rightsquigarrow V$

   $\vdots$

3. $\therefore H^c$

When such an inference is explanationist in character, the sublogic of the consequence relation must provide an explanationist interpretation of $\rightsquigarrow$.

We have already noted the tension between philosophical and AI approaches to the idea of abduction. The dominant philosophical idea, both in epistemology and studies in the foundations of science, is that abduction is inherently explanationist. What this comes down to in particular is that, within the logic of discharge, the degree to which an H is legitimately selected is strictly a matter of its explanatory contribution with respect to some target. In AI approaches, including logic programming and logics influenced by computational considerations more generally, the dominant (though not by any means exclusive) idea is not explanatory but proof-theoretic or algorithmic. On this view, an H is legitimately dischargeable to the extent to which it makes it possible to prove (or compute) from a database a formula not provable (or computable) from it as it is currently structured. This makes it natural to think of AI-abduction in terms of belief-revision theory, of which belief-revision according to explanatory force is only a part. In the preceding chapter, we suggested that it is preferable to have a logic of abduction that absorbs both conceptions, an approach that derives some encouragement from the fact that the AI-conception already subsumes the philosophical conception.

We return to this matter here, not to re-argue the issue, but to point out that the proposed ecumenism receives additional support from conceptions of explanation that smudge the intuitive distinction between AI and explanationist conceptions of abduction. A case in point is the deductive-nomological (D-N) theory of explanation, to which we turn in the following section. As we will see, it is possible to have D-N explanations of a phenomenon without having an iota of explanatory command of it in the intuitive sense.

In its broadest sense, explanationism (a term coined by James Cornman) holds that what justifies nondeductive inference is the contribution it makes, or would make, to the explanatory coherence of the agent's total field of beliefs. Aristotle, Copernicus, Whewell [1840 and 1967], Peirce, Dewey, Quine and Sellars, are explanationists of one stripe or another; but Harman was the first to formulate it as 'the inference to the best explanation', and to argue for it as an indispensable tool of ampliative reasoning [1965]. More recent defences of the principle can be found in Thagard [1992; 1993], Lycan [1988], Josephson and Josephson [1994] and Shrager and Langley [1990]. In contrast to explanationist approaches are probabalisitic accounts. Apart from the inventors of probability theory (Pascal, Bernouilli, and others), a probabilistic approach to induction is evident in the work of Laplace and Jevons. It is also at work epistemologically in Keynes [1921], Carnap [1950], Jeffrey [1983], Levi [1980], Kyburg [1983] and Kaplan [1996]. Lempert [1986] and Cohen [1977] bring a probabilistic perspective to legal reasoning, and concerning how juries reach decisions, Allen [1994] and Pennington and Hastie [1986] combine explanationism and probabalism. (This issue is also discussed below in chapter 5 and in [Thagard, 2000].) In the philosophy of science, there is a continuing rivalry between probabilist theories of scientific reasoning Achinstein [1991], Hesse [1974], Horwich [1982], Howson and Urbach [1993], Maher [1993] and explanationist accounts Eliasmith and Thagard [1997], Lipton [1991], Thagard [1988; 1992] and [Magnani, 2001a].

Inference to the best explanation is perhaps the dominant current expression of explanationism [Harman, 1965]. On this view an inference to the best explanation is one that concludes that something is so (or may be taken to be so) on the grounds that it best explains something else the reasoner also takes to be so.

At best, best-explanation abductions are a proper subclass of abductions. As Hintikka notes,

> Even so, the discovery of the general theory of relativity was not an inference to the best explanation of the perihelion movement of the planet Mercury and of the curvature of light rays in a gravitational field during a solar eclipse, even though they were the first two new phenomena explainable by the general theory [Hintikka, 1999a, p. 95].

The special theory also provides an interesting kind of exception to best-explanation models. The goal of the special theory of relativity was to integrate Maxwell's electromagnetic theory with Newton's mechanics. In terms of the abductive schema sketched earlier, K^n is Newton's theory, K^m is Maxwell's and K^e is the special theory. K^e is such that every law of K^n and every law of K^m are covered by some law or laws derivable in K^e and no such derivation is valid in K^n or in K^m or in their extensional union $\bigcup(K^n, K^m)$. This gives an abductive trigger, the target of which T is to specify a set of laws underivable in K^n and in K^m but which cover the laws derivable there. The target was hit with the construction of K^e. But it is stretching things unreasonably to suppose that in achieving this objective, Einstein was explaining either Maxwell's theory or Newton's. (See also [Hintikka, 1999a, p. 96].)

Lycan distinguishes three different grades of inference to the best explanation. He dubs them 'weak', 'sturdy' and 'ferocious'.

> *Weak IBE*: The statement that H, if true, would explain some desired proposition V and *could* provide inferential justification for concluding that H.

On the other hand,

> *Strong IBE*: The sturdy principle of inference to the best explanation satisfies *Weak IBE*, and is itself a justificatorily independent principle. That is, it need not be derived from, or justified in terms of, more basic principles of ampliative reasoning such as probability theory.

Finally,

> *Ferocious IBE*: The ferocious principle of inference to the best explanation satisfies *Weak IBE*, and is the sole justified method for ampliative reasoning.

Van Fraassen demurs from *Weak IBE*; in this he is joined by Nancy Cartwright [1983] and Ian Hacking [1983]. Cornman [1980] and Keith Lehrer [1974] support *Weak IBE*, but part company with *Strong IBE*. Concerning *Ferocious IBE*, Peirce appears to have been, apart from the young Harman, the sole supporter of note; although, as we have seen, Peirce is not always constant in his ferocious inclinations. Harman himself defends the ferocious version in [1965], but has moderated views in his later writings (e.g., [1986]).

In the sections to follow, we pursue an obvious task for any account of explanationist abduction. It is the task of saying what explanation *is*. As might be expected, there are rival and incompatible answers to this question.

4.3.1 The Covering Law Model

In the *covering law* tradition of the 1940s and onwards, something similar to Hume's idea of a causal law survives in the notion of a general scientific law as developed by Frank Ramsey [1931]. On this view, laws are true generalizations that are a 'consequence of those propositions which we would take as axioms if we knew everything and organized it as simply as possible in a deductive system' [Ramsey, 1931]. Mere correlations on the other hand are true generalizations which would not pass the Ramsey-test.

We owe the term 'covering law model of explanation' to William Dray [1957]. Also known as Hempel's Model, the Hempel–Oppenheim Model, the Popper–Hempel Model and the Deductive-Nomological model, it is a view anticipated by John Stuart Mill [1974]. It involves two main factors — general laws and the subsumption of the *explanandum* thereunder.

> An individual fact is said to be explained by pointing out its cause, that is, by stating the law or laws of causation, of which its production is an instance [1974].

Furthermore,

> a law or uniformity in nature is said to be explained when another law or laws are pointed out, of which that law itself is but a case, and from which it could be deduced.

In 1948 Carl Hempel and Paul Oppenheim expanded on this basic idea, schematizing Deductive-Nomological (D-N) explanation as follows:

$$\left.\begin{array}{l} C_1, C_2, \ldots, C_k \\ \text{(statements of initial conditions)} \\ \\ L_1, L_2, \ldots, L_k \\ \text{(general laws or lawlike sentences)} \\ \hline \mathrm{E} \qquad \text{(description of empirical data to be explained).} \end{array}\right.$$

E may be either a particular event, calling for a singular explanation, or a uniformity, calling for an explanation of laws. Explanation is a deduction offered in answer to the question, 'Why E?' The question is answered (and the intended explanation is rendered) when it is shown that E is deducible from general laws L_i and initial conditions C_i. If the deduction is valid it is assumed that E is nomically forseeable from those laws and those conditions. A useful account of the logic of the covering law model is Aliseda-LLera [1997, sec. 4.2] and [Magnani, 2001a, p. 129ff].

In strictly logical terms, a D-N argument is indistinguishable from predictive and postdictive demonstrations. Even so, explanation is contextually and pragmatically distinguishable from the others by the assumption that E is already known to the investigator. A further condition on acceptable D-N explanations is that the *explanans* must be true and showable as such by experiment or observation. The strength of this requirement, together with the requirement of subsuming laws, suffices to distinguish D-N explanations from *reason-giving explanations*. Reason-giving explanations are answers to the question, 'What reasons are there for believing that E?' If E is 'The reals are nondenumerable', it is a reason to believe E that yonder Field Medalist said that Cantor proved this theorem. But short of the laureate's *universal* reliability, this argument fails both the general laws and the deductive consequence conditions of D-N explanation.

If the requirement that the *explanans* be true is dropped and yet the requirement of experimental testability or observability is retained, then an argument in the form of a D-N argument is said to afford a *potential* explanation of E (*cf.* Aliseda-LLera [1997, p. 136]).

The D-N account of the covering law model admits of various extensions. The *explanans* may admit statistical or probabalistic laws, and the consequence relation on the explanans- explanandum pair could be inductive. This produces four different types of covering law explanations: (1) deductive-general; (2) deductive-probabalistic; (3) inductive-general; and (4) inductive-probabalistic [Aliseda-LLera, 1997, p. 121].

It is possible to soften the Why-question in the form 'How possibly E?' It is also possible to produce *corrective* explanations in answer to both hard and soft Why-questions. Here the answer indicates that sentence E is not entirely true. Such explanations occur in approximate explanations in which, for example, Newton's theory implies corrected versions of Galileo's and Kepler's laws.

The Hempel–Oppenheim version of the covering law model is in several respects too weak. It permits auto-explanations such as 'Water is wet because water is wet.' It suffers from monotonic latitude, permitting both, 'Harry's nose is bleeding because Bud punched Harry or Copenhagen is in Denmark', and 'This powder pickles cucumbers because it is salt mined from under the Detroit River'. Also missing from the Hempel-Oppenheim model is the essential directedness from causes to effects. This is shown by Sylvain Bromburger, who pointed out [van Fraassen, 1980] that the length of a flagpole's shadow is explained by the length of the flagpole, not the other way around.

It is extremely doubtful whether D-N explanations are canonical for all science. But there is no debate as to whether they are canonical for all explanations. We take up this claim in the section to follow. But we note in passing that if abduction were dominantly explanationist and if the D-N model were canonical for explanation, then for practical agents abduction would be a rarity, not a commonplace, as would appear to be the case.

There is, however, an even more systemic difficulty. By the ignorance condition, an abductive hypothesis must lack epistemic virtue. This means that it cannot have an epistemic standing that would admit it to the abducer's knowledge-base K. This sits uneasily with the structure of a D-N explanation in which an explanation E is provable from a law L together with some initial conditions C. Let E^* be some unexplained proposition. By the lights of the D-N model, there is no law L^* and initial conditions C^* such that $\{L^*\} \cup C^* \vdash E^*$. Now let P be any proposition such that in conjunction with $\{L^*\} \cup C^*$, the target E^* is in fact deducible. But this is not a D-N explanation of E^*. It is required by the D-N model that the explanans be the conjunction of a law and initial conditions; and by the fact that $\{L^*\}\cup C^* \nvdash E^*$, we have it that P is not L^* and not C^* or a part of C^*. This allows us to mark a distinction between what we might call *logical* and *epistemic* versions of D-N explanation. According to the logical interpretation, E is explained by any L and C that imply it irrespective of whether L or C is known by anyone to be the case. In the epistemic variation, the *deduction* $\{L\}\cup \vdash E$ is an *explanation* only if L and C are objects of knowledge. Either way, it is problematic whether an abductively successful H can be reconciled to this model. In its epistemic variation, H would be required to be an object of knowledge, which violates the ignorance condition on abduction. But now let P be any proposition that is both hypothesized by some would-be abducer and is such that in conjunction with L and C entails E. These two factors are linked, of course. The *reason* that P is conjectured is that — on the present telling of the logic of explanation — in conjunction with L and C it entails E. But if the logical variation obtains, then $\{P\} \cup C \cup L \vdash E$ holds or does not hold irrespective of what is known of the explanans. Clearly, then, a third kind of case needs to be considered, in which the abducer reasons counterfactually as follows:

> C and L hold
>
> E is not deducible from $\{L\} \cup C$
>
> $C^* = C \cup \{P\}$
>
> If the initial conditions in question were not C but rather C^*, then $C^* \cup \{L\}$ would be an explanation of E.

> Therefore, it may be hypothesized that P is indeed an initial condition in relation to that explanation.

Attractive as it may be, this counterfactual line of reasoning doesn't present us with an inference to an explanation. It is not therefore any confirmation of explanationism as we presently have it. What it yields is explanationism in a weaker sense. For if P were an initial condition, it would be the case that P transforms a non-explanation $(\{L\} \cup C \nvdash E)$ into an explanation $(\{\{L\} \cup C\} \cup \{P\} \vdash E)$. Let us call the explanationism of which this is an instance *subjunctive explanationism*. Accordingly,

Proposition 4.6 (Subjunctive explanationism) *If explanation is taken in the D-N sense, then no successful explanationist abduction can embody the D-N notion of explanation unless the explanationism in question is subjunctive.*

This is an interesting development. The D-N model of explanation has an entirely classical logical structure. The explanations it captures also stand or fall independently of the epistemic states or interests of any agent (so the epistemic version is not intrinsic to the model). But when abduction enters the picture, it does so with the requisite structure and the necessary impairments or omissions of the agent's K-set. The requirements necessary for the explanationist hitting of an abductive target T make it impossible for the abducer to produce a D-N explanation of the state of affairs embraced by this target. The best he can do in this regard is produce a subjunctive D-N explanation, putting it loosely. Even so, it is a fateful turn. The very reasons that prevailed in the D-N case also prevail where explanation of any stripe is in the offing, and to the same constraining effect. There are senses of explanation for which the constraints of subjunctivity are redundant. Another way of saying this is that there are conceptions of explanations that are themselves subjunctivist in character, and for which, therefore, the subjunctivizing consequences of abductive employment amount to the transportation of coal to Newcastle. Thus the importance of Proposition 4.6. Explanations that are not subjunctive in their own right must be held to subjunctivist variation in order to play a role in explanationist abduction.

4.3.2 The Rational Model

We have noted that answers to reason-questions, e.g., 'What reasons are there for believing that E?', are not in general satisfactory answers to D-N Why-questions, such as 'What brought E about?' The distinction between the two sorts of questions — reason-questions and cause-questions — lies at the heart of a conception of explanation that re-emerged in the first half of the past century among philosophers of history, and which has flourished ever since. Its earlier proponents included the neo-Kantians and thinkers such as Dilthey, who argued against the scientific model of explanation, insisting that in some areas of enquiry, history for example — and the *Geisteswissenschaften* more generally — explanation is a matter of irreducible factors specific to the enquiry in question. To give an historical explanation is not to deduce the event in question from general laws but rather is a matter of specifying attendant conditions which enable us to understand the event.

This conflict was revived at the mid-twentieth century by R.G. Collingwood [1946] and William Dray [1957], who challenged the applicability of the covering law model to the study of history. The anti-covering law position invests heavily in the putatively irreducible character of the distinction between reasons and causes. It holds that it is possible to know the cause of any event without understanding the event caused by that cause. They hold with equal force that a satisfactory explanation of an event in terms of the reasons behind it is always a contribution to the event's *intelligibility*. There is little chance of the cause-reason distinction lining up so tightly with the distinction between explanations that needn't and explanations that must increase the intelligibility of their *explananda*. However, these two distinctions show better promise of concurrence when reasons are understood in a certain way. As Collingwood saw it, to explain an historical event is to find the reasons for its occurrence, and to do that is to re-enact the circumstances and states of mind of the actual agents whose actions brought about the event in question. The historian gives 'the' reasons for that event when he gives 'their' reasons for it. Dray calls this the *rational model of explanation*. In Collingwood's version of it, a high premium is put on the historian's empathy with his historical subjects. There was not then a satisfactory epistemological theory of empathy. This occasioned stiff criticism of Collingwood's re-enactment theory. In Dray's hands, the rational model retains some Collingwoodian insights, while downplaying express reliance upon empathy. (Rightly; see chapter 8 below.) According to Dray, historical explanations are *typically* better explanations when they conform to the rational model. In that model, the historical investigator seeks to reconstruct the reasons for the historical subject's deliberate actions, in the light of their known, desired or expected consequences, appear justified or defensible to the agent himself, rather than deducing those actions from true or inductively well-supported uniformities.

It is one thing to say that an event is explainable independently of the uniformities it falls under. It is another thing to deny the very existence of such uniformities. Part of the rational modelist's disenchantment with covering law explanantions in history, and the *Geisteswissenschaften* in general, is the tendency to think that the domain of deliberative, purposive human action is not one governed by laws in the sense required by the D-N model. Supporters of the rational model of historical explanation tend to doubt that there are in this sense genuine laws of psychology, sociology, political science, economics and so on. In the nearly fifty years since *Laws and Explanation in History* first appeared, skepticism about laws of human performance has abated considerably — a not surprising development given the inroads made by materialism during those four and a half decades. Another respect in which the Dravian account may strike the present-day reader as over-circumspect has to do with the role of empathy in ascriptions to others. In recent developments, empathy has emerged as a load-bearing factor in Quine's later reflections on translation [Quine, 1995], and even more so has a central place in the important work of Holyoak and Thagard [1995] on analogical reasoning. We return to the issue of empathy in chapter 9 below. A third point we have already touched upon. It is the comparative success of Davidson's assault on the presumed exclusiveness of the reason-cause distinction.

It is evident that the rational model is rooted in the *story* — the most primitive form of explanation-giving. Stories drive the cosmogonic tradition of making sense of events big and small. Cosmogony is one of two forces that shape Western cultural and intellectual sensibility. Cosmogony derives from Jerusalem, whereas the second tradition, *logos*, derives from Athens. These two inheritances competed with one another for standing as the better way of accounting for the world and its significant events. To this day this competition survives not only in the rivalry between the rational and covering law models, but also — and more broadly — in the contrast between narratives which strike us as satisfying and plausible, and accounts which lay a strong claim on truth or high probability, even at the cost of plausibility and — *in extremis* — intelligibility. Philosophers of history also opt for the narrative approach with notable frequency. As they see it, understanding human conduct in general, and past behaviour in particular, is a matter of weaving a coherent narrative about them. On this view, history is a species of the genus story (see [Gallie, 1964] and more recently [Velleman, 2003]). It is a view that attracts its share of objections. Perhaps the most seriously intended criticism is that effective and satisfying narratives about events do not in general coincide with accounts of those same events that are arguably true. Plausibility is one thing, and truth is another. So it is questionable whether history considered as story qualifies as a *bona fide* mode of truth-seeking and truth-fixing enquiry. Philosophers of history have responded to this criticism in two basic ways. On the one hand, Louis Mink [1969] has argued that historical enquiry stands to the duties of truth-fixing enquiry as theory *construction* in science stands to this same objective. Mink's suggestion will not still the disquiet of every critic, needless to say; but it performs the decidedly valuable task of emphasizing the importance of distinguishing in a principled way the theories a scientist constructs beyond the sway of directly observable evidence, and the narratives constructed by the historian — also beyond the sway of directly observable evidence. Mink's implied question for the scientific theorist is 'Are you not also telling stories?'

A second reaction to the criticism that historical narratives make a weak claim on probativity, is that of writers such as Hayden White [1968], who emphasize the close structural fit of historical narratives with fiction, as well as their susceptibility to fictional taxonomies, such as tragedy, comedy, romance and satire. White concedes that the fictive character of historical explanation calls into question its capacity as a reliable representation of the reality of the past.

In recent decades, historians have altered their practice in ways that makes it more difficult to reconcile the narrative model to what historians actually do. These days historians place less emphasis on the role of individuals, and stress instead the importance of social, cultural and economic forces, as witness Fernard Braudel's likening of the doings of kings, generals and religious leaders to mere surface irritations on patterns of events constituted by these more slowly moving impersonal historical forces. Even so, the narrative model is far from defunct, and is not without its significant supporters (e.g., [Ricoeur, 1977]).

It seems rather clear that if explanationist abduction is a commonplace form of reasoning for practical agents, then, since the rational model is consistent with this presumption, that alone is some (abductive) reason to take the rational model as canonical for practical reasoning. It is not, of course, a proof that this is so; but it does give the nod to the rational rather than covering law model of explanation in the domain of practical reasoning. What is more, in its emphasis on the story, the rational model pivots on what is clearly a counterpart to the abductivist's notion of plausible conjecture.

An important difference between the rational and the D-N models is that the rational model displaces the D-N model's pivotal relation of deducibility with a relation of explanatory coherence. It is, on the face of it, a displacement with higher costs than benefits, since it would seem that a good deal more is known about deducibility than about explanatory coherence. Perhaps this is so, but we doubt that the gap between costs and benefits constitutes a decisive advantage for the covering law model. For one thing, we are not wholly in the dark about explanatory coherence [Thagard, 2000] below. Beyond that, although we know a lot about classical deducibility, classical deducibility must, as we have seen, defer to subjunctive deducibility in contexts of D-N explanationist abduction. Although we aren't wholly in the dark either about subjunctive deducibility, our command of it is less sure than that of its classical counterpart.

4.3.3 Teleological Explanation

According to rational model theorists, the DN-model of explanation fails to fit the structure of explanations in disciplines such as history. According to functional or teleological theorists, the D-N model also fails to fit the explanatory structure of various sciences, e.g., biology.

To the best of our knowledge it was John Canfield who first distinguished functional explanation from those that fit the covering law model [1964]. As Canfield sees it, functional explanations do not place a phenomenon or a feature in the ambit of a general law; rather they specify what that phenomenon or feature does that is useful to the organism that possesses it. (It has proved a seminal insight, giving rise to a whole class of scientific explanations called *explanation by specification*), concerning which see Kuipers [2001, ch. 4]. Canfield writes,

> The latter [explanations of the covering law model] attempt to *account* for something's being present or having occurred, by subsuming it under a general law, and by citing appropriate 'antecedent conditions'. Teleological explanations in biology... do no such thing. They merely state what the thing in question does that is useful to the organisms that have it [1964, p. 295].

As with those who distinguish the rational model from the covering law model, philosophers such as Canfield distinguish the functional model from the covering law model on the ground that the covering law model doesn't answer the type of question that a functional explanation does answer. Here is Ruth Millikan on the same point:

> [I]f you want to know why current species members have [trait] T the answer is very simply, because T has the function F [1989, p. 174].

A second reason for separating them from D-N explanations is that

> although teleological explanations have the same logical form as explanations in terms of efficient causes, it would be highly misleading to call them "causal" [Pap, 1962, p. 361]. (See also [Nagel, 1977, pp. 300–301].)

This criticism has brought to the fore a further distinction, which some covering law theorists find attractive, between causal and inferential explanations. This distinction is especially well appreciated by Wesley Salmon [1978], notwithstanding that Salmon [1984] is the classic statement of the causal theory of explanation. As Salmon suggests [1978, pp. 15–20], the inferential conception pivots on a concept of nomic expectability, according to which the *explanandum* is something that would have been expected in virtue of the laws of nature and antecedent conditions. (Thus inferential explanation is the converse of a Peircean abduction trigger, which is a *surprising* phenomenon, i.e., one that would *not* be expected even given the applicable laws of nature and the presently known antecedent conditions.) Causal explanations, on the other hand, are those that disclose the causal mechanisms that produce the phenomenon to be explained.

The inferential-causal pair thus gives us a distinction that invades the question of how functional explanations are supposed to differ from D-N explanations. On the causal approach to explanation, the point of difference would appear to be that with functional explanations attributed functions denote effects rather than causes. Karen Neander is good on this point.

> The general prima facie problem with the teleological explanation is often said to be that they are 'forward-looking'. Teleological explanations explain the means by the ends..., and so the explanans refers to something that is an effect of the explandum, something that is forward in time relative to the thing explained.... Indeed, because teleological explanations seem to refer to effects rather than prior causes, it looks at first sight as though backward causation is invoked.... The prima facie problem gets worse, if that is possible, because many... functional effects are never realized [Neander, 1991, pp. 455–456].

Robert Cummins offers an ingenious solution of this problem [1975, pp. 745 ff.]. He proposes that functional explanations do not, and are not intended to, explain the presence of the phenomenon or feature to which the function is attributed. Rather what functional explanation explains is a capacity of a physical system of which the phenomenon or feature to which a function is ascribed is a part. Thus says Cummins, the covering law model applies to *transition* explanations but not to *property* explanations, whereas functional explanations are a particular kind of property explanation. They are explanations in which the property *explanadum* is a complex capacity and the *explanans* is the specification of less complex capacities in terms of which the *explanadum* can be analyzed.

Although Cummins restricts causal explanations to those that explain transition *explananda*, i.e., where what needs to be explained is how a physical system changed from a given *state* to a successor state, Salmon uses the term more comprehensively, including in its extension capacity-explanations. So terminological care needs to be taken, as witness Larry Wright's approach in which functional explanations are those that attribute functions, but do so compatibly with the causal account [Wright, 1973; Wright, 1976].

Functional explanations subdivide further into three categories: *etiological, survivalist* and *causal*.

Etiological approaches (Millikan [1984, pp. 17–49]; Mitchell [1989]; Brandon [1990, pp. 184–189]; and Neander [1991]) propose that function ascriptions specify effects for which a given feature has been antecedently selected. In a well-known example, hearts have the function of distributing blood if and only if doing so caused them to do well in past natural selection. As its name suggests, the etiological view of functional explanation is a variant of the causal conception.

On the survivalist[12] conception (Canfield [1964], Wimsatt [1972], Ruse [1973], Bigelow and Pargetter [1987], and Horan [1989]), the function of the heart is to propagate the blood because doing so is how hearts facilitate survival and reproduction of organisms that have hearts.

[12] This is our contraction of Wouters' "survival value approach" [Wouters, 1999, p. 9].

4.3.4 The Pluralism of Explanation

On the causal account, to say that the heart has the function of propagating the blood is to say that doing so accounts for the organism's ability to circulate the blood (Cummins [1975], Nagel [1977], Neander [1991] and Amundson and Lauder [1994]).[13]

The inferential and causal conceptions do not exhaust the kinds of functional explanation. A third approach is the *design explanation* conception of functional explanation. In it, the role of explanation is not to make a phenomenon or feature nomically expectable; neither is it to lay bare mechanisms that produce the states, transitions or properties that the explainer wants an account of. On the design explanation approach, it would be an explanation of the hollowness of the heart that this is the way the heart has to be designed if it is to be a blood-pumper. Of the three conceptions, design explanations are comparatively little-discussed in the relevant literature. So we shall tarry with it a bit (and in so doing we follow [Wouters, 1999, pp. 14–15 and ch. 8]).

Design explanations connect the way in which an organism is constructed, how its parts operate and the environmental conditions in which it exists to what is required or useful for survival and reproduction. Design explanations come in two types.

1. Those that undertake to explain why it is useful to an organism to behave in a certain way; e.g., why it is useful for vertebrates to transport oxygen. Explanation of this stripe are two-phased. In phase 1, the explainer specifies a need which is satisfied causally by the behaviour in question. In phase 2, it is explained how this need hooks up with other features of the organism and its environment.

2. The second kind of design explanation shows why it is useful that a given feature or piece of behaviour has a certain character. Here, too, the explanation proceeds in two phases. In phase 1 the feature or behaviour in question is assigned a causal role. In phase 2, it is explained why this causal role, given the other relevant conditions, is better formed in the way that it is performed rather than some other way.

Wouters argues that design explanations are not causal explanations, even though the causal idiom is used in their description. They are not causal because they do not explain how or why a certain state of affairs is *brought about*.

Reflecting on the issues of this and the prior two sections, it is evident that explanation — even explanation in science — is beset by a hefty pluralism. There are two responses to pluralisms of this sort. One is to take the differences at hand

[13] Canfield's theory is non-explanationist. Ruse and Horan are inferentialists and Winmsatt, too, in a somewhat different way. Bigelow and Pargetter hold to the causal account.

as having rivalrous import, and then to seek for reasons to pick a (near-to-) unique winner. A second response is to employ the strategy of ambiguation, in which each difference marks a different (and legitimate) *sense* of explanation. Seen the first way, there is a single most tenable conception of explanation, and the others at best are less tenable than it, if not outright wrong. Seen the second way, the different conceptions are best, if at all, in particular contexts governed by particular objectives, and with particular kinds of access to the requisite cognitive resources. Seen this way, it is relatively easy to see why abduction is so commonplace a practice among practical agents. Even if explanationist abduction is not all there is to abduction, it is a considerable part of it. Its ubiquity is explained by the availability to abducers of a multiplicity of conceptions of explanation, each contextually and resource-sensitive. From the point of view of a theory of abduction that recognizes its commonplaceness, there is a lesson to propose to theorists of explanation. It is to analyze explanation in the spirit of reconciliation fostered by the ambiguation strategy.

4.4 Assessing *IBE*

At the beginning of this book we expressed an interest in constructive empiricism. We said that this was a philosophy of science that afforded smooth access to the logic of abduction. By now the reason for thinking so is easily discerned. The logic of abduction investigates the business of reasoning well in the absence of evidence. Constructive empiricism investigates the phenomenon of good scientific theories that likewise exceed the reach of evidence. On the face of it, inference to the best explanation is no contradiction of such an orientation. But appearances can be deceiving. There are defenders of it galore for whom inference to the best explanation cannot be good abductive reasoning if constructive empiricism is true. Correspondingly, there has arisen a substantial controversy about the tenability of inference to the best explanation. A principal source of this doubt is [van Fraassen, 1980; Van Fraassen, 1989]. Van Fraassen is not alone in his criticisms, but the responses to him alone constitute a sizeable literature ([Putnam, 1975a; French and Ladyman, 1997], among many others). On the face of it, this is bad news for abduction. In the spirit of Peirce, it is widely (and rightly) agreed that inference to the best explanation is a significant part of abduction, and some would say that, so far, it is the best understood part of abduction. If it turned out that such a significant and well-understood part of abduction were somehow defective or illegitimate, it might be thought that abduction as such had been called into question. We are not ourselves so minded. Here is why. What van Fraassen rejects is the claim at when a hypothesis H is the best explanation of a target T, then it is *rationally imperative* to accept H [van Fraassen, 1980, pp. 189 and 142]. (Van Fraassen actually says "rationally required rule".) Lying behind the rational necessity claim is the following line of reasoning. If its explanatory force is ever reason to accept a hypothesis, then having the most explanatory force is even better reason to accept it. Both parties to the present dispute concede that explanatory force is indeed a reason to accept a hypothesis (although van Fraassen's concession is tactically motivated). So it might well be argued that the rational necessity claim is correct; that if H explains a target phenomenon better than any rival then it is rationally imperative to select H over its rivals. If this is right, it is a contribution to the engagement-sublogic, although hardly an earth-shattering one. It can also be noted that it is open to the supporter of this view to propose it in a modified form in which the claim of rational necessity is forwarded defeasibly: accept the best explanation unless there is particular reason to do something else. What, then, could the grounds be for the critic's reservations? There are two possibilities to consider, and each involves a misplacement of the factor of rational necessity. In the one case, it is applied in the interpretation of the conclusion operator of the schema for abduction. In the other it operates at the interface between the sublogic of discharge and the role of hypotheses in abductive settings. Here again let T be an abductive target calling for a consequentialist–explanationist response, for which V is the payoff proposition. Then what we have is in truncated form:

1. $K \not\looparrowright V$
2. $K(H) \looparrowright V$
3. $\therefore C(H)$

4 $\therefore H^c$

In the first case, the rational necessity that operates in the engagement-sublogic is transposed (no doubt inadvertently) to the conclusion-sublogic. As it operates there H is held to follow with a strictness that meets the condition of rational necessity. Whether this means that the conclusion is delivered with strictly deductive force might be questioned. But it is beyond doubt that, whatever the details of such an interpretation of $\therefore$, it is much too strong for abduction in the general case, in which the dominant (and correct) interpretation of $\therefore$ is one in which the move to H is held to be plausible. The strict interpretation *mandates* the move to H. The plausibility interpretation *permits* it. We may say, then, that even though the rational necessity claim seems right for the engagement-sublogic, the result of importing the rational necessity factor to the conclusion-sublogic is a mistake.

A second case to consider also involves misplacement. It is a case in which the factor of rational necessity is transposed from the engagement-sublogic to the discharge-sublogic. The discharge-sublogic attempts to determine when the conclusion H of a piece of abductive reasoning is properly de-hypothesized, that is to say, forwarded in the form H^c. Taking its cue from the prior idea that explanatory force is strongly probative, what the rational necessity claim delivers here is the proposition that drawing H as the conclusion of an abduction is sufficient for its unqualified assertion. As we have already made clear, not only is this not the dominant view — it would not have appealed either to Peirce or Popper — it is also a mistaken. We conclude, therefore, that van Fraassen's (and our own) criticisms of inference to the best explanation turn on an interpretation of what such inferences involve that not only deserves these criticisms but, what is more, is implausible on its face.

Interesting though these reflections may be, we haven't yet dealt with the central issue of explanationist abduction. The fundamental question is whether explanations are probative or evidential. (Again, whether van Fraassen's affirmative nod is for tactical reasons is an issue that takes us beyond the scope of our present discussion). It is not a question that we considered settled by our discussion so far. This leaves us with one final point to make. If explanation is probative, if it is evidence for the truth of its explicanda, then there are two possibilities to consider. Either the evidence conferred by an explanation on an explicans raises it to the level of epistemic virtue exhibited by K, where K alone fails to explain the explicandum, or it does not. If an explanation does confer upon its explicans the same degree or higher of the epistemic virtue possessed by K, then

Proposition 4.7 (Explanation and abduction) *If inference to a best explanation is evidentially clinching, then it is not abductive.*

If the explicans of a successful explanation doesn't thereby rise to the epistemic heights of K, inference to the best explanation remains abductive, but now the debate between constructive empiricists and best-explanation inferentialists becomes harder to take seriously. For if both agree that explanationist abductions fail to hit the epistemic standards evinced by K, how much of a stretch is it for the constructivist to say that theoretical postulates fail to reach the evidential standards appropriate for *observation*? And equally, how plausible is it now to say that constructivism is hostile to realism?[14]

4.5 Characteristicness

For all their considerable differences, covering law, rational, functional-causal and functional-acausal (or design) models of explanation answer to a common idea. In each case, one achieves an explanation of a phenomenon when one sees that it would happen in the circumstances at hand. Suppose that Sarah has lost her temper. Even if this is something that (generically) Sarah would never do, it is explicable that she would do it here because the object of her wrath has vulgarly

[14] There is, however, the case, earlier discussed, of circumstantial conviction in a criminal trial, in which we would seem to have it that explanation is indeed probative. We return to this issue later.

insulted Sarah's adored mother. There is a sense of characteristicness in which Sarah's doing this utterly uncommon thing is not out of character for her in those circumstances.

Even the covering law model could be brought into the ambit of this point if we allowed its embedded notion of law the status of generic propositions rather than that of the brittleness of universally quantified conditional statements. This alone would go along way toward explaining the entrenched practice of scientists not to abandon a law in the face of a single observational disconformity. It would also capture what it is about how things function that explains how they behave or the further states they come to be in. For it is characteristic of things that function in those ways that they behave in those ways. If we are right to see the explicable tied thus to the generic and the characteristic, then we have reason to think that, given the centrality to abduction of the explicable, a like centrality could be claimed for the generic and the characteristic. We pick up this suggestion in subsequent chapters.

The suggestion that the covering law model could be made more user-friendly by the simple expedient of genericizing its laws will not strike everyone as a good idea. There are two particularly telling strikes against it. One is that for certain domains of enquiry universality is literally exhaustive, and thus is mathematically expressible by classical quantification. If its laws were genericized, a theory of this sort would understate what it has the wherewithal to demonstrate. A second reservation turns on the emphasis the theorist is prepared to give the falsifiability condition. If a theory's generalizations are formulated as universally quantified conditionals, then not only does he (and we) know precisely in what falsification consists, but falsification is achieved with absolutely minimal contrary turbulence — a single true negative instance. Neither of these features attaches to the falsification of generic statements.

We do not dispute these points. We have already said repeatedly that the bar of cognitive performance is set in conformity with the type of cognitive agency involved. Theories typically are in the ambit of theoretical (i.e., institutional agency), where exhaustive generality is not only an attainable but also an affordable goal. In disciplines (such as certain of the social and life sciences) in which literal exhaustiveness is not always attainable, and in contexts in which an individual seeks to solve a pressing particular cognitive problem, generic universality is a blessing that it would be foolish to do without. Our point of lines above was simply that in such cases the covering law model can often be adapted in such a way that here, too, explanatoriness pivots on considerations of characteristicness.

4.6 Hanson

In Hanson [1958], an attempt was made to establish a logic of discovery (in our terms, a combined logic of generation and engagement), although later on he came to have second thoughts. In 1958, the 'received view' of scientific theories was that they are physically interpreted deductive systems. Hanson points out that when a scientist is trying to take the theoretical measure of a set of data or a body of evidence, he does not search for a physical interpretation of a structure within which those data are deductive consequences. Rather he seeks for theories which 'provide patterns within which the data appear *intelligible*' [Hanson, 1958, p. 71]. Finding such patterns enables the theorist to explain the phenomena which they subsume.

Fundamental to Hanson's project is pattern recognition. Theories that are constructed in Hanson's every are not discovered by inductive generalizations from the data in question, but rather, as Hanson says, *retroductively*, by inferring plausible hypotheses from 'conceptually organized data' [Suppe, 1977, p. 152]. As it happens, Hanson's whole retroductive approach pivots on the view that data are not pure but rather 'conceptually organized' or 'theory-laden'; so we should pause a while over this issue.

Hanson imagines Tycho Brahe and Kepler up at dawn, awaiting a certain event concerning the Sun. Tycho observes the Sun as rising, which has been the view of common sense to the present day. Kepler sees it as progressively revealed by the downward orbital slippage of the Earth on which he stands. What did they observe? A sunrise, for Tycho, and an Earth-slip for Kepler? Or was there some core sensory experience which they both shared? Hanson's reasoning is that since the latter alternative is false, the former must be true. We won't here rehearse Hanson's full case for the theory-ladenness of observation.[15] It suffices to give the gist of his argument. There is a clear sense in which one can see an object without knowing what it is, without knowing what one is seeing. Hanson is drawn to the view that this is not *real* seeing, that the concept of seeing analytically involves the concept of seeing *as*. The present authors are of the opinion that Hanson needn't be right in this claim in order to be right about a claim that matters considerably for his theory. The claim that is right is that *bare* observations — observations such that the observer does not know what he is seeing — are of no use to science, just as they are wholly inadequate guides to action in practical affairs generally. We receive insufficient instruction from 'Oh look, there's a whatchamacallit' or 'Watch out for the thing I'm presently looking at, whatever it is!' The observations

[15] It is a contentious case. Peirce saw Kepler's reasoning as abductive[Peirce, 1931–1958, pp. 1.71, 2.96]. Mill read it differently, as a straightforward description of relevant facts[Mill, 1959, bk III: ch. II.3]; [Peirce, 1931–1958, pp. 171–174]. Thagard doubts that the heliocentric hypothesis was essential to Kepler's discovery [Thagard, 1992]. This is also a view developed in [Simon *et al.*, 1981] and [Langley *et al.*, 1987].

that matter are those that carry appropriate descriptive signatures. For this to be true, it needn't be the case that the Tycho-Kepler story is strictly true. There is plenty of taxonomic slack with which to characterize a heavenly body as the Sun short of commitment to its status as a planet, or otherwise. But it remains the case that useful observations are observations of objects *under descriptions*, and that it is not uncommon that such descriptions fail to be theoretically neutral. Whether they are or not will depend on the context in which the observations in question have been made. It does not bear on the observation, 'The Sun appeared at 5:10 A.M., whether the Sun is a planet or not, if the context is the question whether it is presently winter in London. In other contexts, the descriptions under which things are seen carry theoretical freight (granted that the dipping Earth example is a trifle forced).[16]

Hanson himself comes close to exposing this slighter view of the theory-ladenness of observation when he emphasizes that terms are theory-laden, and must be so, if they are to have any explanatory function. There may, he concedes, actually be pure sense-datum languages where terms are entirely free of theoretical taint, but no such language can serve the explanatory functions of science or of ordinary life [Hanson, 1958, pp. 59–61]. For our purposes in this book, the greater importance of the theory-laden account of observation is the credence it lends to the project of developing a genuine logic of discovery. Here is Hanson on the point:

> Physical theories provide patterns within which data appear intelligible. They constitute a 'conceptual Gestalt'. A theory is not pieced together from observed phenomena; it is rather what makes it possible to observe phenomena as being of a certain sort, and as related to other phenomena. Theories put phenomena into systems. They are built up in 'reverse' — retroductively. A theory is a cluster of conclusions in search of a premise. From the observed properties of phenomena the physicist reasons his way toward a keystone idea from which the properties are explicable as a matter of course [Hanson, 1958, p. 90].

What, then, will the proferred logic of discovery look like?

Hanson distinguishes

(1) reasons for using and/or accepting a hypothesis H (i.e., engaging H)

from

reasons for thinking of H in the first place (i.e., generating H).

[16] We note the widespread view among perception theorists that perception is always perceiving *as*, and the related view that, since unconceptualized perception is impossible, perception is inherently epistemic. (See, e.g., Hamlyn [1990, ch. 4 and 5].)

The reasons subsumed by (1) are for Hanson reasons for thinking that H is true (rather than, for example, as with Newton and his action-at-a-distance theorem, reasons for *using* H, where there is no question of thinking it true). Reasons subsumed by (2) are reasons for thinking that H is a plausible conjecture. Here, too, a further distinction is required, owing to the afore-noted ambiguity of 'plausible'. In one sense, a plausible conjecture is one whose truth can plausibly be supposed. In another sense, a conjecture H is plausible when it is a plausible candidate for consideration. However this distinction is drawn in fine, it pivots on the essential difference between the plausibility of conjecturing something even when there isn't the slightest reason for judging it true, and conjecturing something precisely because the idea that it is true is a plausible one. In the first sense, it is plausible to make all other family members suspects in a crime of violence against a family member even without a shred of evidence for the guilt of any of those individuals. In the second sense, it might become plausible to concentrate on Uncle Harry as the miscreant, not that there is any direct evidence of his guilt, but rather that, given the various facts of the case, Harry's involvement would not at all be out of character.

Hanson wants to preserve a logical distinction between reasons for thinking that an hypothesis, H, is true and reasons for thinking up H in the first place. But reasons of this latter kind are, for Hanson, always reasons in the form 'H is the right *kind* of hypothesis for the problem at hand, irrespective of the *particular* claims it may succeed in making.' We ourselves join with Thagard in thinking that putting it this way is tantamount to giving up on the idea of a *bona fide* logic of discovery [Thagard, 1993, p. 63]. Reasoning of this kind is "retroductive" reasoning, which Hanson schematizes as follows:

1. Some surprising, astonishing phenomena $p_1, p_2, p_3, \ldots$ are encountered.[17]

2. But $p_1, p_2, p_3, \ldots$ would *not* be surprising were a hypothesis of H's type to obtain. They would follow as a matter of course from something like H and would be *explained* by it.

3. Therefore there is good reason for elaborating a hypothesis of the type of H; for proposing it as a possible hypothesis from whose assumption $p_1, p_2, p_3, \ldots$ might be explained [Hanson, 1961, p. 33].

Hanson's abductive schema hooks up with his theory-laden view of observation in a straightforward way. For if it is true that observations involve a certain conceptual organization of the phenomena in question, that fact alone will preclude various possible kinds of explanation and will encourage, or at least be open to, other

[17] Note the Peircean cast.

kinds of possible explanation. If an observer sees some data as having the conceptual organization that makes Keplerian explanations the right kind to consider, then Tychoean explanations will be of the wrong kind for serious consideration.

While our brief resumé doesn't do full justice to Hanson's abductive project, the same can be said of Hanson's own writings which, while more extensive than the account sketched here, nevertheless present a highly programmatic research programme, with a good deal more promised than actually delivered. In this we agree with Achinstein.

> Hanson [... describes] one such pattern [i.e., actually employed in the sciences], retroduction. This consists in arguing from the existence of observed phenomena that are puzzling to a hypothesis, which if true would explain these phenomena. I do not believe that Hanson describes this type of inference in a way that is free of difficulties, but I do believe there is an explanatory type of inference often used in science and that it can be correctly described. Moreover, contrary to what Hanson sometimes suggests and to what has ... been stated by Gilbert Harman, I believe that there are certain *nonexplanatory*, nondeductive patterns of inference actually employed in the sciences [Achinstein, 1977b, p. 357; (emphasis added)].

Before bringing the present section to an end, it would be well to consider an objection which is levelled against the entire project of discovery abduction. Where, the critic might ask, do these abductive hypotheses come from?

And is it not the case that whatever the precise answer to this question, it brings psychological factors into the picture? No doubt, this is so. When Kepler was in process of thinking up his laws on the basis of Tycho's data, causal factors were at work. Perhaps Kepler would not have been able to make these discoveries had he not had the formal education that he did in fact have or had he had different interests, and so on. It is, even so, a grave error — indeed, a special case of the genetic fallacy — to suppose that if how a theorist came to think up a law is subject to a causal explanation, then it must follow that he could not, in so doing, have been involved in a reasoning process. It remains an open — and good — question as to whether there are any systematic connections between the causal conditions under which reasoning is produced and the logical conditions in virtue of which it counts as correct or reasonable. But there appears to be little reason for thinking that the logic of hypothesis formation cannot proceed apace without exposing the etiology of the reasoning process that is in question. (See also Achinstein [1977a, p. 364].)

Of course, thinking things up is often a matter of intuition (although saying it just this way doesn't offer much of an explanation). Still, no one doubts that having the right intuition can be crucial for the theorist's larger purpose. Consider, for example, a theory in which two definitions are equivalent, but only one generalizes.

There was a point at which the physicist could have chosen either $F = m\frac{dv}{dt}$ or $F = \frac{d(mv)}{dt}$. If his intuitions were 'reliable', he would have opted for the second, since it is the second that generalizes to relativity.

Hanson, the promissory-note abductive logician, is also an explanationist. Deductive-nomological explanations are fine as far as they go; they may even be necessary conditions on explanations that do the full job. But D-N explanations do not as such serve in the cause of 'rationally comprehending the 'go' of things ...' [Hanson, 1958, p. 45].

> What can be wrong with our seeking examples of scientific theory which are capable both of explaining *à la* Hempel and of providing understanding and illumination of the nature of the phenomena in question? Even if distinguishable, the two are genuinely worthwhile objectives for scientific enquiry; they are wholly compatible. And, it may be noted, the second is unattainable without the first. So although Hempel's [*D-N*] account of scientific explanations may not be sufficient, it seems to be necessary [Hanson, 1958, p. 45].[18]

What is more — and note well:

> Ontological insight, unstructured by quantitatively precise argument and analysis is mere speculation at best, and navel-contemplatory twaddle at worst [Hanson, 1958, p. 45].

Even in these later posthumous writings (Hanson was killed in a flying accident in 1967, at age 43), there is nothing in the way of 'quantitatively precise argument and analysis' about what it is to make sense of something, to figure it out, or make it comprehensible. The account may not be mere speculation, but it is decidedly promissory.

We see, then, Hanson's approach to abduction implicitly recognizes the distinction between the generation and engagement of hypotheses. As with Peirce, the trigger is an astonishing event or state of affairs that can't be explained with available resources. T is the target, whose attainment repairs the astonishment. Hypotheses are generated when they are seen to be of a type that would explain T. Hanson does not explain how hypotheses of that type are recognized (so strictly the generational sublogic is idle), but it is clear from context that he sees analogy at work here. Crudely, phenomena like the trigger are explained by hypotheses of such-and-such type; therefore, an hypothesis of *like type* is a good bet to explain T. This is all but what we call engagement (Hanson calls it *elaboration*). Once elaborated, it is subjected to a discharge. Hanson retains Peirce's idea that triggers must

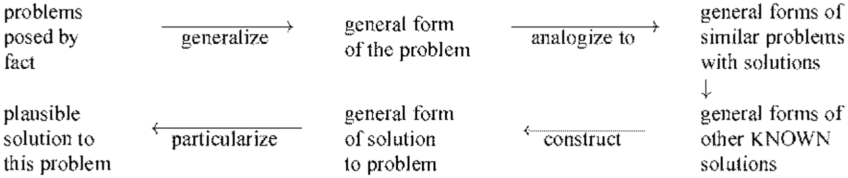
[18] There is a slip here. If a satisfactory Hansonian explanation implies a satisfactory D-N explanation, then it cannot be true, as Hanson strenuously avers, that explanations are sometimes non-predictive.

surprise or astonish, but he doesn't accept Peirce's faith in the abducer's innate capacity for guessing right. Peirce saw perception as the limit of abduction. For Hanson, too, observations are essentially bound up with abduction, but differently. Observations always involve conceptual organization. Thus how a phenomenon is perceived shapes what it is about it that may call out for explanation. Central to this approach is the idea that abductive inference is inference from considerations of like-typeness. We read this as analogical inference, to which we shall return in chapter 8.

4.7 Darden

Hanson's favourable disposition to a logic of discovery has spurred much follow-up work among philosophers of science. A notable exception is Darden [1974]. Central to her approach is the role of analogy.

Darden proposes 'the following general schema for a pattern of reasoning in hypothesis-construction:


problems posed by fact
generalize
general form of the problem
analogize to
general forms of similar problems with solutions
general forms of other KNOWN solutions
construct
general form of solution to problem
particularize
plausible solution to this problem

The pattern of reasoning is used for each of the facts in turn. The use of the same explanatory factor in each one results in a hypothesis made up of a set of postulates all of which involve the same explanatory factors' [Darden, 1976, p. 142].

Darden's approach to the logic of discovery emphasizes the use of hypotheses for the purpose of achieving explanationist abduction, but there is no reason in principle why, if the scheme works for explanationist abductions, it can't be generalized (in its own spirit, so to speak) so as to apply to other forms of abductions, such as those that turn on the predictive yield of the hypothesis. We meet here for the first time a sketch of abduction in which analogy plays a central role. In chapter 8 we investigate this idea further.

4.8 Fodor

In a widely discussed essay Fodor [1981], it is proposed that abduction is intrinsic to conceptualization. The work in question is lengthy and highly nuanced, but its main features can briefly be set out, beginning with the following question.

"Everyone agrees that human agents conceptualize their experience, but where do concepts come from?" Fodor's answer is that all concepts are innate or are computationally derivable from those that are. Derived concepts are constructed; and constructing a concept is like constructing a theory with which to explain certain facts. Such theories need to be cognitively prior to the data that are to be explained in their terms. The same is true of constructed concepts. They are put together out of existing resources for the purpose of having some application to experience.

Some will not see it as plausible to think of highly technical and original concepts as already 'there", so to speak (e.g., quark; *cf.* Thagard [1988, p. 71]). But Fodor can plausibly plead that the critic has confused lexical unfamiliarity with conceptual novelty.

It is arguable that Fodor's is the most extreme form of nativism since Plato's famous example in the *Meno*, in which Socrates elicits from the slave-boy the Pythagorean Theorem. (See here [Magnani, 2001a, pp. 3–8]) But it must be said that Plato is a nativist about knowledge, whereas Fodor's nativism extends to concepts (and, if having a concept of Q is believing what it is to be a Q-thing, to such beliefs as well).

For our part, we grant that even Fodor's nativism is counterintuitive, but we are not prepared to infer from this its indefensibility. For our purposes here, it isn't necessary to settle the nativist-empiricist controversy about concept acquisition. It suffices to note the sheer extent to which on Fodor's account, the human agent is an abductive animal.

4.9 Adaptive Explanationism

In chapter 3, we touched briefly on the so-called logic-based approach to abduction, pointing out that it exhibits — in the distribution of its modal operators — a kind of conformity with the ignorance condition on abduction. In all its variations, the adaptive approach is explanationist. We briefly illustrate with reference to how a modal extension of the adaptive system $CP1$ models the discovery of Uranus [Meheus *et al.*, forthcoming]. In a widely cited passage, Thomas Kuhn describes the discovery as follows. [Kuhn, 1977, pp. 171–172]:

> On the night of 13 March 1781, the astronomer William Herschel made the following entry in his journal: "In the quartile near Zeta Tauri ... is a curious either nebulous star or perhaps a comet." ... Between 1690 and Herschel's observation in 1781 the same object had been seen and recorded at least seventeen times by men who took it to be a star. Herschel differed from them only in supposing that, because in his telescope it appeared especially large, it might actu-

> ally be a *comet*! Two additional observations on 17 and 19 March confirmed that suspicion by showing that the object he had observed moved among the stars. As a result, astronomers throughout Europe were informed of the discovery, and the mathematicians among them began to compute the new comet's orbit. Only several months later did the astronomer Lexall suggest that the object observed by Herschel might be a planet. And only when additional computation, using a planet's rather than an comet's orbit, proved reconcilable with observation was that suggestion generally accepted.

Following [Meheus *et al.*, forthcoming], it is possible to model this reasoning in their system $CP1$. In the interests of brevity, we will reproduce Herschel's original inference, using the following notation. Let a be an object, O the occurrence predicate, r_i and t_i places and times, L a predicate for largeness and M for movement, T_i trajectories, S, C and P predicates designating the properties, respectively, of being a nebulous star, a comet and a planet. Four modal operators are deployed, each producing modal formulas of first degree. $\Box_1$ applies to observation sentences and $\Box_2$ to background sentences. Likewise, $\Diamond_1$ applies to derivations from observation sentences and $\Diamond_2$ to derivations from background sentences.

According to Kuhn's account, Herschel's reasoning encompasses four premisses:

1. a appeared large
2. a occurred at a certain time t_o and place r_o.
3. Nebulous stars appear large
4. Comets appear large

From these premisses he abduced that a "is a curious either nebulous star or perhaps a comet". This is modelled as follows.

1.	$\Box_1 La$	premiss
2.	$\Box_2 Oar_o t_o$	premiss
3.	$\Box_2 \forall x(Sx \supset Lx)$	premiss
4.	$\Box_2 \forall x(Cs \supset x)$	premiss
5.	$\Diamond_2 Sa$	from (1) and (3)
6.	$\Diamond_2 Ca$	from (1) and (4)

In this fragment of the reconstruction, $\Box_1 La$ is the explicandum. It reports an observation for which an explanation is sought. Intuitively, $\Box_2 Oar_o t_o$ likewise records an observation. But since the observation it reports is not the object of an explanation, the system registers it as background information. What our present fragment does not encompass is the dynamical nature of the total discovery. The

$CP1$ reconstruction of this is rather impressive. But for the purposes at hand, all that we need consider is the original fragment.

As mentioned in chapter 3, the $CP1$ reconstruction marks a distinction between the force of what the premisses assert and what the conclusion concludes. In the GW-schema, the premisses are forwarded assertively and the conclusion is subject to two interpretations. In one, the conclusion $C(H)$ is also forwarded assertively, but what it asserts is not that H is the case, but rather that H is worthy of conjecture. In the other the 'C' in $C(H)$ is a modal operator having the force of a deontic assertion: "It may be conjectured that H". In both cases the CPi approach and the GW-schema, there is structural recognition of the ignorance condition — the conclusion is modally ($CP1$) or epistemically (GW) weaker than the premisses.

Even so, there are differences. In $CP1$ the conjectural character of abductive conclusions and the tentative character of abductive inference is recognized in only two ways. One is the slighter modal character of abductive conclusions. The other is their suspectibility to subsequent reconsideration in the dynamical context of $CP1$-reasoning. In the GW-approach, a third factor of provisionality is recognized. The conclusional operator of abductive inference is construed as plausible and presumptive conclusionality. This leaves an interesting question. Is the plausibilisitic (and presumptive) character of the abductive conclusion operator $\therefore$ adequately catered for by representing it as deductive consequence constrained by the standard AKM constraints, viz., where V is the payoff sentence, $K(H) \vdash V, K \nvdash H, T \nvdash H, V \nvdash H$, and H is minimal?

By our lights, the answer is clearly not. Right or wrong, it is not something to go to war over. $CP1$ follows established practice from Aristotle onwards. If what's wanted is a softer notion than strict deductive conclusionality, then apply the right softening constraints. Taken in this standard way, soft conclusionality is a species of strict conclusionality. It is an attenuation of strict conclusionality. This is a strategy to be applauded, as far as it goes. In the present case, our own view is that it doesn't soften strict conclusionality enough.

A further point of difference is $CP1$'s representation of background generalities. There they are taken as universally quantified material conditionals. As such, they cannot abide additional premisses in the form of negative instances. we see this is as over-strict. The generalities on which abductions very often turn have the character of generic statements or, at times, normalic claims. In neither case does admittance of a negative instance necessitate the disturbance of their premissory roles. So, notwithstanding its considerable virtues, we find the $CP1$ reconstruction to be a bit stiff at the joints. We return to the issue of inconsistency in chapter 7 below.

4.10 Non-abductive Conjecture

In the standard inference-to-an-explanation model, the core of abduction is reflected in the schema

1. $K(H) \looparrowright V$
2. $\therefore H^c$

Attaching to such inferences is a standard epistemic pattern. The premisses are taken to be known to the abducer, but the conclusion is not. Accordingly, the premisses are asserted categorically (albeit with the conditional premiss (1) in subjunctive form) and the conclusion is not asserted, but is forwarded as a conjecture. It is easy to see that the inference is wrought by dropping what is not known, i.e., the hypothesis H, into the antecedent of premisses (1), whereupon (1) itself is taken as known. Another way of saying this is that in all explanationist abductions that conform to the present schema, a *relatum* is unknown, whereas a *relationship* (between it and its other relatum) is known. In marking this contrast between what is known and unknown in an inference to-an-explanation abduction, we intend no particular set of epistemological standards. It suffices that what is known and unknown in an abducer's inference to an explanation is relative to the epistemological standards, whatever they are in detail, to which his K-set conforms.

We have already seen that an abductive context precludes inference to an explanation, when explanation is taken in a "pure" D-N rather than subjunctive sense. But more radical preclusions flow from the variants of explanation we have loosely classified as teleological, as well as a wide range of explanations in the manner of the rational model. If we stay with our example, of the etiological, survivalist, causal and design variations of teleological explanation give rise to a quite different pattern of conjecture, as follows

1. E_1
2. E_2
3 $\therefore E_1 \looparrowright E_2$

in which factors E_1 and E_2 are antecedently known, and the explanation itself conjectured. In the classical model, there is a hypothetical element embedded in the premisses which the inference detaches. In the pattern presently in view, nothing is conjectured but the explanatory link in question. In the classical model, the reason for the conjecture is the (subjunctive) truth of the relational premisses. In the teleological model, the relational proposition is conjectured, but it can hardly be said that the reason for the conjecture is the mere truth of its relata. In fact, it is far from clear whether the three steps of our teleological schema even represent an *inference*.

Much the same can be said of rational explanations of the Dravian sort. For large ranges of cases, these too are pieces of conjecture in which one of a pair of knowns is conjectured as explaining the other.

Accordingly, we have it that

Proposition 4.8 (Non-abductive conjectures) *Conjectural or hypothetical reasoning of the explanationist sort does not exhaust the class of explanationist abductive inferences.*

Corollary 4.8(a) *Whereas inference-to-an explanation in the manner of Harman is intrinsically abductive, explanationist conjecture (teleological or rational) is not.*

Chapter 5

Non-Plausibilistic Abduction

> Prove all things: hold fast that which is good.
>
> *Thessalonians* 5, 21

> Today I have made a discovery as important as that of Newton.
>
> Max Planck, in conversation with his son, 1900.

5.1 Introductory remark

The preceding chapter explored the structure of explanationist abduction. Explanation is an ambiguous and context-sensitive concept. It might be expected therefore that there exist resolutions of abduction problems in which the target is delivered non-explanatively and yet this very fact explains or helps explain some collateral feature of it. Accordingly we are content to offer the explanationist-non-explanationist contrast for what it is, namely, as a loose and contextually flexible distinction. A further distinction, also of great salience to abductive logic, is that between abductions that advance propositionally plausible hypotheses and those that advance propositionally implausible hypotheses.

But we should say in passing that we intend no particular concurrence between the explanationist/non-explanationist distinction and the plausiblist/implausiblist distinction. We allow that some understandings of the concept of explanation that recognize the possibility of *implausible explanatory hypotheses*. Of course, we are speaking here and throughout the chapter of propositional plausibility and implausibility, unless otherwise noted.

5.2 Newton

There is an error of substantial durability as to what Newton meant by his celebrated dismissal of hypotheses. The error is that when Newton said that he deigned not feign hypotheses, he was rejecting the hypothetical method in physics and that he would have no truck with inferences to the best explanation. Newton's dismissal is not a general indictment, but rather a response to a particular case. In Newton's theory of gravitation, the gravitational force acts instantaneously over arbitrary distances. Newton's own view was that this was a conceptual impossibility, and he hinged his acceptance of it on the extraordinary accuracy of the relevant equations. '*Hypotheses non fingo*' is a remark which Newton directed to the particular proposition that gravitation acts instantaneously over arbitrary distances. By this he meant that he regarded the action-at-a-distance claim as inexplicable. (It is, of course, widely believed that Einstein's field theory eliminates the source of the inexplicability.)[1]

Although '*hypotheses non fingo*' does not stand as general indictment of the hypothetical method, there is an air of general disapproval evident in Newton's writings, well-summarized by Duhem:

> It was this . . . that Newton had in mind when in the 'General Scholium' which crowns his *Principia*, he rejected so vigorously as outside natural philosophy any hypothesis that induction did not extract from experiment; when he asserted that in a sound physics every proposition should be drawn from phenomena and generalized by induction [Duhem, 1904–1905, pp. 190–191].

Here is the relevant part of the General Scholium:[2]

> But hitherto I have not been able to discover the cause of those properties of gravity from phenomena, and I feign no hypotheses; for whatever is not deduced [sic] from phenomena is to be called an hypothesis; and hypotheses, whether metaphysical or physical, whether of

[1] Tom van Flandern denies this claim in van Flandern [1999]. He argues that there are pulsars whose observations entail that the speed of the gravitational influence is not less than 2×10^{10} times the speed of light. This striking development, which exceeds the scope of the present book, is well-discussed in [Peacock, 2001].

[2] Newton was an inductivist in the spirit of Bacon. Bacon condemned speculation in the absence of data, but he was also aware that a given set of data can be made to align with more than one explanation. Bacon rejected the method of hypothesis — which he called *Anticipation of Nature* — when it was rashly or prematurely resorted to. Bacon also thought that 'there are still laid up in the womb of nature many secrets of excellent use, having no affinity or parallelism with anything that is now known, but lying entirely out of the best of the imagination, which have not yet been found out' [Bacon, 1905, p. 292]. But this is nowhere close to an outright condemnation of the scientific imagination. The debate between inductivists and conjecturalists rages to this day. Consider, for example, the sometimes heated exchanges between biologists (e.g., [Rose and Rose, 2000]) and evolutionary psychologists (e.g., [Miller, 2000]).

> occult qualities or mechanical, have no place in experimental philosophy [Newton, 1713, p. 546]. [3]

That the *Principia* should have been formatted axiomatically bears on our issue in an interesting way. Newton inclined towards the generally current view of axiomatics in supposing that axioms afforded a *kind* of explanation of the theorems that are derivable from them, that the demonstrative closure of a set of scientific axioms is suffused with the explanatory force that originates in the axioms and is preserved under and transmitted by the theory's deductive apparatus. For Newton, this kind of explanatory force is axiomatically secure, a view held in contrast to abductivists in the manner of Laplace [1951], for example, for whom a significant part of a science's explanatory force must come from hypotheses that resist efforts at refutation.

The action-at-a-distance theorem proved a methodological embarrassment for Newton. If the axioms are considered as having explanatory force, and if explanatory force is preserved by a theory's deductive apparatus, then it ought not be the case that *anything* should turn out to be inexplicable in the deductive closure of the first principles of gravitation theory. Given the unintelligibility of the action-at-a-distance theorem, then by Newton's own lights either axioms are not explanatory as such, or the deductive apparatus is not explanation-preserving.

The hypothetical method bears on what the Port Royal logicians would call the *method of discovery*.

> Hence there are two kinds of methods, one for discovering the truth, which is known as *analysis*, or the *method of resolution*, and which can also be called the *method of discovery*. The other is for making truth understood by others when it is found. This is known as *synthesis*, or the *method of composition* and also can be called the *method of instruction* [Arnauld and Nicole, 1996, p. 232].[4]

Arnauld and Nicole, and the other Port Royal logicians, were of the view that analysis always precedes synthesis. They pressed this point in the context of an unsympathetic discussion of syllogistic logic, but it is clear that they see their remarks as having a more general application. They assert that, in as much as the

[3] "Newtonian gravitation burst upon the scene like a bombshell. Newton's supporters simply stonewalled. Roger Coates explicitly denied there was a problem, arguing (in his preface to the second edition of Newton's *Principia*) that nature was *generally* unintelligible, so that the unintelligibility of forces acting without contact was nothing specifically worrisome. However unpalatable Cote's position may seem as a precept for science, ... there is something to be said for it — not, to be sure, as science but as metascience, for we cannot hold the science of tomorrow bound to the standards of intelligibility espoused by the science of today." [Rescher, 1996, p. 75]

[4] These notions derive for ancient Greek geometry. Synthesis is deduction, or a mode of reasoning from causes to effects. Analysis reasons backwards from theorems to axioms, or from effects to causes, and hence is a kind of hypothetical reasoning. See here[Hintikka and Remes, 1974].

challenges of real life are much more to *discover* what things are true or probable than to demonstrate them, it follows that the rules of syllogistic, which are rules of demonstration, can have only a limited utility, and that they will soon be forgotten by young student-logicians because of their general inapplicability to the main problems of life. We see here a considerable debt to Descartes. Not only do Arnauld and Nicole accept the Cartesian distinction between the logic of discovery and the logic of demonstration, they are shrewd to notice (what others don't) that the logic of discovery resists — even if it does not outright preclude — detailed articulation, which is a challenge for the would-be logician of hypothesis-generation.

> That is what may be said in a general way about analysis which consists more in judgement and mental skills than in particular rules [Arnauld and Nicole, 1996, p. 238].

Notwithstanding his Baconian allegiances, Newton was not uniformly hostile to the use of analysis in science. He writes,

> By this way of analysis we may proceed from compounds to ingredients, and from motions to the forces producing them; and in general, from effects to causes, and from particular causes to more general ones, . . . and the synthesis consists in assuming the causes discovered, and established on principles, and by them explaining the phenomena proceeding from them, and proving the explanations [Newton, 1713, p. 380ff.].

The Port Royal logic appeared in 1662 and was widely noted, not only in France; it also had a considerable impact upon Locke. Newton's *Principia* was published in 1687, by which time the method of analysis was sympathetically received throughout Europe, and its influence was discernible in various developments in the new sciences, if not in the formal treatment of the *Principia*. Hypothesis formation is a substantial part of the method of analysis; so perhaps it is somewhat surprising not to find it in Newton. But as the Port Royal logicians insisted, the method of analysis, abduction included, involves the exercise of the scientific imagination; hence is more a matter of 'judgement and mental skills' than of 'particular rules'. So neither is it surprising to find little in the methodology of the new sciences ensuing from Bacon in the way of a theoretical development of the scientific imagination

It is not uncommon for philosophers to speak of the contribution made by the hypothesis of action-at-a-distance as one of explaining otherwise unexplainable observational data. This is a matter calling for some care. Like numerous instances of D-N explanation, Newtonian explanations need convey no elucidation of their explicanda. They need confer no jot of further intelligibility to them. The action-at-a-distance equation serves Newton's theory in a wholly instrumental sense. It allows the gravitational theory to predict observations that it would not otherwise be able to predict. Like those various cases of D-N explanation, Newtonian explanations take liberties with the common sense understanding of what explanations are. There is reason, therefore, to think of them as associated with that class of abductions which we have classified as generically proof-theoretic. All the same, it is a well-entrenched habit, especially among philosophers, to speak of these things as explanations. Here, too, we have no interest in merely semantic wrangles. Suffice it to say that, Newtonian explanations lay rightful claim to recognition in a chapter on *non-plausible* accounts of abduction.

In what is perhaps its most basic sense, conjecture is *epistemically agnostic*. In making a conjecture H one hypothesizes that H possesses a level of epistemic virtue which one does not know it to have. But when conjectures are made in the context of abduction, the ignorance condition is in play. It provides that at one point in the process, the abducer has the belief that H is not in K. This is not agnosticism; it is epistemic atheism by the reasoner's own lights. But given what it is to conjecture such an H, it is a fallibilist atheism, in which the abducer takes what he presently thinks he does not know and ventures that if released for further promissory work it might fare well; indeed might even fare well epistemically. More simply put, we may say that in selecting H the abducer expresses the hope that eventually he will be proved mistaken about H's expistemic standing. Like hopes of any kind, the abducer's hopes for H may turn out to be dashed. This leaves the abducer with a further choice. Pending the presentation of events that will in time either redeem the hope or dash it, is the abducer content to employ H on instrumental grounds? In most cases, abductions secured by instrumental considerations are those in which the hopes of vindication have yet to be either redeemed or dashed. But there are limiting cases, such as Newton's, in which from the beginning of the abductive exercise, it is conceded that, *epistemically* speaking, there is no hope for the hypothesis of action-at-a-distance.

All abductions embed instrumental factors. In the general case, one accepts H because doing so enables one's target to be hit, notwithstanding that H lacks the relevant epistemic virtue. However, in cases such as Newton's, H is selected notwithstanding that it is taken to be epistemically hopeless. Accordingly,

Proposition 5.1 (Radical instrumentalism) *An abduction is radically instrumentalist when it is made in the belief that the embedded H is epistemically hopeless.*

It is well to note that radically instrumentalist abductions constitute only a proper subset of reasonings made contrary to fact. Although radical, they are not especially uncommon. Rich assays of contrary-to-fact scientific pronouncement are ably scrutinized in such works as Nancy Cartwright's *How the Laws of Physics Lie* [Cartwright, 1983]. Beyond that, contrary-to-fact assertion is at the heart of any empirically directed discipline that discharges it principle claims in ideal models.

So we should not be unduly alarmed by radically instrumentalist abduction.

Let us sum up these points.

Proposition 5.2 (Newtonianly explanationist abduction) *In Newtonian contexts, the concept of explanation is considerably stretched. Newtonian abductions are more perspicuously described as radically instrumental.*

Proposition 5.3 (Instrumentalism in abduction) *Since all abduced hypotheses H*

lack the requisite epistemic virtue, all abductions forward their Hs on correlatively instrumental grounds.

Proposition 5.4 (Explanationist abduction) *In explanationist abduction, H's instrumental value is its (subjunctive)explanatory force.*

Proposition 5.5 (Implausibilist abduction) *It is not a condition on the abductive success of a hypothesis H that it be propositionally plausible.*

Corollary 5.5(a)*There are senses of "explanation" in which propositionally implausible propositions can have explanatory force.*

Proposition 5.6 (Discharging the implausible) *Since discharging a hypothesis is releasing it for future premissory use, implausibility is no bar to hypothesis-discharge.*

It is easy to see that radically instrumentalist abduction have an important bearing on the logic of hypothesis discharge.

Proposition 5.7 (Discharging radically instrumentalist hypotheses) *Since the hypothesis of a radically instrumentalist abduction fails all tests that would reveal it as having the requisite epistemic value, such hypotheses are not subject to discharge except for their instrumental value.*

Current disputes in the philosophy of science between cognitivists and instrumentalists can be seen as disagreements about the extent to which the antecedent of the corollary is true, if at all.

It is interesting to note in passing that the axiom of choice[5] in the ZFC theory of sets has an abductive role, in the Newtonian sense, in proofs of the Löwenheim-Skolem theorem. The theorem guarantees to any true first-order theory a model in the natural numbers. Löwenheim's original proof of 1915 turns out to be equivalent to the axiom of choice. Skolem's more streamlined proof of 1920 makes

[5] The axiom of choice asserts that for any set of non-empty sets there exists a set containing exactly one member from each. The choice axiom is equivalent, among other things, to the well-ordering theorem (which asserts that every set can be well-ordered) and Zorn's lemma (which asserts that if every chain in a partially ordered set has an upper bound, then the set possesses a maximal element).

explicit use of the axiom. In 1922 Skolem produced a proof which averted the necessity to assume choice. Choice was postulated in the 1920 paper precisely because it was thought necessary for the successful derivation of the theorem in question. In that proof, choice had beyond doubt the force of a Newtonian explanation. Even so, this does not yet tell us whether the role of choice in the 1920 proof might also have had plausibilistic explanatory force. If so, it enhances our understanding not of the fact that the theorem follows from the 1920 proof and would not do so in the absence of choice, but rather of the fact that true first-order theories are true in the natural numbers. It would seem that the descriptively adequate answer that reflects how people are in fact affected by the proof, and by its conclusion, is that its explanatory force ranges variously over a range of real-life reactions. At the one extreme we find people who find choice more intelligible than the conclusion of the proof, and who find that conclusion no more intelligible with the axiom than without it (*cf.* the discussion of Russell and Gödel in section 5.6 below). At the other extreme we find those who find the theorem comfortably intelligible, and who may do the same for the axiom. Given that post-1920 developments show that choice is not needed for proving the theorem, the question of the axiom's explanatory force in post-1920 proofs doesn't arise. It does arise for the 1920 proof, however. Here, too, there is a distinction to make.

(a) Did the axiom of choice elucidate the conclusion of the 1920 proof?

(b) Did the axiom of choice contribute to the elucidary force that the *entire* proof had with regard to its conclusion?

The answer to (a) would appear to be "No". The answer to (b) is less obviously in the negative. But it remains the case that even for someone who finds both the axiom and the premisses of the proof richly intelligible, it might not be the case that the proof makes the conclusion more intelligible to anyone satisfying the antecedent condition.

The issue of plausibilistic explanation-potential is made somewhat more complicated by the fact that the theorem is equivalent to the axiom. That is, a true first-order theory is true in the natural numbers if and only if for any set of non-empty sets there is a set containing one and only one member from each.

5.3 Planck

Planck's discovery of the quantum arose out of his study of black body radiation, and turned on the question of the interrelation among temperature T, frequency of radiation v, and the energy density per unit interval of wavelength u_v. Experimental results were thought to suggest two laws, the Rayleigh-Jeans Radiation Law for very low frequencies

$$u_v = \frac{8\pi v^2 kI}{c^4} \tag{5.1}$$

and Wein's Radiation Law for high frequencies,

$$u_v = \alpha_v^3 \exp(-\beta_v/T). \tag{5.2}$$

(In fact, there was little by way of experimental confirmation for Wein's Law; but let that pass.) It is significant that (1) fails for high frequencies, and (2) for low.

Planck undertook to eliminate this kind of dichotomy. It was on its face a forbidding task, given that they represent such different functions mathematically. How was Planck to show that each is a limiting case of some deeper mathematical equation? Of course, in a momentous turn in the history of science, the answer lies in the postulation of minimal separable quantities of energy, or quanta.

It is a matter of some irony that Planck abjured the very idea of them. On the one hand, he recognized that accepting them would trigger a revolution in physics on a scale that would displace Newton; but even so, Planck spent years trying to think up alternatives to the postulation of quanta.

We see in this the structure of an abduction problem. The physics of the day could not produce a unified law of black body radiation. This was Planck's abductive trigger. His target was a physics that would make possible the unification of these laws. His conjecture was the quantum hypothesis. Apart from the fact that Planck thought it untrue (shades of Newton), the quantum hypothesis was indeed momentous. Planck's case is one of the most impressive examples in the history of science of the tail wagging the dog, hence of the enormous costs abductive reasoners are sometimes prepared to pay to hit their targets.

The invention of quantum physics also resembles the later discovery by Gell-Mann of quarks. As we earlier remarked, Gell-Mann saw quarks as artifacts of the mathematics that drove his theory. Quarks were epiphenomena to be tolerated, because the mathematics that generated them was indispensable to the general physics. In the early days Gell-Mann was openly skeptical of quarks, and stoutly denied their empirical reality. When experimental confirmation was eventually forthcoming, Gell-Mann was happy to forget his earlier hostility.

5.4 Physical Dependencies

Philosophers of science tend to side with Bacon in supposing that a central feature of science — indeed one of its greatest triumphs — is the explanation of phenomena by way of the lawlike structures within which they occur. Examination of the history of science, especially that of the physical sciences, show this to be a misconception. In the physical sciences the dominant target is the revelation of *physical dependencies*. It lies in the nature of such dependencies that for large ranges of cases they admit of high levels of evidential and theoretical support, well short of explanatory effect. Some writers tend to blur the distinction between physical dependency and explanation by aligning it with a putative contrast between different *kinds* of explanation. In [Thalos, 2002], the purported congruency is with the explanation/causal explanation contrast, with physical dependency likened to a sort of non-causal explanation – or anyhow a causally neutral kind of explanation.[Magnani, 2001a, p. 17] Our own view is that such congruencies are a trifle forced, and that it is preferable to acknowledge outright that it is physical dependencies that, as such, are explanatively underdetermined. If this is so, it further confirms our claim that

Proposition 5.8 (Abduction and Explanation) *Abduction is neither intrinsically or typically explanationist.*

5.5 The Superstring Controversy

> Superstring theory starts off by proposing a new answer to an old question: what are the smallest, indivisible constituents of matter? For many decades, the conventional answer has been that matter is composed of particles — electrons and quarks — that can be modelled as dots that are indivisible and that have no size and no internal structure. Conventional theory claims, and experiments confirm, that these particles combine in various ways to produce protons, neutrons, and the wide variety of atoms and molecules making up everything we've ever encountered. Superstring theory tells a different story. It does not deny the key role played by electrons, quarks, and the other particle species revealed by experiment, but it does claim that these particles are not dots. Instead, according to superstring theory, every particle is composed of a tiny filament of energy, some hundred billion billion times smaller than a single atomic nucleus (much smaller than we can currently probe), which is shaped like a little string. And just as a violin string can vibrate in different patterns, each of which produces a different musical tone, the filaments of superstring theory can also vibrate in different patterns. But these vibrations don't produce different musical notes; remarkably, the theory claims that they produce different particle properties. A tiny string vibrating in a different pattern would have the requisite properties to identify it as a quark, a neutrino, or any other kind of particle. All species of particles are unified in superstring theory since each arises from a different

> vibrational pattern executed by the same underlying entity [Greene, 2004, pp17–18].

A prime motivation of superstring theory is the deep tension between modern physics' two most dominant theories, the general theory of relativity (GT) and quantum mechanics (QT). GT is at bottom a theory of gravity applied to classical objects. Its gravitational field is a curvature of space-time, in which Humpty-Dumpty's fall is characterized as a response to space-time curvature induced by the earth's mass. GT's objects are classical, of a size that affords the quantum uncertainties no real bearing. In contrast, quantum objects — atoms and the elementary particles — are riven with the quantum uncertainties. Here too, size is the principal factor. Quantum objects are too small to be uninfluenced by quantum forces and also too small to create significant gravitational fields. Given the present state of mainstream physics, GT and QM divide the workload into two empirically well-confirmed theories — one for the large and the other for the small — that are mathematically and conceptually incompatible. For GT, space-time is a plasticity drawn and repelled by classical objects. For QM, space-time is a rigid framework within which quantum events occur.

As things now stand physics is disunified. Neither GT nor QM, or any plausible extensions of them, offer any promise of overcoming their mathematical and conceptual incompatabilities. A good part of what makes superstring theory important lies precisely in its claim to repair this difficulty by way of a theory of *quantum gravity* that works for classical and quantum objects alike. If it passes scientific muster, quantum gravity is the (one and only) theory that unifies physics, and thus attains *en large* the same kind of target that Planck set for himself with regard to the laws of black body radiation. As with Planck's original hypothesis; there isn't the slightest evidence currently or forseeably at hand that the superstring hypothesis is experimentally testable.

At the heart of the theory of quantum is the postulation of *gravitons*, which by analogy with the photon as a quantum of light, is a quantum of gravity. Photons are detectable by observational methods. They involve the expulsion of electrons from metal surfaces by light beamed on them. Part of what makes these detections possible is that electromagnetic interactions are considerably stronger than gravitational interactions. If one attempted a similar electron-expulsion test for incident gravitational waves, the universe itself would have expired before the requisite observations presented themselves. At least that would be so for detectors of normal density. If the detector were compressed to a high enough density to be observationally effective, it would collapse into a black hole. At present, there is no evidence that these difficulties can be averted, hence no evidence that gravitons can be observed. If present indications are accurate, the graviton hypothesis is observationally contestable. The positivistically-minded will see this as tantamount to demonstration of the physical unreality of gravitons. Those less positivistically inclined may pin their hopes on the theory's prospects for empirical adequacy. Those at ease with radical instrumentalism will allow that a theory's scientific adequacy is not exhausted by its observational success. In its starkest terms, the contention over superstring theory reflects a tension between unity without observability and observability without unity. It is a conflict which abductive logicians must take note of, and accommodate, but need not solve. Accordingly,

Proposition 5.9 (Radical instrumentalism and experimental testability) *It lies in the nature of radically instrumental abduction that winning hypotheses not be held to the requirement of experimental testability.*

5.6 Russell and Gödel

Newton strives to be a faithful inductionist, as the General Scholium of 1713 makes clear. '*Hypotheses non fingo*' may have been directed at the action-at-a-distance theorem specifically, but it is clear that Newton disdained the method of hypotheses in a quite general way. What is interesting about Newton's decision to retain the incoherent theorem is that it resembles a form of abduction, or a variant of it, that is not much recognized in the literature, but which in fact plays an important role in the foundations of mathematics. For reasons that will shortly declare themselves, we will call this sort of abduction *regressive abduction*, after Russell's discussion of it in a paper delivered in Cambridge in 1907, but which remained unpublished until 1973 [Russell, 1973]. The title of this work is 'The Regressive Method of Discovering the Premises of Mathematics'.

Russell's regressive method forms an important part of his logicism, which, as correctly observed in Irvine [1989], is widely misunderstood. For our purposes it suffices to concentrate on one particular feature of Russell's mathematical epistemics; readers interested in the fuller story of Russell's logicism could consult Irvine [1989] with profit. The specific issue which concerns us here is the problem of determining in what sense 'a comparatively obscure and difficult proposition may be said to be a premise for a comparatively obvious proposition' [Russell, 1973, p. 272]. A further problem is that of explaining how such obscure and difficult propositions are discovered and justified. One thing that Russell certainly did think that logicism involved is the doctrine that '. . . all pure mathematics deals exclusively with concepts definable in terms of a very small number of fundamental logical concepts, and that all its propositions are deducible from a very small number of fundamental logical principles . . .' [Russell, 1973, p. xv]. However, as Russell freely admits,

> There is an apparent absurdity in proceeding, as one does in the logical theory of arithmetic, through many recondite propositions of symbolic logic to the 'proof' of such truisms as $2+2=4$: for it is plain that the

> conclusion is more certain than the premises, and the supposed proof therefore seems futile [Russell, 1973, p. 272].

Consider such a 'proof', in which the premisses are recondite propositions of logic and the conclusion is an arithmetic truism such as '$2 + 2 = 4$'. The purpose of such a proof is not to demonstrate the truism, but rather to demonstrate that it follows from those logical premisses. Russell then proposes that in those cases in which an antecedently accepted conclusion can be shown to be deducible from a given logical premiss or set of premisses, then the fact of such deducibility tends to confer some justification, not on the antecedently accepted *conclusion*, but rather on the premiss or premiss-set of the deduction in question.

Russell calls the method of reasoning by which such justifications are proposed the *regressive method*, which, as Irvine points out (citing Kleiner), is a 'method similar to Peirce's abduction' [Irvine, 1989, p. 322, n. 26].[6] As Russell goes on to say,

> We tend to believe the premises because we see that their consequences are true, instead of believing the consequences because we know the premises to be true. But the inferring of premises from consequences is the essence of induction; thus the method in investigating the principles of mathematics is really an inductive method, and is substantially the same as the method of discovering general laws in *any other science* [Russell, 1973, pp. 273–274].[7]

On Russell's regressive account even the most fundamental of logical laws may be only probable [i.e., other than certain]. This matters. If logical laws needn't be attended by certainty or by what Quine calls "obveity", thus the logician is spared the burden of finding for his axioms a *Fregean* justification in which 'they are truths for which no proof can be given ... and for which *no proof is needed*. It follows that there are no false axioms, and that *we cannot accept a thought as an axiom if we are in doubt about its truth*' [Frege, 1914, p. 205; (emphasis added)].[8] Since Russell thought that this burden could not be met, especially in the wake of the antinomy that bears his name and which ended his 'intellectual honeymoon', he is able to plead the case for wholly non-intuitive axioms such as, notoriously, infinity, without having to concede that its truth is more than merely probable.

> In induction [sic], if p is our logical [and obscure] premise and q our empirical [i.e., pretty clearly true] premise,[9] we know that p implies

[6] However, we ourselves demur from the suggestion that Russell's abductions here are explanationist.

[7] Note here that Russell's use of 'induction' is that of a generic concept of non-deductive or ampliative inference of which abduction is a kind.

[8] For a more detailed discussion of the difference between Frege and Russell concerning the epistemology and logic and mathematics, see Woods [2003].

[9] Russell's use of 'empirical' is metalevel. He here means *clearly* or *obviously* true.

> q, and in a text-book we are apt to begin with p and deduce q. But p is only believed on account of q. Thus we require a greater or less probability that q implies p, or, what comes to the same thing, that not-p implies not-q. If we can *prove* that not-p implies not-q, i.e. that p is the only hypothesis consistent with the facts, that settles the question. But usually what we do is to test as many alternative hypotheses *as we can think of.* If they all fail, that makes it probable more or less, that any hypothesis other than p will fail. But in this we are simply betting on our inventiveness: we think it unlikely that we should not have thought of a better hypothesis if there were one ([Russell, 1973, pp. 274–275]. Emphasis added in the second instance.)

Note in passing Russell's use of *ad ignorantiam* reasoning in a form which AI researchers call *autoepistemic*, and which is a far cry from always being a fallacy. As we have already remarked, in its most elementary form,

1. It is not known that p

2. Therefore, not-p,

the *ad ignorantiam* is fallacious. In certain autoepistemic variations, it is not, as in our previous example:

a. If there were a Departmental meeting today, I would know it.

b. But I have no knowledge of such a meeting.

c. Therefore, the Department doesn't meet today.

Thus,

i. If there were a better hypothesis we would know it by now.

ii. But we don't

iii. So there isn't.

We shan't here press the question of how safely autoepistemic Russell's elimination argument is. But we note that it is of a type which often plays a central role in abductive reasoning.

We have it, then, that it is Russell's view that more often than not axioms are accepted abductively, rather than by way of their self-evidence.

> The reason for accepting an axiom, as for accepting any other proposition, is always largely inductive [sic], namely that many propositions which are nearly indubitable can be deduced from it, and that no equally plausible way is known by which these propositions could be

> true if the axioms were false, and nothing which is probably false can be deduced from it [Whitehead and Russell, 1910, pp. 59 and 62 of 1st ed.].

Russell is clear in seeing regressive abduction as instrumental. He goes so far as to concede that his axiom of reducibility has a 'purely pragmatic justification' [Whitehead and Russell, 1910, p. xiv]. In this, it is clear that Russell anticipates Gödel, who held that

> . . . even disregarding the intrinsic necessity of some new axiom, and *even in case it has no intrinsic necessity at all*, a probable decision about its truth is possible also in another way, namely inductively by studying its 'success'. Success here means fruitfulness in consequences, in particular, 'verifiable' consequences, i.e., consequences demonstrable without the new axioms, whose proofs with the help of the new axiom, however, are considerably simpler and easier to discover, and make it possible to contract into one proof many different proofs [Gödel, 1990b, pp. 476–477 (emphasis added)].[10]

Gödel acknowledges the Russellian precedents of his own regressive views, and he joins with Russell in emphasizing the epistemological similarity between mathematics and the natural sciences. 'The analogy', says Gödel, 'between mathematics and a natural science is enlarged upon by Russell also in another respect . . . the axioms need not necessarily be evident in themselves, but rather their justification lies (*exactly as in physics*) in the fact that they make it possible for these 'sense perceptions' to be deduced . . .' [Gödel, 1944, p. 127].

Irvine has provided a valuable service in having unearthed Russell's notion of regressive abduction. On the face of it, it is a dubious idea when applied to the foundations of mathematics. One might be forgiven for thinking it to be a quirk of Russell's early thought and something that he had the sense to abandon in good time.[11] In fact, however, Russell never did wholly give up on regressive abduction in mathematics and logic. It remained a part of his thinking in much the way that a more general thread of pragmatism is discernible in many of his writings, early and late. Beyond that, the idea of regressive abduction has been taken up with considerable gusto in the last twenty-five years, in what is called, by Harvey Friedman and others, *regressive mathematics* [Friedman and Simpson, 2000]. The

[10] *Cf.* '. . . probably there exist other [axioms] based on hitherto unknown principles . . . which more profound understanding of the concepts underlying logic and mathematics would enable us to recognize as implied by these concepts . . . [and] so abundant in their verifiable consequences . . . that quite irrespective of their intrinsic necessity they would have to be assumed' Gödel[1990a].

[11] Gödel's subscription to it is even more quixotic-seeming. How is it possible for so radical a Platonist about mathematical objects and the manner of our knowing them be such an instrumentalist about mathematical axioms? This question is well-discussed in[Rodych, 2005].

central idea of regressive mathematics is purely Russellian. Propositions can be adopted as axioms entirely for the contribution they would then make to axiomatic proofs of target theorems.

A further example of regressive abduction can be found in the Jordan Curve Theorem, which establishes that if there is a closed curve in the plane satisfying certain conditions, then it divides the plane into 'inside the curve' and 'outside the curve.' Call this proposition F. It is interesting to note that the proof of the Theorem was preceded by seven years of effort and failure. When the proof was finally achieved it did not establish the truth of proposition F. That proposition had long been known and did not require *establishment*. The seven-year lull prior to the proof was instructive. It showed that there was something not quite right with the proof apparatus of plane geometry, since it couldn't deliver the obvious truth F. In a certain way, the eventual success of the proof resembles the situation involving Galileo's telescope. Before its apparent disclosures of canals on the Mars could be believed, the instrument was first trained on a mid-distant but well-known church to verify features it was already known to have. When the telescope 'proved' these features, it was considered a generally reliable optical instrument. In the same way, when geometry was finally tricked out in ways that enabled the proof of the Jordan Curve Theorem (complex variables and topology), this served abductively to indicate the general reliability of the theory as an instrument of which investigates the mathematical structure of space.

5.7 The Consequence Relation

We have made some progress with the details of explanationist and plausibilist abduction, but we have done little by way of direct interpretations of $\leadsto$ itself in those contests in which it plays a role. Here is why. There already exists a large and somewhat disputatious literature concerning the semantics of consequence in various contexts. In the case of abduction, when it comes to interpreting the $\leadsto$-relation, we might simplify by saying that it all depends on the target. If the target is to achieve a proof of some proposition, then $\leadsto$ must be a proof relation such that K alone doesn't imply the payoff proposition V and $K(H)$ does. If the target is to provide a D-N explanation of some proposition, or a Popperian prediction of it, then $\leadsto$ imbibes the deductive character that such explanations possess intrinsically. If the target is to increase the conditional probability of some proposition, then $\leadsto$ will be interpreted inductively, in ways that capture the greater-than relation from prior to posterior probability values. For these and other cases, theoretical accounts of these relations already exist in abundance, and they do not always agree. Our position is that unless at various junctures the analysis of abduction *particularly* requires it, it is unnecessary for us to re-invent the wheel or to embroil ourselves in disputes about the "true nature" of deduction (or whatever else). But we should not shy away from declaring a bias. It is that in the matter of interpreting the consequence relation in abductive contexts, we favour interpretations as classical as the abductive circumstances permit. This stands in marked contrast to what is required by interpretations of the conclusion-operator $\therefore$. Forewarned by the spectre of affirming the consequent, the conclusion operator cannot be classically interpreted.

Whether we are strictly correct in our bias for renderings of $\looparrowright$ as classical as can be managed, there is, even so, an important difference between the respective logics of consequence and conclusion in abductive contexts. If we consider once again the core part of the consequentialist–explanationistabduction schema,

1. $K(H) \looparrowright V$
2. $\therefore H^c$

we see the presence of two conditionals, not one. There is the conditional displayed in the premiss, $H \looparrowright T$; and there is the conditional corresponding to the argument itself, viz., $((K(H) \looparrowright V) \looparrowright H)$ in which the rightmost occurrence of $\looparrowright$ does not by and large bear the same interpretation as the leftmost occurrence.

Proposition 5.10 (Interpreting conditionals) *Whatever the interpretation of the* $\looparrowright$ *in* $K(H) \looparrowright V$, *the interpretation of the corresponding conditional,* $((KCH) \looparrowright V) \supset H^c$, *cannot be not stronger than it.*

It may strike some readers that the claim of Proposition 5.10 is a bit excessive. Consider some cases. Suppose that we have an abduction problem, for the attainment of whose target T, it is required that for some V and K there is an H such that V is a deductive consequence of $K(H)$. Suppose that this condition is met. Then there exists between $K(H)$ and V the strongest possible kind of consequence relation. There is no valuation verifying $K(H)$ that fails to make V true. If we take it that the truth of this conditional suffices for the attainment of abductive target T, what then? According to the abduction schema we conclude $C(H)$. What is the force of these conclusions? It is not easy to say in a general way. But in the present case it is obvious that the force of both conclusions is weaker than the strength of the $\looparrowright$-relation. In refusing the claim that H is a justified conjecture, or that H is ready for further premissory work, one might be making an abductive mistake — even a very serious abductive mistake. But it can hardly be said that one has contradicted oneself.

Here is a case at the other extreme. Let T be an abductive target, K a knowledge-module, H a hypothesis, V a proposition, and $\looparrowright$ a consequence relation. Now give to $\looparrowright$ the weakest interpretation under which the truth of

(1) $K(H) \looparrowright V$

suffices for the attainment of T. For concreteness, suppose that T calls for a not wholly implausible (or a weakly plausible) explanation of the state of affairs expressed by V. Given the truth of (1), we may say informally that H explains V in a weakly plausible sense. Suppose finally that apart from the truth of (1), V is wholly inexplicable and T is completely unmet. Could it not be concluded with a stronger force than the force with which H explains V with weak plausibility that H might be conjectured as true? We grant that opinions might divide in this question. But provided that it is clear that

(a) conjecturing that H

and

(b) determining that H is a good bet for testing.

are logically independent judgements, our answer is "No" with respect to (a) and "Yes" with respect to (b). Proposition 5.10 remains intact.

We now turn our attention to the question of hypothesis-discharge, about which, after Peirce, it may be said that Popper's views are best known. Popper is unique among 20th century philosophers of science. His theories were, and still are, taken seriously by some working scientists, and to some extent they have influenced the work of a few of them. Further evidence of the regard in which he was held by scientists is the work Popper did in collaboration with them (see, e.g., [Popper and Eccles, 1983]).

There is, apart from Popper's own prodigious writings, a substantial literature on his views. It is not our purpose to add to that total. The basics of Popper's approach to science will be known to most readers of this book and, in any went, we wish here to concentrate only on those features of it that bear on abduction. Those features are well summed up in the title of one of his more important works, *Conjecture and Refutation* [1963].

Popper rejects inductivism in science. He thinks that Hume's Problem of Induction has no solution save evasion. Evade it he does, declaring that scientific laws derive none of their security from high degrees of confirmation by way of true positive instances. The security of scientific laws is provisional at best. It derives from the fact that persistent and sophisticated attempts to refute them have so far failed. When this condition is met, Popper is prepared to say that such a law is *corroborated* by its positive instances. But he emphasizes that corroboration (also a provisional attribute) has no confirmatory significance. Thus the corroboration of a scientific law or theory consists in its predictive (as opposed to explanatory) success to date coupled with the failure to refute it to date.

Popper is a deductivist about science. One knows a scientific theory through its deductive consequences. Refutation imbibes the spirit of Popper's deductivism. A scientific law or theory is refuted even by a single true negative instance. Such a notion of refutation flies in the face of scientific practice but, according to Popper, this would show only that such scientific practice is philosophically misconceived. Even if we allow that an entire scientific theory is falsified by any given true negative instance, there is nothing in Popper's conception of refutation that enables us in a systematic way to localize the theory's defect, so that comparatively minor repairs might be considered. This raises the question of whether alternative accounts of refutation might be more hospitable to the idea of localized diagnosis. (See here [Gabbay and Woods, 2003a].)

There is an odd asymmetry between Popper's conception of conjecture and his account of refutation. Given what he takes refutation to be, scientific theories are maximally brittle and unstable. There is no latitude that can be given a theory in the face of even a single true negative instance. Conjecture, on the other hand, is a liberally latitudinarian notion. We have seen that Peirce thinks that conjecture is constrained instinctually. Others require that conjectured hypotheses meet various conditions of conservatism, of what Quine calls the "maxim of minimal mutilation". Popper is not so-minded. Conjectures need not be constrained in any of these ways. There is in the whole literature on abduction no freer rein given to the generation and engagement of hypotheses. In some respects, Popper thinks that the wilder-the-better is the applicable standard.[12] Thinking so would make Planck's quantal speculation a conjecture of the very best kind, of a kind to which scientific revolutions are frequently indebted. Implausible, instinctually alien, mutilating, the quantum conjecture blasted physics from its classical embrace into the most successful physical theory in the history of human thought. This is not to say that Popper is a total conjectural promiscuist. Hypotheses cannot be entertained unless they are falsifiable. This is a point that returns us to an earlier difficulty. Let H be a hypothesis that we are considering making. If we intend to do thing Popper's way we must drop H if we think that it is not falsifiable. This we might not know until we find a place for H is a complex scientific theory of a sort that we take to be falsifiable. Suppose later that a true negative instance is unearthed. Then the theory is false. But we still might not know whether H is false. If this is all that we have to go on, we might not even know whether H is falsifiable.

In the logic of generation and and engagement, there is little that Popper has to offer, except negatively.

> My view of the matter ... is that there is no such thing as a logical method of having new ideas, or a logical reconstruction of this process [Popper, 1934, p. 31].

[12] Not to forget Russell: "But usually what we do is test as many alternative hypotheses as we can think of" [Russell, 1973, p. 275].

Indeed,

> the act of conceiving or inventing a theory seems to me to be neither to call for logical analysis nor to be susceptible of it [Popper, 1934, p. 32].[13]

So plausibility is not a constraint; conservatism is not a constraint; intuitiveness is not a constraint; and so on. And the one positive requirement, falsifiability, is problematic. In the logic of discharge, as Popper will go only half-way. Strictly speaking, nothing about a scientific theory is categorical. All of science is on sufferance; and the best that we can hope for is that our favorite theory hasn't yet been put out of business. When it comes to discharging a hypothesis, Popper recognizes that the closest that one decides to leave it in place with whatever confidence it deserves to have on the basis of its corroboration record to date. But in so doing.

Popper confuses defeasibility with hypotheticality. Popper is assuredly correct to emphasize that scientific theories and the hypothesis that they embed are forwarded defeasibly. They are forwarded in ways that recognize the fallibility of all science. But this is no bar to categorical assertion in science it; is not required that scientific utterance be restricted to hypothetical utterance. There is no equivalence between

1. P, although I may be wrong

and

2. I conjecture that P.

When we say that Uncle Frank will live to a hundred, we may be wrong and we may say so. This is compatible with our strong conviction that, already very old, he will nonetheless achieve his century. "My conjecture is that he will live to a hundred" is much weaker.

Popper's confusion of defeasible utterance with conjectural hypotheses might be occasioned by his insistence that no conjecture is ever justified.

> [T]hese conjectures . . . can neither be established as certainly true or even as "probable" (in the sense of the probability calculus) . . .. these conjectures can never be positively justified [Popper, 1963, preface].

Even so, we repeat the point that insusceptibility to justification does not preclude categorical utterance in the form H^c.

Although both Peirce and Popper refuse the idea of abduction as an enhancer of probability, it would be a mistake to think that this commits them to reject what

[13] For a discussion of how faithful Popper was to these rather uncompromising strictures, see [Aliseda, forthcoming].

we earlier called probabilistic abduction. A probabilistic abduction problem is one in which a target T cannot be hit from propositions one takes to be probable to at least some degree k. The problem is solved by finding a proposition H with no higher epistemic value than that of a proposition of degree of probability $k - 1$, such that H together with what one already took to be probable to degree k or higher does hit T. In that case, we abduce H; that is to say, we conclude — as Peirce also insists — given that H is a proposition worthy of conjecture, it is all right to release it for premissory work. We do *not* conclude that H has any given probability value, although we are free to ruminate about that after suitable trials, after which it may be found to have a probability value of k or higher. Accordingly,

Proposition 5.11 (Probabilistic abduction) *Since probabilistic abductions do not confer a probability value upon their conclusions, they are inferences of a kind that comport with Peirce's and Popper's insistence that abduction does not* establish *probabilities. (See the discussion of so-called "Bayesian" abduction in 5.12.1 below).*

5.8 Lakatos

It will repay us to touch briefly on some ideas loosely developed by Imre Lakatos [1970] and [1968], which can be seen as a corrective to Popper's conception of refutation. Popper understood a scientific theory as a set of deductions derived from conjectures. If the deduced consequences are false, the theory that implies them is false. Although Popper's general methodological orientation was something Lakatos agreed with, he finds fault with Popper's account in two respects. Popper erred in conceiving of the methods of science too narrowly and too strictly. The narrow error was that of thinking that individual theories are the intended targets of scientific testing. In fact, Lakatos claimed science tests not individual theories but *families* of theories, which he called 'research programmes'. The error of excessive strictness was the mistake of supposing that a research programme is automatically toppled if it licenses a false prediction. Lakatos writes:

> The main difference from Popper's is that in my conception criticism does not — and must not — kill as fast as Popper imagined. Purely negative destructive criticism, like 'refutation' or demonstration of an inconsistency does not eliminate a [research] programme. Criticism of a programme is a long and often frustrating process and one must treat budding programmes leniently. One may, of course, shore up the degeneration of a research programme, but it is only *constructive criticism* which with the help of rival research programmes, can achieve real success [Lakatos, 1970, p. 179 (original emphases omitted in some cases)].

As Lakatos sees it, a research programme is what Kuhn *should* have understood a paradigm to be [Kuhn, 1967; Kuhn, 1970]. Against Kuhn's idea that normal science embeds a paradigm, Lakatos holds that what 'Kuhn calls 'normal science' is nothing but a research programme that has achieved monopoly', and 'one must never allow a research programme to become a *Weltanschauung*, or a sort of scientific rigour' [Lakatos, 1970, p. 155 (emphasis omitted)].

The idea of research programmes is central to how Lakatos conceives of a logic of discovery. The basic concept of a logic of discovery is not the theory, but rather families or series of theories. 'It is a succession of theories and not one given theory which is approved as scientific or pseudo-scientific' [Lakatos, 1970, p. 132]. The central issues of the logic of discovery require for their consideration the framework of the methodology of research programmes. A research programme can be defined by way of *problem shifts*. Suppose that $T_1, T_2, T_3, \ldots$ T_n is a sequence of theories in which every subsequent member results from its predecessor by addition of supplementary statements or by providing its predecessor with a semantic interpretation with which to take account of an observational anomaly. A problem shift is *theoretically progressive* if each subsequent theory is empirically richer than its predecessor, if, in other words, the successor theory predicts a new fact beyond the reach of the predecessor theory. A problem shift is *empirically progressive* if the new predictions of a successor theory are empirically corroborated. A problem shift is progressive if it is both theoretically and empirically progressive. A problem shift is degenerative if it is not progressive [Lakatos, 1970, p. 118].

When a science is in a state of immaturity, problem shifts are a matter of trial and error. Mature theories have *heuristic power* which guides the science in the more orderly development of its research programmes. A positive heuristic gives counsel on what research strategies to pursue. A negative heuristic guides the researcher away from unfruitful research protocols. Lakatos sees the negative heuristic as framing the key insight of a problem shift, its conceptual core, so to speak; and 'we must use our own ingenuity to articulate or even invent auxiliary hypotheses which from a *protective belt* around this core' [Lakatos, 1970, p. 133]. Thus the conceptual core is *irrefutable* and is made so by the theorist's methodological decision, an idea that calls to mind Quine's jape that theories are free for the thinking up, and Eddington's that they are put-up jobs. Concerning the matter of what enters into such decisions, to say nothing of what makes them scientifically rational, Lakatos has nothing to say.

There is value in Lakatos' idea of heuristic power. It enables him to stick with Popper's insistence that explanatory success is not intrinsic to scientific progress. Even so, science cannot be done well — it cannot be done at all — without some understanding of what the issues are and of the ways and means of exploring them. Lakatos has it that one of the virtues of a mature theory is that it has this capacity to illuminate the research programme. In this it has a clearly explanative function; yet it has it without the necessity it also being an explanation of the phenomena it successfully covers. It also has this function without it being necessary that the hypotheses that drive future research have explanatory force with respect to the modified theory's predictions.

A science's positive heuristic specifies its range of refutability. It identifies the 'refutable variants' of the research programme and makes suggestions about how to modify the 'refutable protective belt' of the programmes core insights [Lakatos, 1970, p. 135]. The positive heuristic

> sets out a programme which tests a chain of ever more complicated models simulating reality: the scientist's attention is riveted on building his models following instructions which are laid in the positive part of his programme; he ignores the *actual* counterexamples, the available data [1970, p. 135].

A theoretically progressive problem shift arises from attempts to falsify theories and then to save them by supplementary hypotheses or semantic reinterpretation or both. A theory can be seen as a negative heuristic together with auxiliary hypotheses worked up as a model. The only reason to abandon a research programme is that it is degenerate (a fact which is often enough apparent only in hindsight). The idea that a theory should be abandoned if its fundamental insights are false or if they lack inductive support. Lakatos thinks it likely that *all* core ideas of research programmes are false, and he joins with Popper in insisting that inductivism is a wholly misbegotten idea in the methodology of science [Lakatos, 1970, pp. 155–173, 99–103]. It follows, therefore, that it is irrational to suppose that *any* given scientific theory is true; and this, says Lakatos, is a powerful and correct inducement to scientific pluralism and the proliferation of research programmes.

Hypotheses are the work of a programme's positive heuristic. Their function is to protect the programme's core insight from observational discouragement or from predictive impotence. The hypothesis is justified (or rational) to the extent that it facilitates novel predictions that turn out to have observational corroboration, if not immediately, then soon enough in the evolving successiveness of the research programme and before there is occasion to attribute degeneracy. Core ideas, on the other hand, while not strictly refutable, can over time be discredited by the failure of the programme's positive heuristic to protect it. A core idea is in trouble if the research programme to which it is central is degenerating.

Sketchy though it surely is, Lakatos has some conception of how to evaluate a prior claim in the form H^c. Hypothesis-generation and engagement are another matter. Here Lakatos has little to say except — in ways that call Hanson to mind — that hypotheses should be selected for their suitability to the positive heuristic of a research programme. Thus those hypotheses should be considered which are of a type that might do well if made part of such a heuristic.

Lakatos' views have attracted a large number of criticisms, three of which are especially important [Suppe, 1977, pp. 667–668]. One is that in the research programme defined by a series of theories T_1, T_2, ..., the changes wrought by a successor theory may not, according to Lakatos, include the deletion of non-core claims made by the predecessor theory. Another is that in his account of how a research programme arises, Lakatos over-concentrates on the modification of theories within a problem shift, and accordingly pays insufficient attention to what Dudley Shapere [1977] calls 'domain problems'. A domain problem is the problem of determining with sufficient precision the target questions of an emerging research programme; it is the question of what the programme is appropriately *about*. A third criticism, linked to the first, is well developed by Toulmin [1972]. The idea of what theoretical strategies are appropriate is seriously underdetermined by an absence of anchoring *concepts* of requisite type. Lakatos admits no notion of conceptual adequacy in theory construction save for the device of semantic reinterpretation in the production of problem shifts. But as Toulmin points out, the history of science is dotted with ideas of such inappropriateness that no amount of semantic reinterpretation (fairly so-called) would put things right.[14]

The vigorous disagreements that Lakatos' work has occasioned leave at at least one point undisturbed (although, ironically, for reasons that Lakatos' may not have entertained). It is that

Proposition 5.12 (Refutation) *If* (H) *is the conclusion of an abductive inference and H is subsequently shown to be false, this discredits neither the conclusion* $C(H)$ *nor the conclusion* H^c.

Corollary 5.12(a) *This shows the importance of recognizing that the conclusions of abductions imbed (often implicitly) temporal parameters. The abducer concludes that it is reasonable* now *to conjecture that H and to release it for premissory action . A subsequent falsification of it shows only that it would not be reasonable to conjecture that H* any longer *and to* retain *it for premissory work, not that it wasn't reasonable to do so* then.

[14] It is interesting to note that Lakatos' philosophy of science is a model of a form of non-cooperative dialogue logic known as *MindClosed* [Gabbay and Woods, 2001c; Gabbay and Woods, 2001a]. In its pure form *MindClosed* offers an impenetrable form of defence of a player's thesis T, in which every critical move is rejected out of hand. The reasoning underlying these rejections is that since T is true, (or obviously, self-evidently or necessarily true), the criticisms not only must fail, but can be seen to fail without detailed consideration of their purported weight. We join with Lakatos' critics in thinking that he presses too far what, in effect, is the *MindClosed* model. Still, it seems clear to us that a science defends its organizing conceptualizations and core insights in ways that approximate to the *MindClosed* strategy.

5.9 Hintikka

During the past twenty-five years and more, Jaakko Hintikka and his collaborators have been pursuing a fundamental reform of logic. Hintikka's view is that a psychologically real and technically adequate account requires us to think of logic as intrinsically game-theoretic and erotetic. It is game-theoretic in as much as it assigns a central role to strategic factors. It is erotetic in as much as it is a logic of interrogative inquiry [Hintikka *et al.*, 2002].

Hintikka's approach to abduction is likewise game-theoretic and erotetic. It is an approach that is neither explanationist nor probabilist, although it bears some affinity to a class of diagnostic logics which themselves admit of probabilistic interpretation (e.g., [Peng and Reggia, 1990]).

Hintikka finds intimations of his account in Peirce, who defines logical operators by dialogical rules that yield decidable proof-protocols. [15]

> I call all such inference by the peculiar name, *abduction*, because it depends upon *altogether different principles from those other kinds of inference* [Hintikka, 1999b, p. 97, quoted from Peirce. Emphasis added in the second instance].

These principles are summarized in Kapitan [1997, p. 479].

1. Inference is a conscious voluntary act over which the reasoner exercises control [Peirce, 1931–1958, pp. 5.109, 2.144].
2. The aim of inference is to discover (acquire, attain) new knowledge from a consideration of what is already known (MS 628:4).
3. One who infers a conclusion C from a premiss P accepts C *as a result* of both accepting P and approving a general *method* of reasoning according to which if any P-like proposition is true, so is the correlated C-like proposition [Peirce, 1931–1958, pp. 7.536, 2.44, 5.130, 2.773, 4.53–55, 7.49].
4. An inference can be either *valid* or *invalid* depending on whether it follows a method is conductive to satisfying the aim of reasoning, namely, the acquisition of truth [Peirce, 1931–1958, pp. 2.153, 2.780, 7.44], Peirce (MS 692:5).

Hintikka's general logic distinguishes two types of inference rules which regulate the moves of players in games of interrogative enquiry. *Definatory* rules specify moves that are permissible. *Strategic* rules specify moves that are advisable. Hintikka sees evidence of the distinction between definatory and strategic rules in Peirce [Hintikka, 1999a, p. 98 ff.]. He attributes to Peirce the view

[15] See also [Lorenzen and Lorenz, 1978], which extends this approach to a constructive foundation for classical logic.

that the two classes of rules have different justification conditions, and he thinks that his own strategic rules are a good candidate for Peirce's 'altogether different principles'. A definatory rule is valid in so far as it 'confers truth or high probability to the conclusion of each particular application . . . ' [Hintikka, 1999a; Sintonen, 1993, pp. 98–99][2002] and [Wisniewski, 1995]. Strategic rules, on the other hand, can be justified even though they may fail in individual cases. They are justified not by their universal success but by their propensity for success. They are game-theoretic principles. They honour game theory's insight that for the most part the acquisition of utilities is best achieved by way of general strategies, rather than individual moves. Hintikka's own deep insight is that measures for guaranteeing truth-preservation or probability-enhancement greatly underdetermine what an investigator should do, move-by-move. Strategic rules describe a general method which pays off in the long run even though it may land the investigator in local turbulence. Hintikka sees his strategic rules prefigured in Peirce's description of the contrast between induction and abduction. Induction is a type of reasoning

> which professes to pursue such a model that, being persistent in, each special application of it ... must at least indefinitely approximate to the truth about the subject in hand, in the long run. *Abduction* is reasoning, which professes to be such that in the case there is ascertainable truth concerning the matter in hand, the general method of this reasoning though not necessarily each special application of it must eventually approximate the truth [Eisele, 1985, p. 37].[16]

Where Hintikka parts company with Peirce is in claiming that his definatory-strategic distinction *cuts across* the deduction-induction-abduction trichotomy. Thus a bit of reasoning is not made abductive just because it involves the use of strategic rules. Every move of a reasoner, deductive, inductive or abductive, involves the presence of both types of rules. The definatory rules tell him whether the move is permissible. The strategic rules tell him whether it is *smart*. A definatory rule must be 'conductive to the acquisition of truth.' A strategic rule need only be part of a general method which may have local failures (hence is a type of rule tailor-made for our conception of resource-strapped individual agents).

It is Hintikka's contention that it is only through the providence of strategic rules that reasoning is truly ampliative [Hintikka, 1999a, p. 101]. And Peirce contends that whenever the objective of the reasoning involves new hypotheses, a move is ampliative only if it is abductive. Hintikka's proposed 'solution to the problem of abduction,' is that '[a]bductive 'inferences' must be construed as answers to the enquirer's explicit or (usually) tacit question put to some definite source of answer (information)' [Hintikka, 1999a, p. 102]. Note here the simi-

[16] And: *Deduction* is reasoning which proposes to pursue such a method that if the premisses are true the conclusion will in every case be true [Eisele, 1985, p. 37].

larity of question-generating diagnostic models, for example, KMS.HT. (See Peng and Reggia [1990, pp. 40–41], discussed in the next chapter.) Hintikka discerns this method of reasoning in the Socratic *elenchus*, and in much later writers such as R.G. Collingwood [1946] and Hans Gadamer [1975].

It is clear that the interrogative element of abduction is explicitly recognized by Peirce.

> The first starting of a hypothesis and the entertaining of it, whether as a simple interrogation or with any degree of confidence, is an inferential step which I propose to call abduction [Peirce, 1931–1958, p. 6.525].

Then, as we have seen,

> This will include a preference for any one hypothesis over others which would equally explain the facts, as long as this preference is not based upon any previous knowledge bearing upon the truth of the hypothesis, nor [sic] on any testing of any of the hypotheses, after having admitted them on probation. I call such inference by the peculiar name, abduction, because its legitimacy depends on altogether different principles from those of other kinds of inference.

In addition,

> It is to be remarked that, in pure abduction, it can never be justifiable to accept the hypothesis otherwise than as an interrogation. But as long as that condition is observed no positive falsity is to be feared [Buchler, 1955, p. 154].

Peirce did not explicitly identify an abductive inference with a question-answer step in an interrogation. Consequently we lack Peircean answers to fundamental questions. Can we say in a grounded way what the best questions are in such an enquiry? How is one to select the best questions from a set of possible questions?

In attempting to answer these questions, Hintikka proposes abandoning the idea that abduction is inference of any kind. Rather, he says, 'Abduction should be conceptualized as a question-answer step, not as an inference in any literal sense of the word' [Hintikka, 1999a, p. 105]. A virtue of the suggestion is that it entirely vindicates Peirce in his claim that abduction is 'altogether different' from deduction and induction.

An accessible treatment of Hintikka's model of question-answer enquiry is Hintikka and Bachman [1991], Hintikka and Halonen [1995], as well as the aforementioned [2002]. A similar approach is developed in the erotetic logic of Kuipers and Wisniewski [1994].

Hintikka's approach offers a number of attractions. It is surely correct to note that abductive triggers can be likened to questions. Hintikka is also right to see the limitations of definatory rules and the more central importance of strategic rules. Strategic rules are paradigms of the cognitive tools employed by beings like us, that is to say, by practical agents. Hintikka sees that sound reasoning is often a matter of backing procedures that have a good record, short of truth-preservation and probability-enhancement. There are proceeds that focus on the importance of the characteristicness and of common knowledge. In Hintikka's scheme, the characteristic and the commonplace are oracles, where oracles are the source of answers to questions. To the extent that Hintikka's abducer favours strategic rather than definatory rules, the ignorance condition is, if not outright acknowledged, then tacitly present. If the question put to the abducer is a request for knowledge about some matter P, and if the answerer (the abducer) draws his response from the oracle of characteristicness, then the answer he gives may well not give the questioner the knowledge he asked for; but this doesn't stop the respondent from having given him as an adequate answer. Since in the abductive model, questioner and answerer are one and the same, Hintikks's model preserves the satisficing and transformational traits of the GW-schema. Wanting epistemic attainment of T, the GW-abducer settles for presumptive attainment of it. What is more, Hintikka's definatory rules are sometimes the wrong rules to use, and abduction furnishes a context in which their employment is inappropriate. For abduction aims at neither truth-preservation nor probability-enhancement, as Peirce saw so well.

A further virtue of Hintikka's approach to abduction is, as we had earlier occasion to remark its express recognition of the point that abduction is not intrinsically explanationist. A related attraction is that strategic rules carry risks. Even when employing a mode of reasoning that seems generally appropriate in circumstances similar to those of the case at hand, it can fail in certain respects. In abductive contexts, one of the ways in which such a failure might express itself is in the generation of a propositionally implausible H. Hintikka recognizes that this would not discredit the decision to use the rule. We would add: Neither would it necessarily discredit the choice of H.

5.10 Empirical Progress

An interesting approach to abduction is Theo Kuipers' theory of empirical progress [Kuipers, 1999]. While explicitly explanationist in character, it also accommodates propositionally implausible hypotheses. So it is appropriate to consider it here. Suppose that an agent's target T is to produce a revision $K(H)$ of a given theory K under the condition that $K(H)$ represents a degree of empirical progress over K. Let us also take it that a necessary condition on the attainment of T is that $K(H)$ exceeds the explanatory power (or reach) of K and that in so doing it is no worse off than K on the score of observational anomalies. In other words, the agent's goal is to find a H such that the explanatory improvements afforded by $K(H)$ are not offset by observational degradation. Given these facts, we may assume that our agent infers H on the grounds that it engenders the requisite $K(H)$.

The agent's reasoning is both backwards-chaining and explanationist. But it is not successfully abductive. To see why, it is necessary to observe that our agent is performing not one but two cognitive tasks and that they are linked. In the first instance, he infers H. In the second, he infers that $K(H)$ is a better theory than K, and does so only because he has the required confidence in his inference to H. If his inference of H is abductive, then H is cognitively junior to K. Accordingly, so too is $K(H)$. But if $K(H)$ is cognitively junior to K, how can $K(H)$ be better science than K? Accordingly,

Proposition 5.13 (Kuiper's Dilemma) *Either the backward chaining reflected Kuipers revision of K to $K(H)$ is not abductive or it is not successful.*

We see, then, that

Proposition 5.14 (Backwards Chaining) *Backwards chaining reasoning is not intrinsically abductive.*

5.11 Semantic Tableaux

Semantic tableaux constitute a refutation method for designated formal languages [Hintikka, 1955; Beth, 1969]. An attractive exposition is that of Smullyan [1968]. The basic method has been adapted to abductive tasks by Aliseda-Llera [1997] and [Aliseda, forthcoming]. She writes,

> To test if a formula Φ follows from a set of premisses Θ, a *tableau* for the sentence $\Theta \cup \{\neg\Phi\}$ is constructed. The tableau itself is a binary tree built from its initial set of sentences by using rules for each of the logical connectives that specify how the tree branches. If the tableau closes, the initial set is unsatisfiable and the entailment $\Theta \models \Phi$ holds. Otherwise, if the resulting tableau has open branches, the formula Φ is not a valid consequence of Θ. A tableau closes if every branch contains an atomic formula β and its negation [1997, p. 83].

According to the tableau rules, double negations are suppressed, conjunctions add both conjuncts; negated conjunctions branch to two negated conjuncts; disjunctions branch into two disjuncts; negated disjunctions add both negated disjuncts; and conditionals reduce via negation and disjunction. Accordingly, *Negation*

$\neg\neg\Phi \rightarrow \Phi$

Conjunction

$\Phi \wedge \Psi \rightarrow \frac{\Phi}{\Psi}$

$\neg(\Phi \wedge \Psi) \rightarrow \neg\Phi \mid \neg\Psi$

Disjunction

$\Phi \vee \Psi \rightarrow \Phi \mid \Psi$

$\neg(\Phi \vee \Psi) \rightarrow \frac{\neg\Phi}{\neg\Psi}$

Conditional

$\Phi \rightarrow \Psi \rightarrow \neg\Phi \mid \Psi$

$\neg(\Phi \rightarrow \Psi) \rightarrow \frac{\Phi}{\neg\Psi}$

In languages in which all wffs have either disjunctive or conjunctive normal forms, the above rules reduce to one or other of two types of rule, a conjunctive (α-type) rule, and a disjunctive (β-type rule).

Rule 1. $\alpha \rightarrow \frac{\alpha_1}{\alpha_2}$

Rule 2. $\beta \rightarrow \beta_1 \mid \beta_2$

Suppose that $T(\Theta)$ is a tableau for a theory Θ. Then it is known that

a. If $T(\Theta)$ has open branches, Θ is consistent. An open branch represents a verifying model for Θ.

b. If $T(\Theta)$ has only closed branches, Θ is inconsistent.

c. Semantic tableau constitute a sound a complete system for truth functional languages.

d. A branch of tableau is closed if it contains some formula and its negation.

e. A branch is open if it is not closed.

f. A branch B of a tableau is closed if (recall rules 1 and 2 just above) for every α occurring in B both α_1 and α_2 occur in B, and for every β occurring in B, at least one of β_1, β_2 occurs in B.

g. A tableau is completed if every branch is either closed or complete.

h. A proof of a wff Φ is a closed tableau for $\ulcorner\neg\Phi\urcorner$.

i. A proof of $\Theta \models \Phi$ is a closed tableau for $\Theta \cup \{\neg\Phi\}$.

Aliseda construes abductions as a kind of extended tableau. An extended tableau is the result of adding new formulas to a tableau. Consider the case in which for some theory Θ, Φ is not a valid consequence. In the tableau for $\Theta \vDash \{\Phi\}$ there are open branches. Each such branch is a counterexample to the claim that Φ is a consequence of Θ. Of course, if further wffs were added to the open branches it is possible that they might now close. Consider a least class of wffs that fulfill this function. Then Φ would be derivable from a minimal extension of Θ. Finding such wffs is a kind of abduction.

Accordingly, Aliseda proposes [1997, p. 91] the following conditions.

> *Plain Abduction* $T((\Theta) \cup \{\neg\Phi\} \cup \{\alpha\})$ is closed. (Hence $\Theta, \alpha \vDash \Phi$.)
> *Consistent Abduction* Plain abduction $+T(\Theta \cup \{\alpha\})$ is open. (Hence $\Theta \nvDash \Phi$.)
> *Explanatory Abduction* Plain abduction $+T(\Theta \cup \{\neg\Phi\})$ is open (hence $\Theta \nvDash \Phi$) and $T(\{\alpha\} \cup \{\neg\Phi\})$ is open (hence $\alpha \nvDash \Phi$).

Further restrictions are required. Added wffs must be in the vocabulary of Θ. Each such wff must be either a literal, a non-repeating conjunction of literals or a non-repeating disjunction of literals.

Given a Θ and a Φ, *plain* abductive explanations are those wffs that close the open branches Γ of $T(\Theta \cup \{\neg\Phi\})$. *Consistent* abductive explanations are subsets of wffs which close some but not all open branches Γ of $T(\Theta)$.

A *total closure* of a tableau is the set of all literals that close each open branch Γ. A *partial closure* of a tableau is the set of those literals that close some but not all open branches Γ. (Such closures are definable for both branches and tableau.)

The *negation* of a literal is its ordinary negation of atomic or the embedded atom if negative. Thus $\neg \pm \Phi = \neg\Phi$ or Φ.

We consider now an algorithm for computing plain abductions [Aliseda-LLera, 1997, pp. 102–103].

Input

> A set of wffs representing theory Θ. A literal Φ representing the fact to be explained.
> Preconditions: $\Theta \nvDash \{\Phi\}, \Theta \nvDash \{\neg\Phi\}$.

Output

> Generates the set of abductive explanations $\alpha_1, \ldots, \alpha_n$ such that $T((\Theta \cup \{\neg\Phi\}) \cup \{\alpha_i\})$ is closed and the α_i satisfy the previously mentioned lexical and syntactic restrictions.
>
> *Procedures*
>
> > Calculate $\Theta + \neg\Phi = \{\Gamma_1, \ldots, \Gamma_n\}$. Select those Γ_i that are open branches

Atomic plain explanations

1. Compute $TTC(\Gamma_1, \ldots, \Gamma_n) = \{\gamma_1, \ldots, \gamma_m\}$ = the total tableau closure = the set of literals which close all branches concurrently.
2. $\{\gamma_1, \ldots, \gamma_m\}$ is the set of plain abductions.

Conjunctive plain explanations

1. For each open branch Γ_i construct its partial closure $BPC(\Gamma_i)$ = the set of literals that close that branch but do not close any of the other open branches.
2. Determine whether all Γ_i have a partial closure. Otherwise there is no conjunctive solution (hence go to END).
3. Each $BPC(\Gamma_i)$ contains those literals that partially close the tableau. To construct a conjunctive explanation take a single literal of each $BPC(\Gamma_i)$ and form their conjunction.
4. For each conjunctive solution β delete repetitions. The solutions in conjunctive form is $\beta_1, \ldots, \beta_h$.
5. END.

Disjunctive plain explanations

1. combine atomic with atomic explanations, conjunctive with conjunctive explanations, conjunctive with atomic, and each atomic and conjunctive with Φ.
2. Example. Construct pairs from the set of atomic explanations, and form their disjunctions $(\gamma_i \vee \gamma_j)$.
3. Example. For each atomic explanation, form a disjunction with Φ (viz. $(\gamma_i \vee \Phi)$).
4. The result of all such combinations is the set of explanations in disjunctive form.
5. END.

In the interest of space, we omit the specification of algorithms for constructing consistent abductive explanation. The interested reader should consult [Aliseda-LLera, 1997, pp. 103–106].

There are attractions to Aliseda's approach to abduction, not least of which is its employment of the well-understood machinery of semantic tableaux. Another is that it extends in a natural way to accommodate (most of) the structural features of Theo Kuipers' theory of empirical progress [Aliseda, forthcoming, ch. 6] and [Kuipers, 2000, p. 112]. Let us say that at a given time t a theory Θ_2 is at least as successful as a theory Θ_1 if and only if the set of failures in Θ_1; that the set of successes in Θ_1 is a subset of the set of successes in Θ_2; and at least one of these subsets is proper. Intuitively a theory Θ has a success with respect so some data E, and some background conditions K, if $K \cup \{\Theta\} \vDash E$; and that it has a failure with respect to these same parameters if $K \cup \{\Theta\} \vDash \neg E$. We may read F as confirmation. For ease of reference, we follow Kuipers in characterizing these cases as a success (failure) by E of Θ relative to C. Clearly we also have it that E may sometimes be such that it is neither a success nor a failure of Θ relative to C, but rather is a *lacuna* of H with respect to C.

Progress in science is a matter of moving from given theories to better ones. Roughly speaking, a theory Θ_2 is better than a theory Θ_1, if Θ_2 it has more successes than Θ_1 One standard way of achieving this kind of empirical progress is by revising Θ_1 in ways that produce Θ_2. Kuipers [1999] sets out the following

> *Instrumentalist abduction task*: Search for a revision Θ_2 of a theory Θ_1 such that H_2 is more successful than Θ_1 having regard to the available data.

Aliseda approaches this task by showing that the concepts of lacuna, success and failure admit of precise characterization in the language of semantic tableaux. Where $\mathcal{T}(\Theta)$ is a tableau for Θ, then if $\mathcal{T}(\Theta) + \{\neg E\}$ and $\mathcal{T}(\Theta) + \{E\}$ are both open extensions of $\mathcal{T}(\Theta)$, then E is a lacuna of Θ. Similarly, if E is a success of Θ, then $\mathcal{T}(\Theta) = \{\neg E\}$ is a semi-closed extension and there is an initial condition K such that $\mathcal{T}(\Theta) + \{K\}$ is a semi-closed extension (Similarly for failure, putting E for $\neg E$). A standard way of achieving empirical progress is to convert a theory for which E is lacuna into a theory for which E is a success. This involves finding a set of wffs H to add to the theory such that $K \cup \Theta \cup \{H\}$ is a revision of Θ relative to K for which E is now a success. This particular transformation is achieved when $K \cup \Theta \cup \{H\}$ confirms neither E nor $\neg E$.

5.11.1 Assessing Semantic Tableau Abduction

Semantic tableau abduction considerably resembles enthymeme resolution, (concerning which, see chapter 9). The more abstractly formal its presentation, the closer the similarity is. In enthymeme resolution is to find a P that closes a $\vDash$-connection between some premisses and a conclusion, where $\vDash$ is read as deductive consequence. In semantic tableau abduction, the task is to find a P that closes a $\vDash$-connection between a theory and some empirical data. Although the general theory doesn't require it, Θ in its adaptation to the ends of Kuipers' theory of empirical progress, $\vDash$ may be read as a confirmation relation. What this shows us is the importance of the interpretation given to $\vDash$. Treated as *any* consequence relation, as in Aliseda's core theory, there is nothing abductive about the closure of $\vDash$-connections. Treated as confirmation, things aren't quite so bleak.

So, then, what *do* we make of a situation in which adding P to Θ closes the confirmation-connection to some data E? Is this not some reason not only to send P to trial, but also to make the conjecture that P is true? The answer depends in part on the size and importance of E and of the collateral costs (if any) that attach to conjecturing P's truth. But let us concede what is plain to see. In ranges of cases, this kind of confirmation is reason, albeit sometimes modest reason, to make the conjecture. But this is not abduction either. Justified conjecturability is part of what is required, and we may suppose that we have it for large classes of cases. What is also required is ignorance, and nowhere is that factor addressed in semantic tableau abduction. Of course, it is a requirement of abductions of this sort that the P that turns the trick not be derivable in the original theory Θ. But it is not ruled out that P be true or that it be known to be true, then it is eligible for consideration by semantic tableau theorists even so. For it suffices to add it to Θ as a new axiom. Suppose we do. Where now is the case that P's role in closing the $\vDash$-connection to E offers reasons for conjecturing that P is true. If P is known to be true, there is no room for any such conjecture; and anything purporting to constitute grounds for conjecture will have turned out to be epistemically defective. It hardly needs saying that needed repairs are not hard to make. One might require, for example, that any P selected in a semantic tableau exercise not have a degree of epistemic virtue equal or greater to that evinced by the original theory Θ. Still, there are two admonitions that require gentle sounding. The first we've already met with. It is

Proposition 5.15 (Arbitrary interpretations of $\vDash$) *The freer the range of interpretations of $\vDash$ in semantic tableau abductions, the less genuinely abductive they are.*

Corollary 5.15(a) *Apart from that, in semantic tableau abductions, $\Vdash$ can bear only those interpretations that answer to the relevant closure conditions. In particular, $\Vdash$ cannot be interpreted as* plausible consequence, *since plausible consequence is not closed under negation.*

5.11.2 Is It Abduction?

In our discussion in this section we have been assuming, with Aliseda, that Kuipers' theory reveals empirical progress to have an abductive character. Armed with that assumption, we have been considering Aliseda's interesting attempt to elucidate the abductive structure of empirical progress by adapting Kuipers' account to her own semantic tableau account. It is essential to the success of her project that she be right in making this assumption. Given our discussion of empirical progress in the previous section, we are unable to agree that the assumption is sound. According to what we called *Kuipers' Dilemma*, empirical progress inferences are either not abductive or not sound. Accordingly, we are skeptical of Aliseda semantic tableau accommodation of Kuipers as a contribution to the logic of abduction.

5.12 Inconsistency Again

In Chapter 3 we considered and rejected the standard positions concerning the consistency of K. As it happens, there are three questions about abductive consistency that need to be settled.

1. Should we require that K be consistent?
2. Should we require that $K(H)$ be consistent?
3. Should we require that H be (self-)consistent?

The position taken in chapter 3 answers question (1) in the negative but addresses neither of the remaining questions. We shall attempt to repair those omissions now.

Question (3) asks, in effect, whether we should maintain a strong anti-dialetheic stance toward winning abductive hypotheses. Dialetheism asserts that some but not all contradictions are both true and false. While we have no particular stake in dialetheism's being true, are not prepared to dismiss it out of hand [17] Consider a case. It is all but universally held that the proof of the Liar is a paradox. A paradox is a proof that appears to be sound, but, owing to the transparent falsity of the conclusion, cannot be taken as sound. The problem created by such proofs is that, although they demonstrate the falsity of something in or presupposed by the premiss-set, it is left wholly undetermined as to where in particular to pin the blame. Dialetheism offers a different solution to such puzzles. What best explains the appearance that the premisses are true is that they are true. What best explains the appearance that the proof is valid is that it is valid. What best explains the appearance that the conclusion is false is that it is false. What best explains why these explanations themselves pairwise consistent is that the conclusion is also true.

While we do not ourselves endorse this solution, we fail to see how we would advance the analysis of abduction by ruling it out. In as much as the present solution is itself an abductive solution: it is perhaps appropriate that we leave official room for winning Hs to be self-inconsistent. Accordingly,

Proposition 5.16 (Self-inconsistency) *Not only may a winning H be propositionally implausible, it may also be self-inconsistent.*

Question (2) asks whether we should make it a condition on H that it be consistent with K. Since we already have it that K itself might be inconsistent, the

[17] On the pros and cons of the idea of true contradictions, [Woods, 2005a] and [Armour-Garb, 2004].

issue raised by (2), is whether to tolerate additional inconsistencies occasioned by the conversion of K to $K(H)$. If we were treating inconsistency classically, there would be trouble with the idea of "additional" inconsistencies. But since the toleration of K's negation-inconsistency requires that K be lodged in a paraconsistent logic, $K(H)$ can have more inconsistencies than K, since the negation-inconsistency of neither lands it in absolute inconsistency. So let us consider question (2).

Let H be a winning abductive hypothesis, and let the negation of H be a member of K. Should we prohibit this? The answer is that we should not. We should allow for the case in which, although to the best of our knowledge that P, we nevertheless conjecture that not-P. This we may do provided that membership in K is not required to meet the KK condition. Accordingly, any K at time t is what is taken to be known by an agent (or by the community of knowers) at t. Fallibilism is a discouragement of the KK-hypothesis, even if it is not strictly incompatible with it. The KK-hypothesis opines the impossibility of knowing anything, without knowing that you know it. The more realistic assumption is that often someone might know something without realizing it. This is a natural precursor to fallibilism, the view that anything (or most things) we take for knowledge, we might be mistaken about. By these lights, conjecturing that not-P is compatible with our taking P as known. This is not to say that we have *carte blanche* in such matters. In abducing H, we decide to give it a premissory role in future reasons within or from $K(H)$. Correspondingly, we diminish $\neg H$'s premissory range, notwithstanding that $\neg H \in K$. Even though we would make paraconsistent provision for their joint use as premisses, there are lots of cases in which their joint use would contaminate the reasoning in question. So, $K(H)$ should be attended by the appropriate quarantines.[18] Again, think of the historical origins of Anderson-Belnap relevance. In their attempt in 1959 to give a simple treatment of the truth functions, it happened that Disjunctive Syllogism (DS) is not a valid rule in their treatment. This was a natural occasion to conjecture the actual invalidity of DS while acknowledging that the validity of DS was in K. The unseemly haste with which Anderson and Belnap abandoned DS is an interesting demonstration of how quickly a conjecture that P can be promoted to full membership in a successor to K. But the point remains that it is perfectly tenable to conjecture the opposite of what one takes for knowledge.

What this shows is the importance of the belief/acceptance distinction to abductive logic. While a conjectured hypothesis might be something the agent believes, conjecturing it does not confer that status upon it. Conjecture is acceptance. It is acceptance for premissory work in future inferences, subject to the possibility

[18] See here [Batens, 1980; Arruda, 1989; DaCosta and Bueno, 1996; Jennings and Schotch, 1981; Priest, 2002; Detlefsen, 1986; Jáskowski, 1948; Woods, 2003; Gabbay and Hunter, 1993; Tanaka, 2003].

of recall. Another way of saying this is that conjecture does not report a doxastic state. Rather it expresses a decision. From this we have it that

(1) P, but we conjecture that not-P

does not have the speech-act structure of *blindspot utterances* of the form

(2) P, but I believe that not-P.

The trouble with (2) is that, in the absence of additional information, it is impossible to determine what the utterer's position to P actually is. There is no such problem with (1). The utterer's position is that P is the case and nevertheless that there are adequate reasons to release not-P for properly regulated premissory work in future inferences.

In sum,

Proposition 5.17 (Paraconsistency and dialetheism) *K can be negation consistent but not absolutely inconsistent. The same holds of $K(H)$. This occasions the necessity to lodge abductive reasoning in a paraconsistent base logic. H itself can also be inconsistent. This requires that the underlying logic also have a dialetheic component.*

5.12.1 Bayesian Inference

Bayes' theorem charts the difference between the prior probability of a hypothesis, $P(h)$ and the conditional probability of that hypothesis on the available evidence, $P(h|e)$. The connection is described by the following equation

$$P(h|e) = \frac{P(e|h) \times P(h)}{P(e)} \tag{5.3}$$

in which $P(e|h)$ is the inverse probability of $(h|e)$ and $P(e)$ the probability of e taken alone. Accordingly in a Bayesian approach to the evaluation of probabilistic reasoning the better way of evaluating a hypothesis h is to replace its prior probability $P(h)$ with its conditional probability $P'(h) = P(h|e)$. This is called *Bayesian conditionalization*, Bayesian conditionalization gives us a way to conceive of inductive confirmation. On this view, the confirmation a given h receives from a body of evidence $e(S(h|\frac{h}{e}))$ is its conditional probability on e minus its prior probability (provided that it is not zero) [Howson and Urbach, 1993, p. 117]. Another way of understanding confirmation or evidential force involves a Bayes factor, so-called in [Good, 1983]. A Bayes factor is a likelihood ratio λ defined as

$$\lambda(e|h) = \frac{P(e|h)}{P(e|\ h)} \tag{5.4}$$

It is easily shown that the likelihood ratio exceeds one if and only if the degree of support given h by e exceeds zero. So $\lambda(e|h)$ is interpretable as a degree of evidential support when embedded in the odds-likelihood version of Bayes' theorem:

$$P(h) = P(h)|P(h), \tag{5.5}$$

evidential support construed as changes in odds captures desired changes in probabilities as well as the idea expressed by $S(h|e)$, never mind that $\lambda(e|h)$ is also more easily calculated than $S(h|e)$. Since confirmation theory is mainly involved with the conditions under which it is reasonable to change one's beliefs, $\lambda(e|h)$ has a natural role to play in stating these conditions. Howson [2000] performs the valuable service of showing that conditionalization

$$\frac{P(B|A) = r, P'(A) = 1}{P'(B) = r} \tag{5.6}$$

does not hold unless we also have it that

$$P'(B|A) = P(B|A) \tag{5.7}$$

Howson has shown, however, that there are cases in which subscription to $P'(B|A) = P(B|A)$ is inconsistent [2000], but apart from that it is extremely implausible that conditionalization would hold fast in all contexts. On the other hand, there are ranges of cases for which just such a standfast assumption about conditionalization is justified. This will be so whenever the determination e of an experimental result would not by itself induce us to change the probability of its discovery conditional upon any h. By and large, there is not much scope, simply in determining the particular experimental result e, to revise $P(e|h)$. But, let us be clear. The conditionalization equation sometimes fails and nothing in Bayesian theory tells us when or why. Accordingly,

Proposition 5.18 (Changing conditionalization) *A Bayesian theory of evidential reasoning requires the supplementation of a theory of belief revision explaining the conditions under which conditionalization is caused to lapse.*

A standard complaint against Bayesianism is its apparent inability to say how prior probabilities originate. Except for situations involving gaming devices (e.g., decks of cards, coins and dice), there is truth in these complaints. The truth is that for lots of cases which seem intuitively amenable to probabilistic reasoning, there is no simple or direct way of determining priors. It is hardly surprising therefore that some researchers try to model probabilistic reasoning in a manner that doesn't require the invocation of priors. Standard statistical inferences — for example, significance tests and confidence interval determinations — do not require the specification of priors. Even so, any such inference can be re-expressed as a Bayesian inference and, with the aid of Bayes' theorem in reverse, the prior probabilities can in fact be determined. This is a welcome turn. Sequences of these non-Bayesian inferences are either manifestly incoherent or they determine a prior probability distribution. This constitutes an interesting kind of answer to those who complain that, gaming devices aside, Bayesian methods do not suffice for the determination of priors. For they may now be challenged to say where the incoherence of such inferences lies.

Granted that the Bayesian approach to rational theories of belief dynamics is still a work in progress, it is sometimes asserted, even so, that Bayesian insights offer needed clarity to probabilistic versions of inference to the best explanation. Such inferences have the gross form:

$$\begin{array}{c} P(e|h) \text{ is high} \\ P(e|\ h) \text{ is low} \\ e \text{ is a fact} \\ \hline \text{Therefore } P'(h) \text{ is high} \end{array}$$

Embedded in the schema are the factors of likelihood and priority, both of which respond well to Bayesian construal.

This is too fast by far. This can be seen if we restate the inference above informally, as follows:

1. e is a fact.
2. h makes e's probability higher that it would be if taken alone.
3. h makes e's probability lower than it would be if taken alone.
4. Therefore, e makes h's probability high.

Perhaps the most obvious thing to take note of here is that whereas the argument's conclusion asserts that h's probability is made high by fact e, there is nothing in that state of affairs that constitutes an explanation of e. In inference to the best explanation, it is e, not h, that plays the role of explicandum. So the present inference is not inference to the best explanation. The point generalizes. The inference at hand could not be a *non*-explanationist abduction. For, again, the objective of an abduction is to ground the conjecturing of a hypothesis in a relationship it bears to some fact or state of affairs (or a proposition asserting that fact or state of affairs). But in the example before us these factors are nowhere in view. The conclusion has it that e confers high probability on h, whereas in a genuine abduction, the conclusion can only be that h's bearing some kind of consequence relation to e, constitutes nothing more than reason to conjecture h. In the present case, the form of the conclusion is wrong for abduction, and what it asserts about h is also wrong

[18] Assuming also that the prior $P(h)$ is not too low.

for abduction. For what it says about h is (given that e is a *fact* that h is highly probable). But it would be an oddly conservative epistemology that required us to be no more forthcoming about high probabilities than to assert that they are fit candidates for conjecture. In effect, these complaints are Peirce's own. Abduction he admonishes us, is not a matter of fixing probabilities (especially, high ones).

Proposition 5.19 (Bayesian abduction) *Bayesian inference is not abductive.*

This Page is Intentionally Left Blank

Chapter 6

Diagnostic Abduction in AI

> *Patient*: Doctor, it hurts like blazes if I do this (extending his arm upwards).
>
> *Doctor*: Well, my advice is: Don't do that. That will be $500.00, please.
>
> Groucho Marx

6.1 Explanationist Diagnostics

One of the environments in which abduction is most at home is diagnostics. In its general form, the diagnostician's task is to match a disorder to an array of symptoms. In so doing, he or she is often faced with an unwanted abundance at both ends of the process. At the beginning there is a plurality of candidate disorders. And, although the diagnostician strives to shrink this plurality to one, it cannot be excluded that the end of the process a plurality may still exist, albeit a smaller one. Like all abducers, the diagnostician is faced with the Cut Down Problem. He also faces issues relating to the process-product distinction,which, as before, presents the abductive logician with a problem in the engagement-sublogic. The problem is that, as for any abduction problem, even if it is justified to postulate the existence of a filtration structure in which abductive solutions are cutdowns of up to very large possibility spaces, there is no empirical evidence that real-life abducers achieve their abductive targets by constructing such structures. Apart from these quite general features, diagnostics liberally instantiates a distinction of importance for the logic of hypothesis-generation. On the one hand, it is often the case that all the candidate disorders are known in advance, obviating the need for generation. In such cases, the abductive task is to pick a candidate from a known field each of whose members is underdetermined by the symptoms to which the diagnostician has access. On the other hand, there are cases in which the symptoms are more radically mysterious, since they present themselves in the absence of any candidate disorders.

In this chapter we shall briefly review some representative treatments of diagnostic abduction by AI researchers. Josephson and Josephson take an explanationist tack, whereas Peng and Reggia's[1] approach integrates explanationist and probabalistic elements. In a further section, we shall make some effort to adjudicate the rivalry between explanationism and probabilism.

Much of the recent — and most promising work — has been produced by computer scientists and by logicians who also work in the AI research programme. Most of the contemporary work to date tends to concentrate on hypothesis generation and engagement. Accordingly, this chapter is a contribution to the relevantly associated sublogics. We begin by reviewing a representative system for dealing with a class of problems for which hypothesis-generation and hypothesis-engagement are an efficient way of finding the best solutions (or explanations). The system in question is that of Josephson and Josephson [1994, ch. 7].[2] We should be careful about what we intend by the representativeness claim. Josephson and Josephson [1994] is representative of a the explanationist approach to formal diagnostics. It is an approach that meets with daunting complexity problems. The work is more representative in the former respect than the latter; but it is well to emphasize early in the proceedings the general problem posed by complexity in formal reconstructions of real-life human performance.

Josephson's and Josephson's is an approach with some notable antecedents; e.g., Miller, Pople and Meyers [1982], [Pearl, 1987], de Kleer and Williams [1987], [Reiter, 1987], Dvorak and Kuipers [1989], Struss and Dressler [1989], Peng and Reggia [1990], and [Cooper, 1990], among others.

We here adopt the notational conventions of Josephson and Josephson [1994]. d denotes a fact or a datum; D a class of data; h denotes a particular hypothesis and H a class of these. H itself can be considered a composite hypothesis. An *abduction problem* is an ordered quadruple, $\langle D_{all}, H_{all}, e, pl\rangle$. D_{all} is a finite set of the totality of data to be explained; H_{all} is a finite set of individual hypotheses; e is a function from subsets of H_{all} to subsets of D_{all} (intuitively H *explains* $e(H)$); and pl is a function from subsets of H_{all} to a partially ordered set (intuitively, H has *plausibility* $pl(H)$). In this structure the requirement that unit values of pl be partially ordered leaves it underdetermined as to whether $pl(H)$ is "a probability, a measure of belief, a fuzzy value, a degree of fit, or a symbolic likelihood" [Joseph-

[1] Other important, indeed, seminal contributions to probabilistic AI are Pearl[1988; 2000].

[2] The authors of Josephson and Josephson's chapter 7 are Tom Bylander, Dean Allemang, Michael C. Tanner and John R. Josephson.

son and Josephson, 1994, p. 160].[3] **AP** is a logic which aims at a solution of an abduction problem.

H is complete if it explains all the data; i.e., $e(H) = D_{all}$. H is thrifty if the data explained by H are not explained by any proper subset of it; i.e., $\neg\exists H \subset H(e(H) \subseteq e(H'))$. If H is both complete and thrifty then it is an *explanation*. An explanation H is a *best* explanation if there is no other more plausible explanation; i.e., $\neg\exists H'(pl(H') > pl(H))$. Note that since pl gives only partial orderings, there can be more than one best explanation of a d or a D. A solution to an AP both resembles and yet differs from our earlier conception of a filtration structure. Similar questions also arise. Perhaps, most importantly, is the issue of whether practical agents actually construct such solutions in their own successful abductions on the ground. The system **AP** was designed to model Reiter's account of diagnosis [1987] and Pearl's approach to belief-revision [1987]. We briefly sketch the connection with Reiter's theory.

Reiter on Diagnosis: A diagnosis problem is an ordered triple ⟨SD, COMPONENTS, OBS⟩. SD is a finite set of first-order sentences which describe the diagnostic problematic. OBS is a set of first-order sentences which report observations. COMPONENTS is a finite assortment of constants, of which ab is a one-place predicate meaning 'abnormal'. A *diagnosis* is a least set $\Delta \subseteq$ COMPONENTS such that SD $\cup$ OBS $\cup \{ab(c)/c \in \Delta\} \cup \{\neg ab(c)/c \in \text{COMPONENTS}\backslash\Delta\}$ is consistent. A diagnosis problem can be modelled in AP as follows:

$H_{all} =$ COMPONENTS
$D_{all} =$ OBS
$e(H) =$ a maximal set $D \subseteq D$ such that

$$\text{SD} \cup D \cup \text{ab}(h) \mid h \in H \cup \neg\text{ab}(h) \mid h \in H_{all}\backslash H$$

is a consistent set.

Diagnoses are unranked in Reiter's treatment; hence pl is not needed in the AP reconstruction.

In a good may respects **AP** is a simplified model. For example, both e and pl are assumed to be tractable, notwithstanding indications to the contrary [Reiter, 1987; Cooper, 1990]. Even so, intractability seems to be the inevitable outcome

[3] We give a formal model of abduction in Chapters 12 and 13. In this model, given elements $d \in D_{all}$, an abductive algorithm utilizing the proof theory Π of the logic will yield a family $\{H_d^i | i = 1, 2, \ldots\}$ of possible hypotheses which explain d. From such families it is easy to define the set H_{all} and a function e as described in the Josephson model. In this context, the ordering $pl(H)$ can be meaningfully defined from the logic involved and using the abductive algorithm available.

The Josephson model $\langle D_{all}, H_{all}, e, pl\rangle$ is no longer an abstract model but a derived entity from our abductive mechanisms. As a derived entity, better complexity bounds may be obtained, or at least its complexity can be reduced to that of the abductive algorithms.

above a certain level of interaction among the constituents of composite hypotheses.

We say that an abduction problem is *independent* when, should a composite hypothesis explain a datum, so too does a constituent hypothesis; i.e., $\forall H \subseteq H_{all}(e(H) = \cup_{h \in H} e(h))$. In systems such as **AP**, the business of selecting a best explanation resolves into two component tasks. The first task (1) is to find an explanation. The second task (2) is keep on finding better ones until a best is identified. A number of theorems bear on these two matters, the first seven on (1) and the next three on (2). In the interests of space, proofs are omitted.

Theorem 6.1 *In the class of* independent *abduction problems, computing the number of explanations is #P-Complete (i.e., as hard as calculating the number of solutions to any NP-complete problem).*

In the case of independent abduction problems, sub-task (1) is tractable. Hence we have

Theorem 6.2 *For independent abduction problems there exists an algorithm for specifying an explanation, if one exists. The algorithm's order of complexity is $O(nC_e + n^2)$, where C_e is the true complexity of e, and nC_e indicates n calls to e.*

An abduction problem is monotonic if a complete explanation explains at least as much as any of its constituent explorations; i.e., $\forall H, H' \subseteq H_{all}(H \subseteq H' \rightarrow e(H) \subseteq e(H'))$. Any independent abduction problem is monotonic, but a monotonic abduction problem need not be independent.

Theorem 6.3 *Given a class of explanations, it is, NP-complete to determine whether a further explanation exists in the class of monotonic abduction problems.*

Moreover

Theorem 6.4 *In the class on monotonic abduction problems there also exists an $O(nC_e + n^2)$ algorithm for specifying an explanation, provided there is one.*

Let h be a composite hypothesis. Then an *incompatibility abduction problem* exists with regard to h if h contains a pair of constituent hypotheses h^* and $\neg h^*$. More generally, an incompatibility abduction problem is an ordered quintuple $\langle D_{all}, H_{all}, e, pl, I \rangle$, where all elements but I are as before and I is a set of pairs of subsets of H_{all} which are incompatible with one another. We put it that $\forall H \subseteq H_{all}((\exists_i \in I(i \subseteq H)) \rightarrow e(H) = \emptyset)$. In other words, any composite hypothesis containing incompatible sub-hypotheses is *explanatorily inert*. Any such composite is at best trivially complete and never a best explanation. However, independent incompatibility problems are independent problems apart from the factor of incompatibility. In other words, they fulfill the following condition:

$$\forall H \subseteq H_{all}((\neg\exists_i \in I(i \subseteq H)) \to e(H) = \bigcup_{h \in H} e(h)).$$

Incompatibility abduction problems are less tractable than monotonic or independent abduction problems.

Theorem 6.5 *In the class of independent incompatibility abduction problems it is NP-complete to find whether an explanation exists.*

From this it follows that it is also NP-hard to determine a best explanation in the class of independent incompatibility abduction problems. This class of problems can be reduced to Reiter's diagnostic theory. In fact,

Theorem 6.6 *In the class of diagnosis problems, it is NP-complete to determine whether a diagnosis exists, depending on the complexity of deciding whether a composite hypothesis is consistent with SD ∪ OBS. (Here, a composite hypothesis is a conjecture that certain components are abnormal and the remainder are normal.)*

Independent and monotonic abduction problems are each resistant to cancellation. One hypothesis cancels another if and to the extent that its acceptance requires the loss of at least some of the explanatory force the other would have had otherwise. A cancellation abduction problem can be likened to an ordered sextuple $\langle D_{all}, H_{all}, e, pl, e, e\rangle$, of which the first four elements are as before, and e is a function from H_{all} to subsets of D_{all}, indicating the data 'required' by each hypothesis. This gives an extremely simplified notion of cancellation. Nevertheless, it is enough to be a poison-pill for them all. It suffices, that is to say, for interpretability.

Theorem 6.7 *It is NP-complete to ascertain in the class of cancellation abduction problems whether an explanation exists.*

It follows from Theorem 6.7 that it is NP-hard to find a best explanation in this class of problems.

Thrift is also an elusive property, notwithstanding its methodological (and psychological) desirability. Indeed, it is as hard to find whether a composite hypothesis is thrifty in the class of cancellation abduction problems as it is to determine whether an explanation exists. In other word, both tasks are co-NP-complete in this class of problems.

Up to this point, our theorems address the problem of whether an explanation exists in various problem classes we have been reviewing. The remaining theorems bear on the task of finding best explanations. In systems such as **AP**, finding a best explanation is a matter of comparing plausibilities. There is a *best-small* plausibility rule for this. The rule gives a comparison criterion for classes of hypotheses H and H'. The rule provides that these be a function from H to H' which matches elements of H to not less plausible elements of H'. If H and H' have the same cardinality, at least one element in H must be more plausible than its image in H' if H is to be counted more plausible than H'. On the other hand, if H is larger than H', it cannot be more plausible than H'. Whereupon

Theorem 6.8 *In the class of independent abduction problems, it is NP-hard to determine a best explanation, using the* best-small *plausibility rule.*

However, the *best-small* rule is tractable where the individual hypotheses have different plausibility values $\nu_i, \ldots, \nu_n$ and the ν_i are *totally ordered.* In that case,

Theorem 6.9 *In the class of totally ordered monotonic problems satisfying the* best-small *criterion, there exists an $O(nC_e + C_{pl} + n^2)$ algorithm for determining a best explanation.*

It follows that if there exists *just one* best explanation in the conditions described by Theorem 9, it will be found by an algorithm of the stated type. On the other hand, it is notoriously difficult to determine whether the 'just one' condition is indeed met. In fact,

Theorem 6.10 *In the class of totally ordered independent abduction problems deploying the best-small rule, if there exists a best explanation it is NP-complete to ascertain whether another best explanation also exists.*

These theorems establish that if we take abduction to be the determination of the most plausible composite hypothesis that is omni-explanatory, then the problem of making such determinations is generally intractable. Tractability, such as it may be, requires consistency, non-cancellation, monotonicity, orderedness and fidelity to the best-small rule. But even under these conditions the quest for the *most* plausible explanation is intractable. What is more, these difficulties inhere in the nature of abduction itself and must not be put down to representational distortion. One encouraging thing is known. If an abduction problem's correct answer-space is small, and if it is possible to increase the knowledge of the system with new information that eliminates *large* chunks of old information, then this reduces the complexity of explanation in a significant way; but it also threatens to eliminate the abductive character of explanations thus achieved. By and large, however, since there are no tractable algorithms for large assortments of abduction problems (the more psychologically real, the more intractable), most abductive-behaviour is *heuristic* (in the classical sense of the term; see chapter 11).

6.1.1 Difficulties with AP

AP finds a solution for problems in the form $\langle D_a ll, H_a ll, e, pl\rangle$. Except for a specific role recorded to factors of relevance, an **AP** instantiates the filtration-structures discussed in chapter 3. We've said repeatedly that the empirical record of human cognitive behaviour gives little encouragement to the idea that in solving real-life abduction problems, beings like us actually construct filtration-structures. What's true for the genus is likewise true for the species. Accordingly, we have it that

Proposition 6.11 (The non-execution of APs by practical reasoners) *In solving their real-life abduction problems, practical agents do not in the general case execute an* **AP**.

And, as before, we conjecture that

♡ **Proposition 6.12 (Complexity and non-execution)** *It is plausible to suppose, as with the example of filtration-structures, that an important part of why individual agents do not execute* **AP**s *is the computational complexity of such systems.*

We would do well not to be unduly alarmed by Propositiom 6.11, even assuming it to be true. What this proposition conjectures is that programs such as **AP** aren't realistic for beings like us to implement, and that they run up against limitations that inhere in the comparative paucity of an individual's cognitive resources and the comparative modesty of his (or its) cognitive goals. Virtually everything to date that has been offered as a logic of reasoning or as a model of cognitive performance outreaches descriptive adequacy in this same way. Judging by the empirical record, beings like us don't achieve their ends by implementing such logics or instantiating such models. By far the standard *apologiae* for such gaps is that, at its best, and to what actual behaviour approximates to what the logic stipulates the model sanctions. We have already said our piece about this standard answer. We have difficulties with it. One is that it leaves undealt with the question of how to select those principles laws that are given privileged place in the logic or in the model. The other is that there has been rather little in the way of successful accounts of the requisite approximation relation, even assuming a wholly satisfactory resolution of the first difficulty. But we should be clearer than we have been before what our reservation about approximation comes to. It is, therefore, *not* our view that approximations do not exist about what abductive individuals do and how abduction fares in the likes of an **AP**. or any other appropriately contrived filtration-structure. Moreover, it is *not* our view that such approximations are theoretically intractable. It *is* our view, however, that

Proposition 6.13 (Limits of approximation) *Such approximation relations as are presently within our present capacity to describe do not elucidate relevant details of how the actual abductive behaviour of individuals works in practice.*

It is certainly true that actual abductive behaviour more closely resembles what an **AP** provides than what goes on in a model of the internal combustion engines. It is at a close resemblance, in our view. But this is far from a blanket dismissal. There are a number of reasons for taking a conciliatory view.

♡ **Proposition 6.14 (Starting small)** *Since inadequate approximations themselves approximate to requisitely realistic approximations, the former may be indulged as attempts to achieve the latter.*[4]

♡ **Proposition 6.15 (Applicability to theoretical agents)** *To the extent that inadequate approximation is a matter of the resource constraints and goal modesty typical of practical agents, it may be conjectured that the actual behaviour of theoretical agents will have a better approximation fit with* **AP** *and filtration structures more generally.*

Corollary 6.15(a). *If Proposition 6.15 is true, then the great value of such structures is that they give accounts of abduction to which the real-world behaviour of actual agents* of a type *more satisfactorily approximates. Thus they describe real abduction, if not the real abductions of individuals.*

We said in our discussion of filtration-structures that there is reason to believe that in the process of solving an abduction, the set of possible candidates takes on the contours of a filtration-structure. This means that, whatever other details might be involved, winning hypotheses stand to original set of candidates in a complex relationship of the relevant-to-the-plausible-to-the most plausible. We can suppose this to be so quite independently of whether an individual abducer ever actually selects his winning hypothesis by activating these filters. Even so,

♡ **Proposition 6.16 (Filtration-structures as constraints)** *Whatever the manner in which an individual abducer actually selects his hypothesis from a set of candidate hypotheses, his* modus operandi *must honour the fact that the winning H has a determinate place in a filtration-structure.*

Approximative adequacy is one thing; explanationist adequacy is another. In chapter three, we argued that the nescience condition has the effect of restricting the type of explanation embedded in explanationist abductions to subjunctive explanations. Let us briefly recapitulate. An abduction problem is triggered when a certain target cannot be hit with the agent's present resources. In a loose and intuitive way, this means that the target cannot be reached on the basis of what the agent currently knows. In strictness, the lack-of-knowledge (or nescience) requirement is not a matter of sheer ignorance, but rather of the unattainability of the target

[4] See here [Simon, 1973, p. 327]: "If there is no such thing as a logical method of having new ideas, then there is no such thing as a logical method of having *small* new ideas".

with what the agent has a certain level k of knowledge of, or higher. Thus in an abduction problem nothing the agent knows at level k or higher enables him to reach his target. Accordingly, a conjecture is necessary. It is a conjecture in the form $C(H)$, i.e., proposition H is conjectured to have a degree of epistemic virtue consonant with the level-k requirement. This is a clearly necessary constraint. For if H is already known, but at a lower degree than k, or is a well-justified belief short of knowledge strictly speaking, then that H facilitates the hitting of a heretofore unhittable target would afford no occasion for the conjecture of H. (Why make what is already known or justifiably believed a matter of conjecture?). The role of conjecture is intimately connected to the ampliative character of abduction, concerning which Peirce gave such emphasis to the requirement that abductive success required "originary" thinking.

Proposition 6.17 (Testing hypothesis) *The requirement that an abductive hypothesis be a conjecture that a given proposition has an epistemic standing of at least degree k, sets the standard for the subsequent* testing *of the proposition. The test is to determine whether it does in fact possess epistemic virtue to that degree.*

Corollary 6.17(a) *makes it perfectly unintelligible for abducers both to conjecture and to test propositions that are objects of their antecedent beliefs.*

It can be seen on inspection that **AP** logics in the manner of Josephson and Josephson, and Reiter are not fully enough developed to meet reasonable adequacy conditions.

6.2 Another Example

In this section we briefly review the parsimonious covering theory of Yun Peng and James Reggia. As was the case with Josephson and Josephson, our discussion will be limited to a description of characteristic features of this approach, as well as representative problems to which it gives rise. An important feature of this account is its sensitivity to the complexities induced by giving the theory a probabalistic formulation. This is something these authors attempt, but with the requisite reduction in computational costs. In our discussion we shall concentrate on the non-quantitative formulation of the theory, and will leave adjudication of the conflict between qualitative and probabalisitic methodologies for later in this chapter.

This is an appropriate place to make an important terminological adjustment. Explanationism and probabalism are not mutually exclusive methodologies. There is a substantial body of opinion which holds that explanations are intrinsically probabalistic; and that fact alone puts paid to any idea of a strong disjunction. Better that we characterize the contrast as follows.

> An *explanationist* about abduction is one who holds that meeting explanatory targets is intrinsic to his or her enterprise.

Probabilists are not natural rivals of explanationists.

> A *probabilist* is one who holds that, whatever his targets are (including explanatory targets), they are only meetable by way of a probabalistic methodology which is intrinsic to the enterprise, or they are best met in this way.

Peng and Reggia are inference-to-the-best-explanation abductionists, who happen to think it possible to integrate, without inordinate cost, their informal explanationism into probability theory [Peng and Reggia, 1990, ch. 4 and 5]. The main target of their theory are procedures for the derivation of *plausible explanations* from the available data [Peng and Reggia, 1990, p. 1].

Parsimonious covering theory is a theoretical foundation for a class of diagnostic techniques described by association-based abductive models. An associative (or semantic) network is a structure made up of *nodes* (which may be objects or concepts or events) and *links* between nodes, which represent their interrelations or associations. Associative models are evoked by two basic kinds of procedure: (1) the deployment of symbolic cause-effect associations between nodes, and a recurring hypothesize-and-test process. Association-based models include computer-aided diagnostic systems such as INTENIST-1 (for internal medicine; Pople [1975] and Miller, Pople and Meyers [1982]); NEUROLOGIST (for neurology; Catanzarite and Greenburg [1979]); PIP (for edema: Pauker, Gorry, Kassirer and Schwarz [1976]); IDT (for fault diagnosis of computer hardware; Shubin and Ulrich [1982]), and domain-free systems such as KMS.HT (Reggia [1981]); MGR (Coombs and Hartley [1987]) and PEIRCE (Punch, Tanner and Josephson [1986] See also [chapter 4]Magnani:2001).

Given one or more initial problem features, the inference mechanism generates a set of potential plausible hypotheses of 'causes' which can explain the given problem features' [Peng and Reggia, 1990, p. 20]. Thus associative networks involve what we can think of as a logic of discovery. Once the hypotheses have been generated, they are tested in two ways. They are tested for explanatory completeness; and they are tested for their propensity to generate new questions whose answers contribute to the selection of a given hypothesis from its alternatives. This hypothesize-and-test cycle is repeated, taking into account new information produced by its predecessor. This may occasion the need to update old hypotheses; and new ones may be generated

Association-based abduction differs from statistical pattern classification, as well as from rule-based deduction. In statistical pattern classification, the inference mechanism operates on prior and conditional probabilities to generate posterior probabilities. In rule-based deduction, the inference mechanism is deduction which operates on conditional rules. In the case of association-based abduction, the hypothesize-and-test mechanism operates on semantic networks. Both statistical pattern classification and rule-based deduction have strong theoretical foundations — the probability calculus in the first instance, and first order predicate logic in the second. The association-based abductive approach has not had an adequate theoretical foundation. Furnishing one is a principal goal of the parsimonious cover theory (*PCT*)

PCT is structured as follows. In the category of nodes are two classes of events or states of affairs called *disorders D* and *manifestations M*. In the category of links is a causal relation on pairs of disorders and manifestations. Disorders are sometimes directly observable, and sometimes not. However in the context of a diagnostic abduction problem disorders are not directly scrutinizable and must be inferred from the available manifestations. Manifestations in turn fall into two classes: those that are present and those that are not. 'Present' here means "directly available to the diagnostician in the context of his present diagnostic problem". The class of present manifestations is denoted by M+

A *diagnostic problem* P is an ordered quadruple $\langle D, M, C, M+\rangle$, where $D = \{d_1, d_2, \ldots, d_n\}$ is a finite, non-empty set of objects or states of affairs called *disorders*, $M = \{m_1, m_2, \ldots, m_n\}$ is a finite, non-empty set of objects or states of affairs called *manifestations*, and $C \subseteq D \times M$ is a relation whose domain is D and whose range is M, and $D \times M$ is the Cartesian product of D and M. C is called *causality*. $M+$ is a distinguished subset of M which is said to be present

For any diagnostic problem P, and any d_i and m_j, *effects* (d_i) is the set of objects or states of affairs directly caused by d_i and *causes* (m_j) is the set of objects or states of affairs which can directly cause m_j. It is expressly provided that a given m_j might have alternative and even incompatible, possible causes. This leaves it open that *P* may produce a differential diagnosis of an m_j

The effects of a set of disorders *effects* (D_I) is $\bigcup_{d_i \in D_1}$ effects (d_i), the union of all *effects* d_i. Likewise, the causes of a set of manifestations *causes* M_J is $\bigcup_{m_j \in M_J}$ *causes* (m_j), the union of all *causes* (m_j)

The set $D_I \subseteq D$ is a *cover* of $M_J \subseteq M$ if $M_J \subseteq$ *effects* (D_I). Informally, a cover of a set of manifestations is what *causally accounts* for it

A set $E \subseteq D$ is an *explanation* of $M+$ with respect to a problem *P* if and only if *E* covers $M+$ and *E* meets a *parsimony requirement*. Intuitively, parsimony can be thought of as minimality, or non-redundancy or relevance

A cover D_I of M_J is *minimum* if its cardinality is the smallest of all covers of M_J. A cover of M_J is *non-redundant* if none of its proper subsets is a cover of M_J, and is redundant otherwise. A cover D_I of $M+$ is *relevant* if it is a subset of *causes* $M+$, and otherwise is *irrelevant*

Finally, the solution *Sol(P)* of a diagnostic problem *P* is the set of all explanations of $M+$

Two especially important facts about diagnostic problems should be noted.

> *Explanation Existence Theorem.* There exists at least one explanation for any diagnostic problem.
>
> *Competing Disorders Theorem.* Let *E* be an explanation for $M+$, and let $M + \bigcap$ *effects* $(d_1) \subseteq M + \bigcap$ *effects*(d_2) for some d_1 and d_2 εD. Then (1) d_1 and d_2 are not both in *E*; and (2) if $d_1 \varepsilon$ *E* then there is another explanation *E** for $M+$ which contains d_2 but not d_1 and which is of the same cardinality or less.
>
> *Cover Taxonomy Lemma.* Let 2^D be the power set of *D*, and let S_{mc}, S_{nc}, S_{re} and S_c be, respectively, the set of all minimal covers, the set of all non-redundant covers, the set of all relevant covers, and the set of all covers of $M+$ for a given diagnostic problem *P*. Then $\emptyset \subseteq S_{mc} \subseteq S_{nc} \subseteq S_c \subseteq 2^D$

We now introduce the concept of a generator. Let $g_1, g_2, \ldots, g_n$ be non-empty pairwise disjoint subsets of *D*. Then $G_I = \{g_1, g_2, \ldots, g_n\}$ is a *generator*. The class $|G_I|$ generated by G_I is $\{\{d_1, d_2, \ldots, d_n\} \mid d_i \varepsilon g_i, 1 \leq i \leq n\}$

Here is an informal illustration of how Parsimonious Cover Theory deploys these structures. We consider the well-known example of a chemical spill diagnosed by *KMS.HT*, a domain-independent software program for constructing and examining abductive diagnostic problem-solving

The problem is set as follows: A river has been chemically contaminated by an adjacent manufacturing plant. There are fourteen different kinds of chemical spills capable of producing this contamination. They include sulphuric acid, hydrochloric acid, carbonic acid, benzene, petroleum, thioacetamide and cesmium. These constitute the set of all possible *disorders* d_i. Determination of the type of spill is based on the following factors: pH of water (acidic, normal or alkaline), colour of water (green or brown normally, and red or black when discoloured), appearance of the water (clear or oily) radioactivity, spectrometry results (does the water contain abnormal elements such as carbon, sulphur or metal?), and the specific gravity of the water. *M* is the set of all abnormal values of these measurements

The diagnostician is called the Chemical Spill System, CSS. When there is a spill an alarm is triggered and CSS begins collecting manifestation data. CSS's objective is to identify the chemical or chemicals involved in the spill

The knowledge base for CSS includes each type of spill that might occur, together with their associated manifestations. For example, if the d_i in question is sulphuric acid, the knowledge base reflects that spills of sulphuric acid are possible at any time during the year, but are more likely in May and June, which are peak periods in the manufacturing cycle. It also tends to make the water acidic, and spectrometry always detects sulphur. If the spilled chemical were benzene, the water would have an oily appearance detectable by photometry; and spectrometry might detect carbon. If petroleum were the culprit then the knowledge system would indicate its constant use, heaviest in July, August and September. It might blacken the water and give it an oily appearance, and it might decrease the water's specific gravity. It would also be indicated that spectrometry usually detects carbon

KMS.HT encodes this information in the following way.

Sulphuric Acid
[Description:
Month of year = May $\langle h \rangle$, June $\langle h \rangle$
pH = Acidic $\langle h \rangle$
Spectrometry results = Sulphur $\langle a \rangle$]

Benzene
[Description:
Spectrometry results = Carbon $\langle m \rangle$
Appearance = Oily $\langle m \rangle$]

Petroleum
[Description:
Month of year = July $\langle h \rangle$, August $\langle h \rangle$, September $\langle h \rangle$
Water colour = Black $\langle m \rangle$
Appearance = Oily $\langle m \rangle$
Spectrometry results = Carbon $\langle h \rangle$
Specific Gravity of Water = Decreased $\langle m \rangle$]

Here a is *always*, b is *high likelihood*, m is *medium likelihood*, l is *low likelihood* and n is *never*. It is easy to see that the knowledge base provides both causal and noncausal information. For example, that sulphuric acid has a manifestation pH = acidic is causal information, whereas the information that sulphuric acid use is especially high in May and June is important, but it does not reflect a causal association between sulphuric acid and those months. Information of this second kind reports facts about *setting factors*, as they are called in KMS.HT.

The set *effects* (d_i) is the set of all manifestations that may be caused by disorder d_i, and the set *causes* (m_j) is the set of disorders that may cause manifestation m_j. In particular, *effects (Sulphuric Acid)* = {pH = Acidic, Spectrometry Results = Sulphur}, and *causes (pH = Acidic)* = { Benzenesulphuric Acid, Carbonic Acid, Sulphuric Acid}.

Since there are fourteen disorders in this example, there are 2^{14} possible sets of disorders, hence 2^{14} possible hypotheses. However, for reasons of economy, CSS confines its attention to sets of disorders that meet a parsimony constraint. If we chose minimality as our parsimony constraint, then a legitimate hypothesis must be a smallest cover of all present manifestations $M+$ (a cover is a set of disorders that can causally account for all manifestations). The objective of the CSS is to specify all minimum covers of $M+$.

When a manifestation m_j presents itself it activates the causal network contained in the system's knowledge base. More particularly, it evolves all disorders causally associated with m_j. That is to say, it evokes *causes* (m_j). These disorders combine with already available hypotheses (which are minimum covers of previous manifestations) and new hypotheses are formed with a view to explaining all the previous manifestations together with the new one m_j.

The solution to *P* is the set of all minimal covers. For reasons of economy, it is desirable to produce the solution by way of *generators*.

As we saw a generator is a set of non-empty, pairwise disjoint subsets of *D*. Let the subsets of *D* be *A, B*, and *C*. Then form the set of all sets that can be formed by taking one element from each of the subsets *A, B* and *C*. The set of those sets is a generator on {*A, B, C*}. If the *i*th set in a generator contains n_i disorders, then the generator reflects $\prod_i n_i$ hypothesis, which in the general case is significantly fewer than the complete list.

CSS now continues with its investigation by putting multiple-choice questions to the knowledge base. For example, it asks whether the spill occurred in either April, May, June, July, August, or September. The answer is June. This enables the system to reject Carbon Isotope, which is never used in June. It then asks whether the pH was acidic, normal or alkaline, and is told that it was acidic. This gives rise to hypotheses in the form of a generator that identifies the alternative possibilities as Benzenesulphuric Acid, Carbonic Acid, Hydrochloric Acid, Sulphuric Acid. Each of these is a minimum cover.

Now there is a question asking whether Metal, Carbon or Sulphur were detected Spectrometrically. The answer is that Metal and Carbon were detected. This now excludes Sulphuric Acid because it is always the case that Spectrometry will detect the presence of Sulphur, and did not do so in this case.

The presence of metal evokes four disorders: Hydroxyaluminum, Cesmium, Rubidium and Radium. None of these occurs in the present generator. So to cover the previous manifestation (acidic pH) and the new manifestation, CSS must produce new hypotheses involving at least two disorders.

The presence of Carbon, evokes six disorders: Carbonic Acid, Benzene, Petroleum, Benzenesulphuric Acid, Thioacetamide and Chromogen. Carbonic Acid and Benzenesulphuric Acid are already in the present generator, so they cover both pH = Acidic and Spectrometry = Carbon. Integrating these into the set of four new hypotheses, gives eight minimum covers for three existing manifestations. These in turn give a generator of two sets of competing hypotheses.

These are { Carbonic Acid, Benzenesulphuric Acid} and { Radium, Rubidium, Cesmium, Hydroxyaluminum}.

The system now asks whether Radioactivity was present. The answer is Yes. This evokes four disorders: Sulphur Isotope, Cesium, Rubidium and Radium. Each hypothesis to date involves one of these disorders, except for the Hydroxyaluminum hypothesis (which is not radioactive). Thus Hydroxyaluminum is rejected and a new generator is formed. It contains six minimum covers for the current manifestation

The system asks whether Specific Gravity was normal, or increased or decreased. The answer is that it was increased. This evokes four disorders: Hydroxyaluminum, Cesmium, Rubidium and Radium, which are also evoked by the radioactivity answer. Accordingly, the solution to the chemical spill problem is given by a generator containing two sets of incompatible alternatives: { Carbonic Acid, Benzenesulphuric Acid} and { Radium, Rubidium, Cesmium}.

6.2.1 Remarks

CSS is a very simple system. (A typical search space involves 2^{50} candidates and higher.) Even so, it is faced with 16,384 distinct sets of disorders. Of these, 91 are two-disorders sets. But CSS identifies the six plausible hypotheses (where plausibility is equated with minimality). This is a significant cutdown from a quite large space of possible explanations. This indicates that PCT does well where the approach discussed in the previous section tends to do badly, namely, on the score of computational effectiveness. In the engineering of this system, finding plausible inferences just is making the system computationally effective.

Peng and Reggia point out that for certain ranges of cases non-redundancy is a more realistic parsimony constraint than minimality, and that in other cases relevancy would seem to have the edge [Peng and Reggia, 1990, pp. 118–120]. Nonredundancy was a condition on Aristotle's syllogism, and it is akin to Anderson and Belnap's full-use sense of relevance and of a related property of linear logic. In standard systems of relevant logic a proof $\prod$ of ϕ from a set of hypotheses $\sum$ is relevant iff $\prod$ employs every member of $\sum$. Relevance is not nonredundancy, however; a relevant proof might have a valid subproof, since previous lines can be re-used. If the logic is linear, all members of $\sum$ must be used in $\prod$, and none can be re-used. Peng and Reggia's nonredundancy is identical to Aristotle's: No satisfactory nonredundant explanation can have a satisfactory proper sub-explanation. Relevance for Peng and Reggia is restricted to causal relevance, and so captures only part of the sense of that broad concept. If we allow nonredundancy as a further sense of relevance, then the following structural pattern is discernible in the Peng-Reggia approach. One wants one's explanations to be plausible. Plausible explanations are those produced parsimoniously. Parsimony is relevance, in the sense of nonredundancy. Thus relevance is a plausibility-enhancer. In later chapters, we will investigate the possibility that relevance and plausibility combine in a different way from that suggested here.

In essence, what CSS does is select a causal explanation of some symptoms from explanation of some symptoms from a comparatively small set of phenomena that are not only known to be possible causes of it but are known to be actual causes under particular circumstances. Since the actual circumstances of the symptoms exhibit some variations from circumstances under which it it known that those symptoms are caused, the abducer's trigger is not those symptoms, but rather that those symptoms do not correlate with a known possible cause sufficiently to meet the abducer's (often implicit) epistemic-level test. In a rough and ready way, the abducing device is attempting to move from knowledge of what causes of these symptoms are known to be *like* to what, in this particular case, the symptoms actually are. In the nature of the case, the hypothesizing diagnostician cannot know the answer to this question with the appropriate degree of knowledge. This allows us to say that CSS systems hypothesize that if a certain H did meet epistemic standards which it presently does not appear to meet, it would (subjunctively) be the (best) causal explanation of the manifestations. The winning conjecture in each case is "originary" and "epistemically challenged" i.e., a conjecture that some proposition has a degree of epistemic virtue which it is now not known to have.

6.3 Coherentism and Probabilism

6.3.1 The Rivalry of Explanationism and Probabilism

We briefly characterized explanationism and probabilism in the previous section. As we say, they are not natural rivals. There are ways of being an explanationist which involves no resistance to probabilistic methods, and there are ways of being a probabilist which are compatible with some ways of being an explanationist. For example, one way of being an explanationist about abduction is to insist that all abductive inference is inference to the best explanation. But there is nothing in this that precludes giving a probabilistic account of inference to the best explanation. Equally, someone could be a probabilist about inference in general and yet consistently hold that explanation is intrinsic to abuductive inference.

Even so, there are explanationists who argue that probabilistic methods are doomed to fail or, at best to do an inadequate job. If this is right, then knowing it endows the enquirer into abduction with some important guidance for the resource-management aspects of his programme. It is negative guidance: Do not give to the calculus of probability any central and load-bearing place in your account of abductive reasoning.

In this section we examine one way of trying to make good on this negative advice. In doing so, we develop a line of reasoning which we find to have been independently developed in Thagard [2000]. Since our case heavily involves Thagard himself, and pits him against the same position that Thagard also examines, we shall not give a fully-detailed version of our argument, since the case we advance is already set out in Thagard [2000].

6.4 Explanatory Coherence

Like the approach of Peng and Reggia, Thagard sees abductive inference as a form of causal reasoning [Thagard, 1992]. It is causal reasoning triggered by surprising events (again, a Peircean theme), which, being surprising call for an explanation. Explanations are produced, by hypothesizing causes of which the surprising events are effects.

Thagard understands explanationism with respect to causal reasoning to be qualitative, whereas probabilism proceeds quantitatively. In recent years, both approaches have been implemented computationally. In Thagard [1992], a theory of *explanatory coherence* is advanced. It can be implemented by ECHO, a connectionist program which computes explanatory coherence in propositional networks. Pearl [1988] presents a computational realization of probabilistic thinking. These two styles of computational implementation afford an attractive way of comparing the cognitive theories whose implementations they respectively are. The psychological realities posited by these theories can be understood in part by way of the comparison relations that exist between ECHO and probabilistic realizations. As it happens, there is a probabilistic form of ECHO, which shows not only that coherentist reasoning is not logically incompatible with probabilistic reasoning, but that coherentist reasoning can be considered a special case of probabilistic reasoning. What makes it so is that ECHO's inputs can be manipulated to form a probabilistic network which runs the Pearl algorithms. Coherentism is not the only way of being an explanationist, and Pearl's is not the only way of being a probabilist. But since the computer implementations of these two particular theories are more advanced than alternative computer models, comparing the Thagard and the Pearl computerizations is an efficient way to compare their underlying explanationist and probabilist theories as to type

The chief difference between explanationist and probabilist accounts of human reasoning lies in how to characterize the notion of *degrees of belief*. Probabilists hold that doxastic graduations can be described in a satisfactory way by real numbers under conditions that satisfy the principles of the calculus of probability. Probabilism is thus one of those theories drawn to the idea that social and psychological nature has a mathematically describably structure, much in the way that physics holds that physical reality has a mathematical structure. Explanationism, on the other hand, is understood by Thagard to include the claim that human causal reasoning, along with other forms of human reasoning, can be adequately described non-quantitatively. It rejects the general view that reasoning has a mathematical structure, and it rejects the specific view that reasoning has a structure describable by the applied mathematics of games of chance.

Disagreements between explanationists and probabilists are not new; and they precede the particular differences that distinguish Thagard's theory from Pearl's. Probabilism has attracted two basic types of objection. One is that the probabilistic algorithms are too complex for human reasoners to execute [Harman, 1986]. The other is that experimental evidence suggests the descriptive inadequacy of probablistic accounts [Kahneman *et al.*, 1982], a score on which explanationist accounts do better [Read and Newhall, 1993; Schank and Ranney, 1991; Schank and Ranney, 1992; Thagard and Kunda, 1998]. See also [Thagard, 2000, p. 94]. The importance of the fact that coherentist reasoning can be seen as a special case of probabilistic reasoning is that it suggests that there may be a sense in which the probabilistic approach is less damaged by these basic criticisms than might otherwise have been supposed. There is ample reason to think that the sheer dominance of human reason is evidence of the presence of the *Can Do Principle*, which we discussed in section 3.3.1. A theorist comports with the *Can Do Principle* if, in the course of working on a problem belonging to a discipline *D*, he works up results in a different discipline *D**, which he then represents as applicable to the problem in *D*. The principle is *adversely* triggered when the theorist's attraction for *D** is more a matter of the ease or confidence with which results in *D** are achieved, rather than their well-understood and well attested to applicability to the theorist's targets in *D*. (A particularly dramatic example of *Can Do* at work, is the decision by neoclassical economists to postulate the infinite divisibility of utilities, because doing so would allow the theory to engage the fire-power of the calculus.)

The conceptual core of the theory of explanatory coherence (TEC) is captured by the following qualitative principles [Thagard, 1989; Thagard, 1992; Thagard, 2000], and [Magnani, 2001a, pp.34 and 138].

1. *Symmetry.* Unlike conditional probability, the relation of explanatory coherence is symmetric.

2. *Explanation.* If a hypothesis explains another hypothesis, or if it explains the evidence, then it also coheres with them and they with it. Hypotheses which jointly explain something cohere with each other. Coherence varies inversely with the number of hypotheses used in an explanation. (*cf.* the minimality condition of Peng and Reggia [1990].)

3. *Analogy.* Similar hypotheses that explain similar bodies of evidence cohere with one another.

4. *Observational Priority.* Propositions stating observational facts have a degree of intrinsic acceptability.

5. *Contradiction.* Contradictory propositions are incoherent with each other. (It follows that the base logic for TEC must be paraconsistent lest an episodic inconsistency would render any explanatory network radically incoherent.)

6. *Competition.* If both P and Q explain a proposition, and if P and Q are not themselves explanatorily linked, then P and Q are incoherent with each other. (Note that explanatory incoherence is not here a term of abuse.)[5]

7. *Acceptance.* The acceptability of a proposition in a system of propositions depends on its coherence with them.

TEC defines the coherence and explanation relations on systems of propositions. In ECHO propositions are represented by *units* (viz., artificial neurons). Coherence relations are represented by excitatory and inhibitory *links*.

ECHO postulates a *special evidence unit* with an activation value of 1. All other units have an initial activation value of 0. Activation proceeds from the special evidence unit to units that represent *data*, thence to units that represent propositions that *explain* the data, and then to units representing propositions that explain propositions that explain data; and so forth. ECHO implements principle (7), the Acceptability rule, by means of a connectivist procedure for updating the activation of a unit depending on the units to which it is linked. Excitatory links between units cause them to prompt their respective activation. Inhibitory links between units cause them to inhibit their respective activation.

Activation of a unit a_j is subject to the following constraint:

$$a_j(t+1) = a_j(t)(t-d)^{+}\text{net}_j(\text{max-a}_j(t)\,\textit{if}\,\text{net}j > 0, \textit{otherwise}\ \text{net}_j(a_j(t))\text{-}\textit{min})$$

in which d is a decay parameter (e.g., 0.5) that decrements every unit at each turn of the cycle; j is the minimum activation (-1); *max* is maximum activation (1). Given the weight w_{ij} the net input to a unit (net_j) is determined by

$$net_j = \textstyle\sum_i w_{ij} a_i(t).$$

In ECHO links are symmetric, but activation flow is not (because it must originate in the evidence unit). ECHO tolerates the presence of numerous loops, without damaging the systems ability to make acceptability computations.

Updating is performed for each unit u_i. For this to happen the weight of the link between u_i and u_j (for any u_j to which it is linked) must be available to it; and the

[5] Propositions are explanatorily linked if one explains the other or jointly they explain something else.

same holds for the activation of u_j. Most units in ECHO are unlinked to most units. But even if ECHO were a completely connected system, the maximum number of links with n units would be $n-1$. Updating, then, requires no more than n*($n-1$) calculations. However, uncontrolled excitation (the default weight on excitatory links) gives rise to activation oscillations which preclude the calculation of acceptability values. Thagard reports experiments which show high levels of activation stability when excitation, inhibition and decay are appropriately constrained. Accordingly, the default value for excitation is 0.4; for inhibition (the default value of inhibitory links) the value is -0.6; and 0.5 for decay [Thagard, 2000, p. 97]. Thagard also cites experimental evidence which indicates that ECHO's efficiency is not much influenced by the size or the degree of connectivity of networks. Larger networks with more links do not require any systematic increase in work required by the system to make its acceptability calculations [2000, p. 97].

6.4.1 Probabilistic Networks

A clear advantage of probabilistic approaches to reasoning have over explanationist approaches is a precise theoretical vocabulary, a clear syntax, and a well-developed semantics. Degrees of belief are associated with members of the [0-1] interval of the real line, subject to a few special axioms. A central result of this approach in Bayes' theorem:

$$\Pr(h|e) = \frac{\Pr(h) \times \Pr(e|h)}{\Pr(e)}$$

The theorem provides that the probability of a hypothesis on a body of evidence is equal to the probability of the hypothesis alone multiplied by the probability of the evidence given the hypothesis, divided by the probability of the evidence alone. The theorem has a certain intuitive appeal for the abduction theorist, especially with regard to the probability of the evidence relative to the hypothesis, which resembles *AKM*-abduction schema of 1.1 in an obvious way. The Bayesian approach to human reasoning is, however, computationally very costly [Harman, 1986; Woods and Walton, 1972]. Probabilistic updating requires the calculation of conjunctive probabilities, the number of which grow exponentially with the number of conjuncts involved. Three propositions give rise to eight different probability alternatives, and thirty would involve more than a billion probabilities [Harman, 1986, p. 25]. It is also possible that coherence maximization is computationally intractable, but provided the system has the semidefinite programming algorithm, TEC is guaranteed that the system's optimality shortfall will never be more than 13 percent [Thagard and Verbeurgt, 1998]. In their turn, Pearl's network, in common with many other probabilistic approaches, greatly reduces the number of probabilities and probability determinations by restricting these determinations to specific classes of dependencies. If, for example, Y depends wholly on X and Z depends wholly on Y, then in calculating the probability of Z, the probability of X can be ignored, even though X, Y and Z form a causal network.

In Pearl's network, nodes represent multi-valued variables, such as body temperature. In two-valued cases, the values of the variable can be taken as truth and falsity. ECHO's nodes, on the other hand represent propositions, one each for a proposition and its negation. In Pearl networks edges represent dependencies, whereas in ECHO they represent coherence. Pearl networks are directed acyclic graphs. Its links are anti-symmetric, while those of ECHO are symmetric. In Pearl networks the calculation of probabilities of a variable D is confined to those involving variables causally linked to D. Causally independent variables are ignored. If, for example, D is a variable whose values are causally dependent on variables A, B, and C, and causally supportive of the further variables E and F, then the probabilities of the values of D can be taken as a vector corresponding to the set of those values. If D is body temperature and its values are *high, medium* and *low*, then the vector (.6 .2 .1) associated with D reflects that the probability of high temperature is .6, of medium temperature is .2, and of low temperature is .1. If subsequent measurement reveals that the temperature is in fact high, then the associated vector is (1 0 0). If we think of A, B and C as giving prior probabilities for the values of D, and of E and F as giving the relevant observations, then the probability of D can be calculated by Bayes' Theorem.

For each of Pearl's variables X, '(x)' denotes the degrees of belief calculated for X with regard to each of its values. Accordingly, BEL(x) is a vector with the same number of entries as X has values. BEL(x) is given by the following equation:

$$\text{BEL(x)} = \alpha \times \lambda(x) \times \pi(x)$$

in which α is a normalization constant which provides that the sum of the vector entries is always 1. λ(x) is a vector representing the support afforded to values of X by variables that depend on X. π(x) is a vector representing the support lent to values of X by variables on which X depends.

It is known that the general problem of inference in probabilistic networks is NP-hard [Cooper, 1990]. Pearl responds to this difficulty in the following way. He considers the case in which there is no more than one path between any two nodes, and produces affordable algorithms for such systems [Pearl, 1988, ch. 4]. If there is more than one path between nodes, the system loops in ways that destabilizes the calculation of BEL's values. In response to this difficulty, procedures have been developed for transforming multiply-connected networks into uni-connected networks. This is done by clustering existing nodes into new multi-valued nodes

Pearl [1988, ch. 4], discusses the following case. Metastatic cancer causes both serum calcium and brain tumor, either of which can cause a coma. Thus there are two paths between metastatic cancer and coma. Clustering involves replacing the serum calcium node and the brain tumor node into a new node representing variables whose values are all possible combination of values of the prior two, viz.: increased calcium and tumor; increased calcium and no tumor; no increased calcium and tumor; and no increased calcium and no tumor. Clustering is not the only way of dealing with loops in probabilistic networks. Thagard points out [2000, p. 101] that Pearl himself considers a pair of approximating techniques, and that Lauritzen and Spiegelharter [1988], have developed a general procedure by transforming any directed acyclic graph into a tree of the graph's cliques. Hrycej [1990] describes approximation by stochastic simulation as a case of sampling from the Gibbs distribution in a random Markov field, Frey [1998] also exploits graph-theoretic inference pattern in developing new algorithms for Bayesian networks.

Recurring to the earlier example of *D* depending on *A, B* and *C*, and *E* and *F* depending on *D*, *D*'s BEL values must be computed using values for *A, B, C, E* and *F* only. To keep the task simple, suppose that *a, b, c, d,* and *e* are, respectively, the only values of these variables. Pearl's algorithms require the calculation of *d*'s probability given all possible combination of values of the variables to which *D* is causally linked. If we take the case in which the possible values of *D* are truth (*d*) or falsity (not-*d*), Pearl's algorithm requires the calculation of eight conditional probabilities. In the general case in which *D* is causally linked to *n* variables each with *k* variables, the system will have to ascertain k^n conditional probabilities. Where *n* is large, the problem of computational intractability reappears, and occasions the necessity to employ approximation procedures. Even if *n* isn't large, the question isn't whether the requisite numbers of calculation of conditional probabilities *can* be done by humans, that rather whether they *are* done (in that number) by humans.

6.5 Pearl Networks for ECHO

There are important differences between ECHO and Pearl. There are also some significant similarities. If ECHO is faced with rival hypotheses, it will favour the ones that have greatest explanatory power. The same is true of Pearl networks. ECHO likes hypotheses that are explained better than those that aren't. Pearl networks have this bias, too. Thus there is reason to think that there may be less of a gap between ECHO and Pearl networks than might initially have been supposed. Accordingly Thagard discusses a new network, PECHO. PECHO is a program that accepts ECHO's input and constructs a Pearl network capable of running Pearl's algorithms. (For details see Thagard [2000, pp. 103–108].) The theory of explanatory coherence which ECHO implements via connectionist networks, as well as by the further algorithms in Thagard and Verbeurgt [1998], can also be implemented probabilistically. That alone is a rather striking result. Some of the older objections to probabilism held that the probability calculus simply misconceived what it is to reason. The integration achieved by Thagard suggests that the inadequacies which critics see in probabilism need to be treated in subtler and more nuanced ways. It remains true that PECHO is computationally a very costly network. Vastly more information is needed to run its update algorithms than is required by ECHO. Special constraints are needed to suppress loops, and it is especially important that the simulated reasoner not discard information about co-hypotheses and rival hypotheses. But given that what coherentists think is the right account of causal reasoning is subsumed by what they have tended to think is the wrong account of it, computational cost problems occur in a dialectically interesting context. How can a good theory be a special case of a bad theory? In general when a bad theory subsumes a good, it is often more accurate to say that what the bad theory is bad *at* is what the good theory is good at. This leaves it open that the bad theory is *good* in ways that the good theory is not. This gives a context for consideration of computational cost problems. They suggest not that the probabilistic algorithms are wrong, but rather that they are beyond the reach of human reasoners.

> If one accepts the view of Goldman [1986] that power and speed are epistemological goals as reliability then explanationist models can be viewed as desirable ways of proceeding apace with causal inference while probabilistic models are still lost in computation. [Thagard, 2000, p. 271].

Thagard goes on to conjecture that 'the psychological and technological applicability of explanation and probabilistic techniques will vary from domain to domain' [Thagard, 2000, p. 112]. He suggests that the appropriateness of explanationist techniques will vary from high to low depending on the *type of reasoning* involved. He considers the following list, in which the first member is explanationistly most appropriate, and the last least:

Social reasoning
Scientific reasoning
Legal reasoning
Medical diagnosis
Fault diagnosis
Games of chance.

We find this an attractive conjecture. It blunts the explanationist complaint that probabilism is *just wrong* about reasoning. It suggests that there is something for probabilism to be right about, just as there is something for explanationism also to be right about. It suggests, moreover, that they are not the same things. The fact that coherentist networks are subcases of probabilistic networks, when conjoined with the fact that what is indisputably and demonstrably troublesome about probabilistic networks is their inordinate informational and computational complexity and, relatedly, their psychological unreality, occasions a further conjecture. Systems that are well-served by explanationism are systems having diminished information-processing and computational power, and have psychological make-ups that systems for which probabilism does well don't have. This, in turn, suggests that the variability which attaches to the two approaches is not so much a matter of type of reasoning as it is a matter of type of reason*er*. In chapter 2, we proposed that cognitive agency comes in types and that type is governed by levels of access to, and quantities of, cognitively relevant resources in relation to the strictness of cognitive goals: information, time and computational capacity. If we take the hierarchy generated by this relation, it is easy to see that it does not preserve the ordinal characteristics of Thagard's list. So, for example, scientific reasoning will be represented in our hierarchy of agencies by every type of agent involved in scientific thinking, whether Joe Blow, Isaac Newton, the Department of Physics at the University of Birmingham, all the science departments of every university in Europe, NASA, neuroscience in the 1990s, postwar economics, and so on. This makes us think that we need not take our hierarchy to be a rival of Thagard's. Ours is an approach that also explains the dialectical tension represented by the subsumption of ECHO by probabilistic networks. It explains why explanationism is more appropriate for reasoning by an individual agent, and why probabilism could be appropriate for an institutional agent. And it explains why it is unnecessary to convict either account of the charge that it simply doesn't know what reasoning is.

On the other hand, even for a given type of agency there can be variation in the kind of reasoning it conducts. Individual reasoners undertake cognitive tasks for which social reasoning is appropriate, but they also try their hands at games of chance. A medical diagnostician need not always operate under conditions of battlefield triage; sometimes he has plenty of time, an abundance of already-sorted information, as well as the fire-power of his computer to handle the statistics. What this suggests is that the fuller story of what model is appropriate to what reasoning contexts will be one that takes into account the type of agent involved and the type of reasoning he is engaged in. It is a story which suggests the advisability of criss-crossing a kinds-of-reasoning grid on a kinds-of-resources grid — of laying a Thagardian hierarchy across one such as our own.

Things are less good, however, when we move from the solo reasoning of agents to interactive n-agent reasoning. It is a transition that bodes less well for the probabilistic approach than for various others. It is a feature of Thagard's hierarchy that, whereas at the bottom objective probabilities are in play, at the top we must do with subjective probabilities. But subjective probabilities are notorious for their consensual difficulties in contexts of n-agent reasoning. This is an invariant feature of them irrespective of whether the interacting agents are Harry and Sarah or the

Governments of Great Britain and France. Then, too, an individual reasoner may want to reason about the state of play in Monte Carlo, which is reasoning of a type for which the probabilistic approach is tailor-made. Even so, the individual gambler cannot run the probabilistic algorithms willy-nilly.

It remains the case that at least part of the reason that explanationism gets the nod over probabilism in the case of individual reasoning is that in general individual reasoners are constitutionally incapable in reasoning in the way that probabilism requires, but is capable of reasoning as coherentism requires. Part, too, of the reason that probabilism is psychologically inappropriate for individuals is that agency-types for which probabilism is appropriate don't (except figuratively) *have* psychologies; hence have no occasion to impose the constraint of fidelity to psychological reality. When PECHO took ECHO into probabilistic form, it endowed PECHO-reasoners with probabilized counterparts of reasonings that an ECHO-reasoner is capable of. But it also endowed the PECHO-reasoner with two qualities which in the general sense are fatal for the individual reasoner, a capacity to run algorithms on a scale of complexity that removes it from an individual's reach, and a hypothetical psychology which no individual human has any chance of instantiating. Thus the ECHO-reasoner can get most things right which PECHO can get right, but it can't get them right in the *way* that PECHO does. The heart and soul of this difference is that the way in which the PECHO-reasoner must operate to get things right involve its having properties which no individual can have, in fact or in principle. Institutional reasoners are another matter, subject to the quantifications we have already taken note of.

As it would now appear, an individual's cognitive actions are performed in two modalities, i.e., consciously and subconsciously. Bearing in mind the very substantial difference in information-theoretic terms between consciousness and subconsciousness, this is a good place to remind the reader of the possibility that accounts that are as computationally expressive as probabilistic models might more readily apply to the subconscious cognitive processes of individual agents, which appear to be processes that aren't informationally straitened in anything like the way of conscious processes. It is less clear that unconscious cognitive systems have structures that are not all realistically represented as structures in the real line. but it is harder to make a change of psychological unreality stick against probabilism precisely because the true psychological character of the unconsciousness of individuals is itself hardly a matter of consensus, to say nothing of theoretical clarity.

It is true that some theorists have no time for the idea of unconscious reasoning (*cf.* Peirce's insistence at Peirce [1931–1958, pp. 5.109 and 2.144] or for unconscious cognition of any sort). Our present suggestion will be lost on those who share this view. On the other hand, if one's epistemological orientation is a generally reliabilist one, it is difficult to maintain that the idea of unconscious reasoning is simply oxymoronic. We ourselves take the reliabilist stance. If cognition and good reasoning are the products of processes that are working as they should (i.e., normally, as in Millikan [1984] there is nothing in this basic idea to which consciousness is either intrinsic or exclusive).

6.6 Neuropharmacological Intervention

Neural disorders are often the result of imbalances of neurochemicals, owing for example, to cell damage or degeneration. When the requisite neurochemical is irrevocably depleted a serious neurological malady is the result. Parkinson's disease is such a disorder.

Pharmacological remedies of disorders such as Parkinson's disease are the object of what is called *rational drug design* (RDD) (see, e.g., [van den Bosch, 2001], from which the present section is adapted). In a typical RDD, information is collated concerning the requisite organic structures and of the success or failure of past pharmacological interventions. In wide ranges of cases, computer models are able to suggest further chemical combinations and/or dosages. RDDs are sometimes supplemented or put aside in favour of hit-and-miss strategies in which large quantities of chemicals are produced and tested *in vitro* for their capacity to influence receptors in the targeted neural structures. A third procedure is the generation of data from laboratory studies of animals.

Following Timmerman *et al.* [1998], van den Bosch develops the Parkinson's disease example with a description of an important class of subcortical nuclei called *basal ganglia*, which are necessary for control in voluntary behaviour. When Parkinson's disease strikes, a component of the basal ganglia, *substantia nigra pars compacta* (SNC), is significantly degraded, a result for which there is no known cause. The SNC furnishes the neurotransmitter dopamine, whose function is to help modulate signals from the cortex. Figure 6.1 schematises the function of dopamine.

In Parkinson's disease the object is to stimulate chemically increases in dopamine levels. L-dopa has had a mixed success in this regard. In the first five years, it is highly effective, but with nausea as a significant side effect; and after five years of use, its therapeutic value plummets dramatically. Parkinson's research is, therefore, dominated by the quest for alternative modes of dopamine receptor agonists, especially those that interact with only particular dopamine receptors. Figure 6.1 charts the effect of versions of these dopamine receptor agonists. Two receptors on the passage from the stimulation to the SNR/GPi are indicated as D1 and D2. D1 occurs on the direct route, whereas D2 occurs on the indirect route via the GPe. Both D1 and D2 are receptive to dopamine, but differently. When dopamine stimulates D1, this excites the relevant cell, whereas the stimulation of D2 inhibits it. A natural question is whether a combined excitation/inhibition convergence is necessary for a favourable outcome. Van den Bosch reports studies that show that

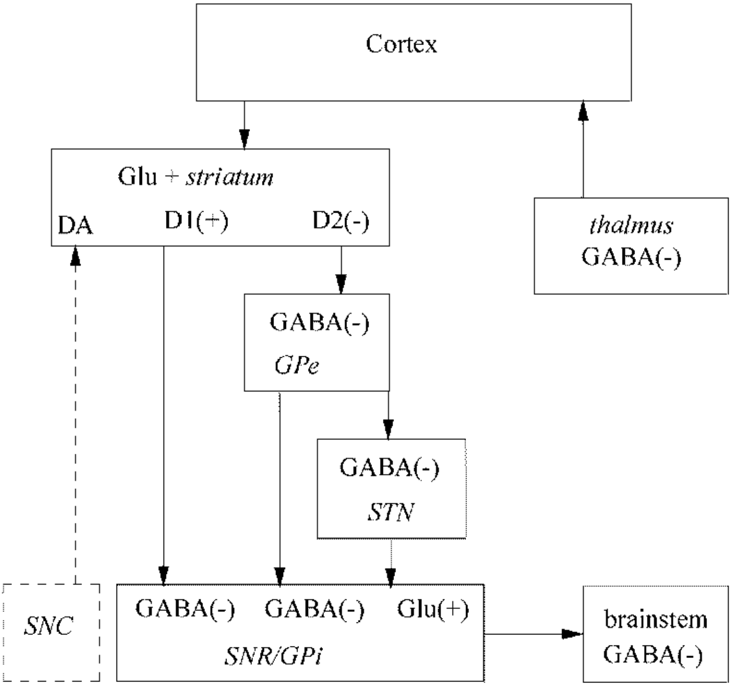


Figure 6.1

compounds that stimulate D1 but not D2 receptors are ineffective [van den Bosch, 2001, p. 32].

The object of a drug design is to discover a successful agonist for the disorder in question. In Parkinson's research, as in many other sectors of neuropharmacology, this is done by modelling what of the patient's neurological endowments are known to be of relevance to dopamine stimulation, together with what our various pharmacological regimes are known to stimulate such receptors. When these two knowledge bases are coordinated in the appropriate way, it is often possible to model the resultant dynamic systems by a qualitative differential equation (QDE). An example of how a QDE might model the basal ganglia as schematized in Figure 6.1 is sketched in Figure 6.2.

M+ and M- are, respectively, monotonically increasing and decreasing functions. Variables have initial values of *high, low* or *normal* and dynamic values of *increasing*, *steady* or *decreasing*. Variables are interconnected in such a way that their differentials determine a value for the combination. These determinacies are subject to a *qualitative calculus*. If a variable v is a differential function over time

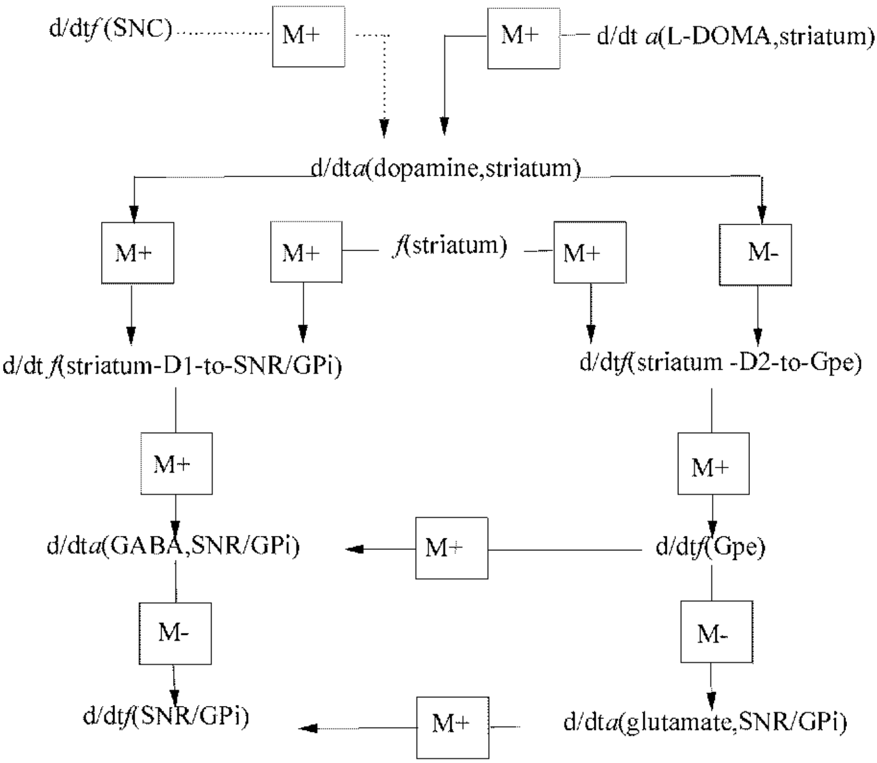


Figure 6.2

of variables v_2 and v_3 and if $v_1 \varepsilon M^+$, then an increase in the value of v_2 and v_3 will drive up the value of v_1, but v_1, will not be assigned a value when v_2 increases and v_3 decreases. Such indeterminacies flow from the qualitative features of QDEs.

> 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 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 not always correct since spurious states may be included as well [van den Bosch, 2001, p. 34].

Figure 6.2 graphs a part of a QDE (which includes a part of the model of basal ganglia schematized in Figure 6.1) together with the operation of dopamine. It charts the firing rates (*f*) of nuclei and neural pathways and amounts (*a*) of neurotransmitters contained in nuclei. Thus if the firing rate of the SNC increases this will increase the quantity of dopamine in the striatum, which in turn depresses activation of the neural pathway that signals to the GPe.

A *drug lead* is a specification of the properties a drug should possess if it is to hit a given desired therapeutic target or range of targets. For every disease for which there is some minimal degree, or more, of medical understanding, a *profile* exists, which is a qualitative specification of the disease. For every profile, the researcher's goal is to supply a drug lead. 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 is lower in the pathological profile, it is included in the goal to increase that variable value [van den Bosch, 2001, p. 35 (emphasis added)].

The (ideal) goal of such a search is to '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 of the goal set' [van den Bosch, 2001, p. 35]. Because of their qualitative nature, QDEs are not typically complete, so that the specification of all possible desired value changes of the goal will not be possible. Accordingly QDE search models are taken as approximations to this ideal [Kuipers, 1999; van den Bosch, 1997; van den Bosch, 1998].

The search tasks can now be described. The starting point in a QDE model and whatever is known of the initial values of variables. Next a goal is selected. A goal is a set of desired values. The the searcher backward chains from the values set up by the goal to 'possible manipulations of the variables' [van den Bosch, 2001, p. 36]. Because QDE methods provide only approximately good (or bad) outcomes, approximation criteria are deployed to measure the closeness of fit between values sought by the goal and values actually produced by the manipulation of variables. The search is successful when it finds a set of manipulations that best approximates to the production of the greatest number of goal values, with least collateral damage. A successful search is thus a piece of backward chaining successfully completed. It is a case of what van den Bosch calls *inference to the best intervention* [van den Bosch, 2001, p. 34].

In the case of Parkinson's disease, one of the goal values is a reduced activation frequency of the SNR/GPi than is present in pathological circumstances. A search of possible manipulations discloses two salient facts. One is an increase (*a*) of L-dopa in the striatum. The other is a reduced firing rate (f) of the indirect channel between the striatum and the GPe induces a suppression of the firing rate of the SNR/GPi. It turns out that a D_2 agonist can produce this decrease, but with lighter consequences for other such (i.e., 'dopaminergics') than dopamine.

QDE searches bear some resemblance to computational diagnostic procedures. In each case, they tell us nothing new about the diseases in question. Their principal advantages are two. One is that these models are means of explicitizing both what is known of the relevant symptoms and how one reasons about them diagnostically and/or interventionistically. Another is that when QDEs are modelled computationally, they, like computerized diagnostic techniques, are especially good at teasing out all consequences compatible with integrity constraints and design limitations. In each case, the backward chaining component is modest. It is little more than manipulation by way of the appropriate monotonically increasing and decreasing functions of variables already known to be interconnected. However, a third feature of the QDE search methodology considerably enhances its status as an instrument of abductive reasoning. For all their incompleteness and, relatedly, their disposition towards misprediction, drug lead exercises can also be used as proposals for experiments, and thus satisfy at least Peirce's idea that decisions about where to invest research resources are a part of the logic of abduction. Claims for research-economic abduction is discussed in the next chapter.

6.7 Mechanizing Abduction

We have attempted to show that explanationist abduction embeds a subjective explanation in the form "If H were to have a degree of cognitive virtue k or higher, this would explain the state affairs at which the explanation is targeted; or for ease of exposition:

> If H were the case, E would also be the case.

Subjunctive conditionals also crop up in another way. In some cases, as we have also seen, an abducer has occasion to consider the explanatory potential of propositions he considers to be false. This is as it should be. One of the tasks of abduction is to set up propositions for trial. One of the purposes of trials is to correct mistakes. The abducer is free to adopt a proposition he now considers false if he also now grants that he might prove to be mistaken in thinking so. Not only does thinking it false preclude a proposition from abductive conjecture; neither does its actually being false. For consider the contrary case. If it were a condition of abductive arguments that their conclusions be true, then twice-over abduction loses its very rationale. Supposing H to be true, the would-be abducer knows this or not. If he knows it there is no abduction possible with regard to H. If he does not know it, he does not know whether a condition on the possibility of what he undertakes to do is met. So he must, at best, be an agnostic about whether what he is doing is abduction. It is one thing — and perfectly in order — for an abducer to be, on occasion, in some doubt about whether his abduction is correct or plausible. But it is not an acceptable account of abduction that the would-be abducer always be in the dark about whether the process he is involved in is abduction *at all*.

Let us grant that essential to the abducer's quest is that for propositions he now takes to be false and concerning which he also allows that he might be mistaken, he be ready to consider not only subjective conditionals, but counterfactual conditionals in the form.

> Even though H is false, it remains the case that were H true then E would be true.

We may take it, then, that real-life abducers routinely deploy counterfactual conditionals. A psychologically real account of what abducers do must take this fact into account.

Computer simulations of what abductive agents do are attempts at producing mechanical models that mimic abductive behaviour. A model gives a good account of itself to the extent that is mimicry approximates to what actually happens in real-life abduction. In particular, therefore, such a model works to the extent that it succeeds in mechanizing counterfactual reasoning. Can it do this? Our answer, which is adapted from [Gabbay and Woods, 2003a] follows closely [Jacquette, 1986]. People who are disposed to give a negative answer to this question are drawn to the following question: What is involved in expressly counterfactual thinking when it is done by real-life human agents? It appears that the human agent is capable of producing some important *concurrences*. For one, he is able to realize that P is true and yet to entertain the assumption that P is not true, without lapsing into inconsistency. Moreover, the human agent seems capable of keeping the recognition that P and the assumption that not-P in mind at the same time. That is, he is able to be aware of both states concurrently. Thirdly, the human agent is capable of deducing from the assumption of not-P that not-Q without in doing so contradicting the (acknowledged) fact that Q might well be true.

When the AI theorist sets out to simulate cognitive behaviour of this sort, he undertakes to model these three concurrences by invoking the operations of a finite state Turning machine. Turning machines manipulate syntax algorithmically; that is, their operations are strictly recursive. The critic of AI's claim to mechanize counterfactual reasoning will argue that no single information processing program can capture all three concurrences. It may succeed in mimicking the first, in which the agent consistently both assents to P and assumes its negation, by storing these bits of information in such a way that no subroutine of the program engages them both at the same time. But the cost of this is that the second concurrence is dishonoured. The human agent is able consciously to access both bits of information at the same time, which is precisely what the Turning machine cannot do in the present case.

It is possible to devise a program that will enable the simulation of the first and the second concurrence. The program is capable of distinguishing syntactically between the fact that P and the counterfactual assumption that not-P, say by flagging counterfactual conditionals with a distinguished marker, for example ⊗. Then the program could have subroutines which has concurrent access to “P” and “⊗not-P⊗”, without there being any danger of falling into inconsistency. Here, too, there is a cost. It is the failure of the program to honour the third concurrence, in which it is possible correctly to deduce “⊗not-Q⊗” from “P”, “⊗not-P⊗” and “if not-P then not-Q”.

Of course, the program could rewrite “If not-P then not-Q” as “If ’ ⊗not-P⊗’ then ’⊗not-Q⊗’”. From the counterfactual assumption “⊗not-P⊗”, the deduction of “⊗not-Q⊗” now goes through, and does so without there being any question of an inconsistency on the deducer’s part.

Still, there is a problem. It is that ⊗-contexts are intensional. There are interpretations of P and Q for which the deduction of “⊗Q⊗” from “not-P”, “counterfactually if P then Q” and “⊗P⊗” fails. Thus it is possible to assume counterfactually that Cicero was a Phoenician fisherman, and that if Cicero was a Phoenician fisherman, then Tully was a Phoenician fisherman, without its following that I assume that Tully was a Phoenician fisherman. The notation “⊗Q⊗” expresses that Q is assumed. Assumption is an opaque context [Quine, 1960], hence a context that does not sanction the intersubstitution of co-referential terms or logically equivalent sentences. [Jacquette, 1986]. Thus ⊗-inference-routines are invalid. Their implementability by any information processing program that, as a finite state Turing machine must be, is strictly extensional dooms the simulation of counterfactual reasoning to inconsistency.

We should hasten to say that there are highly regarded efforts to mechanize reasoning involving counterfactual or belief-convening assumptions. Truth-maintenance systems (TMS) are a notable case in point [Rescher, 1964; Doyle, 1979]. See also [de Kleer, 1986; Gabbay *et al.*, 2003; Gabbay *et al.*, 2002; Gabbay *et al.*, 2004]. The main thrust of TMSs is to restore (or acquire) consistency by deletion. These are not programs designed to simulate the retention of information that embeds belief-contravening assumptions and their presentation to a uniformly embracing awareness. The belief that P is not inconsistent with the concurrent assumption that not-P. There is in this no occasion for the consistency-restoration routines of TMS. Thus ⊗-contexts resemble contexts of direct quotation. Such are contexts that admit of no formally sound extensional logic [Quine, 1960; Quine, 1976]. No strictly extensional, recursive or algorithmic operations on syntax can capture the logic of counterfactual reasoning. Whereupon goodbye to a finite state Turning machine’s capacity to model this aspect of abductive reasoning.

Named after the German word for assumption, ANNAHMEN is a computer program adapted from Shagrin, Rapaport and Dipert [1985]. It is designed to accommodate hypothetical and counterfactual reasoning without having to endure the costs of either inconsistency or the impossibility of the subject's access to belief-contravening assumptions and the beliefs that they contravene. ANNAHMEN takes facts and counterfactual assumptions and conditionals as input. The latter two are syntactically marked in ways that avoid syntactic inconsistency.

This input is then copied and transmitted to a second memory site at which it is subject to deduction. The previous syntactic markers are renamed or otherwise treated in ways that give a syntactically inconsistent set of sentences. The next step is to apply TMS procedures in order to recover a consistent subset in accordance with an epistemic preference-heuristic with which the program has been endowed. In the case before us, the TMS is Rescher's logic of hypothetical reasoning, or as we shall say, the Rescher reduction. From this consistent subset the counterfactual conclusion is deduced by a Lewis logic for counterfactuals and syntactic markers are re-applied. Then all this is sent back to the original memory site. It mixes there with the initial input of beliefs and belief-contravening assumptions. ANNAHMEN can now perform competent diagnostic tasks and can perform well in a Turing test [Turing, 1950]. As Jacquette [2004] observes,

> the functions RESCHER REDUCTION and LEWIS LOGIC call procedures for the Rescher-style reduction of an inconsistent input set to a logically consistent subset according to any desired extensionally definable set of recursive or partially recursive heuristic, and for any desired logically valid deductive procedure for detaching counterfactual conditionals, such as David Lewis' formal system in Counterfactuals [Lewis, 1973].

The problem posed by the mechanization of counterfactual reasoning is that there appeared to be no set of intensional procedures for modelling such reasoning which evades syntactic inconsistency and which allows for what Jacquette calls the "unity of consciousness" of what is concurrently believed and contraveningly assumed.. ANNAHMEN is designed to show that this apparent problem is merely apparent. The solution provided by this approach is one in which the inconsistency that occurs at memory site number two exists for nanoseconds at most and occurs, as it were, subconsciously. Thus counterfactual reasoning does involve inconsistency. But it is a quickly eliminable inconsistency; and it does not occur in the memory site at which counterfactual deductions are drawn. Inconsistency is logically troublesome only when harnessed to deduction. It is precisely this that the ANNAHMEN program precludes. It may also be said that the program is phenomenologically real. When human beings infer counterfactually, they are aware of the concurrence of their beliefs and their belief-contravening assumptions, but they are not aware of the presence of any inconsistency. (Rightly, since the counterfactual inference is performed "at a site" in which there is no inconsistency.)

The ANNAHMEN solution posits for the reasoning subject the brief presence of an inconsistency that is removed subconsciously. It is some interest to the present authors that the program implements the operation Putter-of-Things-Right of [Gabbay and Woods, 2001b]. This is a device postulated for the human information processor. What makes the ANNAHMEN proposal especially interesting in this context is that, in effect, it purports to show that Putter-of-Things-Right is mechanizable.

Whether it is or not, we find ourselves in agreement with Jacquette in the case of an ANNAHMEN approach to *counterfactual* reasoning. Jacquette shows that while ANNAHMEN handles certain types of counterfactual reasoning, it fails for other types. Further even though certain refinements to the ANNAHMEN protocols, in the manner of Lindenbaum's Lemma for Henkin-style consistency and completeness proofs or in the manner of the Lemma for consistent finite extensions of logically consistent sets, resolve some of these difficulties; they cannot prevent others [Tarski, 1956; Henkin, 1950]. We side with Jacquette in thinking that there "is no satisfactory extensional substitute for the mind's intentional adoption of distinct propositional attitudes toward beliefs and mere assumptions or hypothesis". We shall not here reproduce details of Jacquette's criticisms; they are well-presented in [Jacquette, 2004].

Our more immediate interest is in the recurring question of whether the logic of down below is plausibly considered logic. Earlier we briefly noted treatments of abductive insight by connectionist models of prototype activation at the neurological level. (See also [Churchland, 1989; Churchland, 1995; Burton, 1999].) We were drawn to the connectionist model because it see to capture important aspects of abductive behaviour at subconscious and prelinguistic levels. If this is right, there are crucial aspects of abductive practice which on pain of distortion cannot be represented as the conscious manipulation of symbols. As we now see, the manner in which ANNAHMEN is thought to fail bears on this issue in an interesting way. At present there is no evidence that conscious mental functions such as memory and desire, and even consciousness itself, has a unified neurological substructure [Kolb and Whishaw, 2001]. If one subscribes to an out and out connectionist materialism with respect to these matters, we would have it that the phenomenological experience of a unified consciousness is an illusion. If that were so, then it would hardly matter that a mechanized logic of counterfactual reasoning is incompatible with the unity of consciousness.

Even if we allowed that phenomenologically unified manifestations of consciousness were not always or in all respects illusory, we have already seen in section that there are strong information-theoretic indications that the conscious mind is neither conscious enough nor efficient enough for the burdens of a human subject's total cognitive agenda. In the face of mounting evidence that substantial and essential aspects of cognition operate down below, it seems an unattractive dogmatism to refuse to logic any purchase there. Add to that, that plausible mechanization of cognitive processes such as hypothetical reasoning require that we postulate subconscious and prelinguistic performance in a way that downgrades the role of a unified consciousness, not only is the logic of down below given some encouragement, but so to is its algorithmic character. So we think that it must be said by everyone drawn, for reasons of the sort we have been examining, to the logic of down below is in all consistency pledged to reconsider, with more favour than proposed by researchers such as Jacquette, the plausibility of mechanical models of abductive practice.

6.8 Abduction in Neural-Symbolic Networks

We suggested earlier that part of an individual's cognitive wherewithal might well be representable in a connectionist logic. Such was the conjecture of Chapter 3. We have also touched briefly on non-representational systems in which various constraints are satisfied but not by following rules for their satisfaction — yet another comment on the process-product dichotomy. An attraction of such systems, apart from their intrinsic interest, is the hope they offer, albeit with qualification, to logicians who take seriously the massively plain fact of cognitive performance " down below". Connectionist approaches also offer some (conjectural) relief on the score of computational complexity. It must be admitted, however, that the relief is rendered ambiguously. On the one hand, parallel distributed processes are intuitively plausible subduers of complexity, echoing the old saw that many hands make light work. On the other hand, they are often highly complex systems to implement mechanically. The one fact does no discredit to the other. Whatever the complexities of simulating parallel processes, and whatever the complexities in the evolution of them in human beings, when they operate in human beings there is every reason to think that they achieve the economies attendant upon the efficient evasion of complexity-overload.

Connectionist logics are still in their infancy. But enough is already known about them to make it possible to say that, in their standard forms, they are not especially well-suited to abduction. The problem is that the operation of connectionist backwards-chaining is too coarsely grained for the selective refinements of hypothesis-generation and hypothesis-engagement. Our principal task in this final section of the chapter is to consider ways of mitigating this difficulty.

To this end, we will sketch a new parallel model for abductive reasoning based on Neural-Symbolic Learning Systems. Our further purpose, in addition to the benefit of possible parallel speed-ups, is to sketch an integrated reasoning and learning system. This requires the use of simple neural networks to which standard, off-the-shelf learning algorithms can be applied. A third objective is to give the reader an early taste of what lies in store in the next, and final, part of the book, devoted to formal models. The main problem to tackle here is the fact that neural networks work bottom-up, while for abduction we would like to reason top-down. One might be tempted to revert the network in an attempt to reason using abduction, but this would not work in the case of neural networks, as the following example illustrates.

Example 6.18 *Consider the Neural-Symbolic Learning System of Figure 6.3. It encode the logic program $P = \{r_1 : a, b \to x; r_2 : c \to x; r_3 : x \to y\}$. Each rule r_i is mapped from the network's input layer to its output layer through a hidden neuron N_i such that output is activated if the input is satisfied. For example, output neuron x will be activated if either input neurons a and b are both activated, or if neuron c is activated. In addition, input and output neurons having the same label (e.g., x) are linked through a feedback connection with weight 1 connecting the output to the input layer of the network. This is responsible for implementing* chains *such as $a \to b$ and $b \to c$, in the network. In the case of P, this is how, given a and b, the network would have y activated, via neuron x. From using abduction on P, we know that $\{a, b\}$ is a possible explanation for x and so is c. If we were to simply revert the network in an attempt to compute explanations $\{a, b\}$ and $\{c\}$ given hypothesis x, we would have a relation instead of a function from the input to the output of the network. As a result, a standard neural network (which computes functions, and not relations) would not be able to distinguish $\{a, b\}$ and $\{c\}$ as two alternative explanations for x. Instead, $\{a, b, c\}$ would be activated given x.*

An alternative to the problem discussed in this example would be not to reverse the network but to treat different sets of input values to input neurons a, b, c as *hypotheses*. Abduction in this case would be the process of presenting such different input values to the network and inspecting the output for additional hypotheses. For example, if a and b were activated in the input layer of the network of Figure 6.3, and we treated these activations as hypotheses instead of as facts, we would be able to conclude that, since a and b activate x, $\{a, b\}$ is an explanation for x. Similarly, $\{c\}$ would be an explanation for x, and x an explanation for y, but not $\{a\}$ alone, or $\{b\}$ alone. In addition, $\{a, b, c\}$ would still be an explanation for x, but a non-minimal explanation.

The problem with this approach is on the choice of inputs to select. Would we, for example, select all combinations of inputs? In this case, we would have an exponential complexity algorithm with each input being either activated or deactivated, and 2^n input values to check, where n is the number of input neurons (atoms in the corresponding logic program). What we really would like to be able to do is to reason top-down or goal directed way. We would like to be able to activate

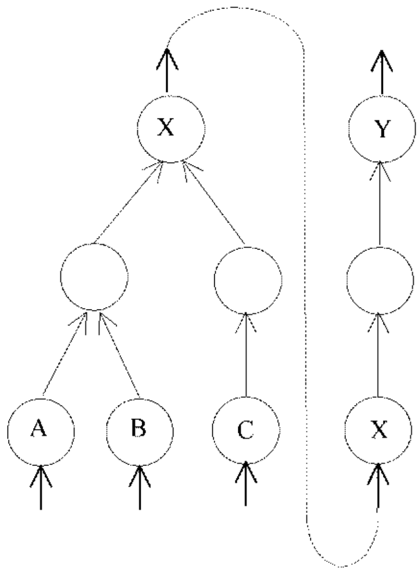


Figure 6.3 A Neural-Symbolic Learning System

y as a hypothesis in the network and obtain $\{x\}$, $\{a, b\}$ and $\{c\}$, (for example), in parallel as alternative possible explanations for this hypothesis.

A solution to this problem lies with *connectionist modal logics* [d'Avila Garcez *et al.*, 2002; d'Avila Garcez and Lamb, 2004]. Modal Logics can be implemented in Neural-Symbolic Learning Systems with the use of an ensemble of neural networks, each network representing a possible world. The use of an ensemble allows for the representation of relations such as accessibility relations, in neural networks. Each network in the ensemble is a simple single hidden layer network like the network of Figure 6.3, to which standard neural learning algorithms can be applied. Learning, in this setting, can be seen as learning the concepts that hold in each possible world independently, with the assessibility relations providing the information on how the networks should interact. In the case of abductive reasoning, we can model the fact that $\{a, b\}$ and $\{c\}$ are possible explanations for x, for example, by having neurons a and b active in a network of the ensemble (say, W_1). Neuron c also active in a different network of the ensemble (say, W_2), whenever neuron x is active in a network W such that $R(W, W_1)$ and $R(W, W_2)$, where R is an accessibility relation. The following example illustrates the idea.

Example 6.19 *Take the same program* $P = \{r_1 : a, b \rightarrow x; r_2 : c \rightarrow x; r_3 : x \rightarrow y\}$. *First, we translate* P *into a modal program by replacing each rule of* *the form $L_1, ..., L_n \to A$ by a modal rule of the form $A \to \Diamond(L_1 \wedge ... \wedge L_n)$. The intuition behind this translation is that $L_1, ..., L_n$ is a possible explanation for A (and thus the use of $\Diamond$). In addition, we label each rule r_i with a world W_i in which r_i holds, and define how the worlds relate to each other (i.e., the accessibility relation $R(W_i, W_j)$), according to the* dependency chains *in P. This will become clearer when we present the algorithm to translate P in the sequel. For now, as an example, suppose we translate r_1,r_2 and r_3 in this sequence. For r_1, we obtain W_1: $x \to \Diamond(a \wedge b)$. For r_2, since there is no chain from r_1 to r_2, we keep r_2 in W_1, and obtain W_1: $x \to \Diamond c$. Finally, for r_3, we define W_2: $y \to \Diamond x$ and $R(W_2, W_1)$, since now there is a chain between r_3 and the other rules using x. Given W_1: $x \to \Diamond(a \wedge b)$, by definition we would like a and b to be true in a world W_0 such that $R(W_1, W_0)$. Similarly, given W_1: $x \to \Diamond c$, we would like c to be true in another world $W_{0'}$ such that $R(W_1, W_{0'})$. Note that, by defining the relation R appropriately, and since in this case we have a different neural network for each world, we can check that when e.g., x is activated in W_1, a and b will be activated in W_0, while c will be activated in $W_{0'}$. Similarly, when y is activated in W_2 then x will be activated in W_1. This allows us to reason top-down, in parallel, and at the same time to keep track of the alternative explanations for our hypotheses.*

Given a modal program, an ensemble of neural networks can be constructed by repeating the procedure for constructing single networks. Figure 6.4 shows networks W_1, W_0 and $W_{0'}$ for the program of Example 6.19. In W_1, whenever input neuron X is activated, we would like neurons $\Diamond A, B$ and $\Diamond C$ also to be activated. This can be easily implemented by properly setting up the connection weights and thresholds of the hidden neurons connecting the input to the output. In addition, whenever output neuron $\Diamond A, B$ is activated, we would like to have output neurons A and B activated in W_0. This is implemented in the same way with the use of hidden neurons in W_0. Similarly, whenever output neuron $\Diamond C$ is activated, we would like to have C activated in $W_{0'}$. Note that the fact that neurons A and B, and neuron C get activated in different networks is responsible for identifying $\{a, b\}$ and $\{c\}$ as two alternative explanations for x.

In addition, similarly to the feedback connections in a neural network (e.g., linking output neuron X to input neuron X in W_1), there might be feedback connections between networks (i.e., from W_0 and $W_{0'}$ to W_1). From W_0 to W_1, there is feedback from A and B to X such that whenever both output neurons A and B are activated in W_0, output neuron X is activated in W_1 (again, this is implemented via an AND gate hidden neuron). Similarly, from $W_{0'}$ to W_1, there is feedback from C to X such that whenever C is activated, X is activated (i.e., output neuron X acts as an OR gate for the hidden neurons that are linked to it). This allows one to reason deductively as well as abductively within the same model. If, for example, for some reason, output neurons A and B are activated in W_0 (say, we force A and B to be activated, or there are other rules and facts in W_0 that make A and B activated), output neuron X will be activated in W_1 (implementing rule $a, b \rightarrow x$). Similarly, if C were to be activated in $W_{0'}$ then X would be activated in W_1. In summary, when activations are propagated forward, according to the accessibility relation, the network computes explanations for hypotheses; when activations are propagated backwards, through the feedback connections in the network, deduction is being performed. This is a very interesting characteristic of the model presented here.

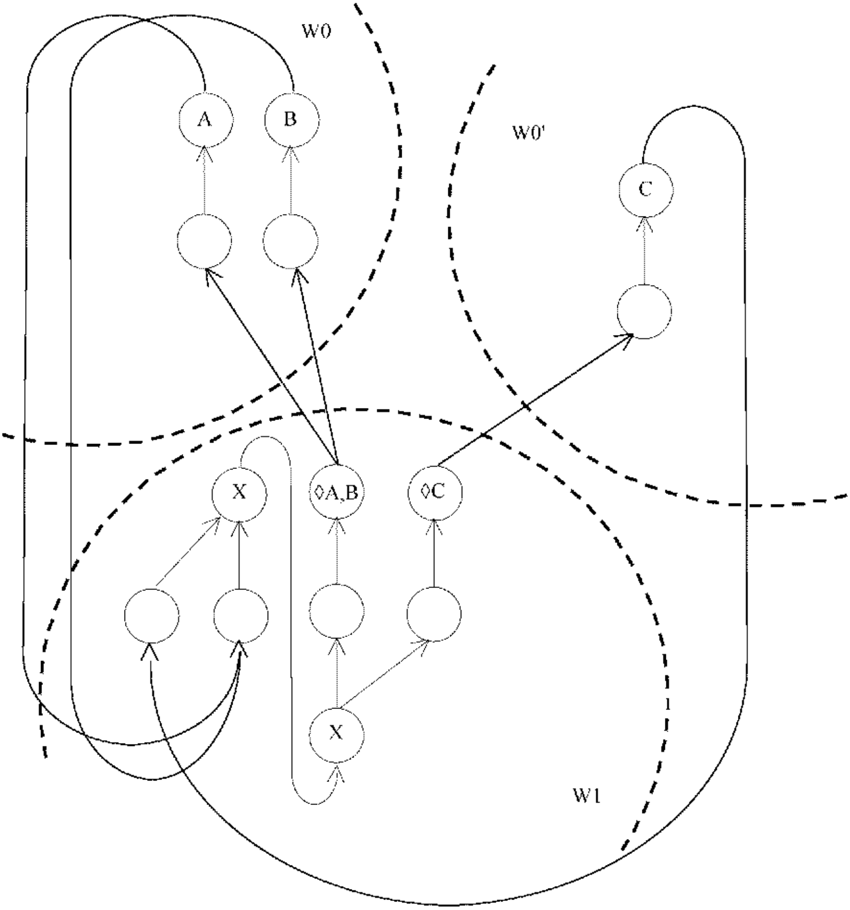


Figure 6.4 Neural Network Ensemble for Abductive Reasoning

Let us now present the algorithms to translate symbolic rules into connectionist networks that reason abductively.

1. For each rule in P of the form $X_1, ..., X_n \to Y$ do:
    (a) If Y has not been assigned a world:
        i. Assign a new world W_i to Y;
    (b) If $X_1, ..., X_n$ have not been assigned a world:
        i. Assign a new world W_j to $X_1, ..., X_n$;
    (c) Make $R(W_j, W_i)$;
2. Make $R(W_k, W_i)$ for any rule W_i in which $\Diamond$ is used; and
3. Call *Modalities Algorithm* to build the network ensemble.

To run the network ensemble to compute explanations (forward), proceed as follows: (a) Activate a number of neurons (hypotheses) at time t_1; and (b) Check neuron activation at times t_2, t_3,... until the ensemble is stable (i.e., until activations at time t_i = activations at time t_{i+1}).

Theorem 6.20 *Each set of activations at time t_i in each world is a possible explanation for the activations at time t_{i-1}.*

To run the network ensemble to compute answer sets (backwards), proceed as follows: (a) Introduce a number of facts to the ensemble by setting certain output neurons as being always active, regardless of the input; and (b) Let the ensemble become stable.

Theorem 6.21 *The set of activations in the ensemble will be the set of what can be deduced from the facts given.*

Connectionist logics are a lot like human infants. They are difficult to raise, and then take quite a long time to grow up. We offer the suggestions of the present section in the hope that it might be said, sooner rather than later, that connectionism in logic has to some extent grown up and that the days of its infancy are numbered, if not ended quite yet.

Chapter 7

The Characteristic and the Plausible

> Snakes are reptiles. Telephone books are thick books. Birds lay eggs. The duck lays eggs. The duck has coloured feathers. Guppies give live birth. Austrians are good skiers. (The) Chinese eat dogmeat. Americans are good baseball players. Germans eat horsement. (uttered after WWII). The tiger even eats grass and soil (under the right conditions). Unicorns have one horn.
>
> Pelletier's Squish

> Plausible reasoning is relatively basic: it reflects a relatively primitive — but not less important — mode of reasoning.
>
> Nicholas Rescher

7.1 The Open Door

As we have seen in the discussion of the Cut Down Problem in chapter 3, it is natural to think of abduction problems as involving what we might call *candidate spaces* and *resolution procedures*. We have already made the point that although both relevance and plausibility are constituent features of what we have been calling filtration structures, there is no empirical evidence that real-life abducers select their hypotheses by constructing filtration structures, or even examining them. We have had ample occasion to assert that a logic of abduction encompasses the five sublogics of consequence, conclusion, generation, engagement and discharge. We have also called attention the the roles played by relevance and plausibility if the construction of filtration structures. There is, of course, a large literature in which a theory of relevance is taken as a part of logic, and there is a lesser, though not insignificant, literature in which the same approach is taken to plausibility. We say again that we have no interest in multiplying logics beyond necessity— or usefulness. Relevance and plausibility logics pre-date any thought of their invocation here. This constitutes a *prima facie* case for the abductive logician to pay them due heed in his own deliberations. We propose to honour this claim in the following way. At those places at which plausibility and relevance are present in the account of abduction, we shall make use of such logics of these things as facilitates their contribution to the understanding of abduction. Subject to this qualification, we shall assume that in his involvement with considerations of plausibility and relevance, the abductive agent reasons in ways that, in principle, a logic of plausibility and a logic of relevance can say something about. As with the other aspects of cognitive practice, we do not suppose that such logics tell the full stories of their respective subject matters. But, equally, we suppose that the parts of the stories that such logics do manage to tell are matters requiring the abductive logician's attention.

There is, however, a prior matter to consider. In chapter 4 we discussed the Peircean idea of surprise. In the present chapter we return to the theme. Consider now the following case.

> It is late afternoon on an overcast November Thursday. Harry arrives home at his usual time, and parks his car in the garage at the end of the back garden. Sarah, his school-teacher wife, always comes home about an hour later. Harry's practice is to enter and leave the house through the back door. But since the garage is only big enough for a single car, Sarah parks in the street in front of the house, and enters and leaves the house through the front door.
>
> Dora helps with the cleaning every Tuesday morning. Apart from her, no one but Sarah and Harry have occasion to be in the house or the means to gain entry; the couple's children are adults and have left the family home long since.
>
> Having parked his car, Harry leaves the garage and makes his way along the path to the back door. He hasn't taken many steps before he sees that the back door is wide open.

7.1.1 The Element of Surprise

The open door is the trigger of an abduction problem. We stipulate that the trigger was *counter-expected* for Harry. It is not just that 'the door is [or will be] open' was not in Harry's K-set before the trigger presented itself; rather the K-set contained at that time the fact that the back door is never open under these conditions (weekday, end of the day, before Harry arrives home, when Dora's not there, etc.).[1] The case, therefore, incorporates the Peircean element of surprise.

The immediate significance of the counterexpectedness of the trigger is that it gets Harry's attention and constitutes a problem for him. It is far from universally the case that the presence of counterexpected situations is a problem for those who discover them. So we must provide that Harry's K-set at the moment of discovery contains not only, 'The door is never open', but also, 'The door is not *supposed* to be open'. We now have the means to motivate the open door as problematic for Harry.

An abductive trigger is not just an occurrence of which an agent is conscious. It is often an occurrence of a type that rises to a second grade of statistical abnormality. It is an event whose occurrence is not only noticed and attended to by the agent; its occurrence is, as we said above, *uncharacteristic* in some sense. Accordingly,

♡ **Definition 7.1 (Surprise, first pass)** *Something is a surprise for an agent, in Peirce's sense of "surprise", if it is both unexpected and uncharacteristic.*

We should not over-blow the element of surprise. A surprising event may astonish us, bowl us over or mystify us, but none of this is essential to Peirce's notion. A surprise in his sense is something unexpected and out of the ordinary. Unexpectedness here is an epistemic notion. Something is unexpected when its occurrence is something that one would not have known about; it is something that could not have been forecast solely on the basis of what one knew at the time. This leaves it open that an event that is unexpected in this epistemic sense might have been expected in some other sense. (Perhaps the agent in question had a hunch that this event would occur). Similarly, a good many more things are uncharacteristic than any agent will actually find to be so. We must amend the definition of surprise to take this fact into account. So

Definition 7.2 (Surprise refined) *An event or state of affairs is a surprise for an agent X when its occurrence is not something that X would have known and whose occurrence X* finds *to be uncharacteristic in some way.*

The factor of uncharacteristicness is necessary if any notion of surprise is to lay claim to a pivotal place in the logic of abduction. Surprising events are those whose occurrence puts an agent at an epistemic disadvantage. Unexpectedness is not in

[1] Formally, let Δ be the non-monotonic database available to Harry at the moment of observation. We have $\Delta \mid\!\sim$ *door closed*. Harry observed $\sim$ door closed. A revision is needed. Compare this case where $\Delta \not\mid\!\sim A$ and $\Delta \not\mid\!\sim \sim A$ and A is observed. Then an explanation is needed. The 'story' will not admit as 'acceptable' taking A *itself* as explanation.

general a marker for this. Most of the things whose occurrence an agent comes to recognize as something he wouldn't have known about are met with a certain passivity. Most of what happens is unknown to everyone. When they become known, most retain the feature that although known now, they wouldn't have been known earlier. The epistemic disadvantage that a surprising phenomenon creates for an agent is not that it is not known. (Now it is known, and that, we should think, represents a change from what wasn't known to what is. How can there be any epistemic disadvantage in this?) The proposed factor of uncharacteristicness is supposed to help answer this question. It is meant to indicate that the occurrence in question places the agent at a disadvantage in respects other than what he presently knows of it. For this to be true, there must be something an agent would desire to become aware of over and above his now present knowledge that the event in question has occurred. Accordingly, the occurrence of an event or the presentation of a state of affairs will count as a Peircean surprise if, in spite of the fact that its occurrence is now known, it presents the agent with an additional cognitive target which cannot be hit with what is now known.

But how does it come to be the case that uncharacteristicness is a marker for this? There is both a general and a particular answer. The general answer is that the acquisition of new knowledge is not typically the generator of this kind of collateral cognitive effort. Consider some cases. Sarah, reading from the paper, tells Harry that the Berlin Philharmonic will play next month. "Good", says Harry. "I'll get tickets". Or, you look out the window and see that it's snowing. "More snow", you mutter (it is January). The absorption of new knowledge is dominantly passive in this sense. The reason is partly economic. No one has time to launch a supplementary enquiry every time something new chances to be known. Collateral cognitive effort is therefore uncharacteristic. What is characteristic in the general case is the passivity of cognitive agents towards the new. In particular cases — this is our second point — the uncharacteristicness pertains more directly to the event itself. Thus Sarah would not be home to-day; it would not be snowing in January in Honolulu; and so on. It suffices for the requirements of the Peirceian notion of abductive surprise that uncharacteristicness in either sense be in play.

We saw from our discussion of Planck the importance of keeping in train the necessary distinction between an occurrence or state of affairs that initiates an abduction and the target at which the abduction is aimed. For Planck, the initiating circumstance was the disunification of the radiation laws for black bodies and the target was to achieve the opposite of this. In the usual psychological meaning of the term, Planck did not find the initiating conditions surprising; but he did find them irritating. But both elements of Peirceian surprise are present in this example. Planck's target could not be hit on the strength of what was then known of physics, and the very fact of his target's existence turns on the fact that it is not characteristic of mature physical science to have disunifications of this sort.[2]

Again we note that not all explanations are solutions of abductive problems. Occurrences can be entirely in-character and yet still enter into stable and sound explanations. And not all abductive problems are of the kind we are about to examine, as we have seen.

Harry wants to figure out 'what's going on here!'. To this end, he determines a *space of candidates* (though his determination, let us note, might well be virtual). The candidates are candidates for the role of what explains or might explain the trigger-event. Since the trigger presents Harry with a problem, we can say equivalently that the space of candidates contains alternative resolutions of Harry's abduction problem. It is an empirically pressing question as to how big, and how varied, candidate spaces actually are or could be.

One fact about Harry's present situation is that his candidate space is very small. Yet the number of states of affairs which, if they obtained, would explain Harry's trigger-datum is very large. Harry's actual candidate space is a meagre proper subset. This the model must take into account. To this end, we put it that Harry has intermingled his interest in explanation with a *relevance logic*. A relevance logic in the sense required here has little directly in common with what relevant logicians occupy themselves with — viz. the consequence and consistency relations (e.g., [Anderson and Belnap, 1975; Read, 1988; Dunn, 1994; Woods, 1989]), but rather is a logic in the sense of a set of algorithms which enable the human agent to discount, for the most part automatically and with considerable alacrity, information unhelpful to a task at hand. (See [Gabbay and Woods, 2003a]). Setting a candidate space for an abduction problem involves performing a reduction from sets of potential explainers to sets of candidates for the status of actual explainer. It is an abductively sensitive instantiation of the more comprehensive operation of cutting down from irrelevant although theoretically applicable information to information relevant to whatever the task at hand happens to be. Again, we emphasize that there is little known empirically about how these processes work.

Relevance in our sense is not a dyadic relation on sentences. It is a set of triples $\langle I, X, A\rangle$, informally interpreted to mean:

> Information I is relevant for agent X to the extent that it advances or closes X's *agenda* A [Gabbay and Woods, 2003a].

Because these processes occur with great speed and mainly automatically, and because they produce such selectively scant outputs, we think of the human reasoner in such circumstances as activating instrumental (and perhaps causal) algorithms. Accordingly, the model posits what we are calling an explanation program,

[2] Which is partly why the present alienation of quantum and relativistic physics rankles so.

and a logic of relevance, a primary function of which is to fix candidate spaces in appropriately economical ways.

We return to our case.

> Harry is perplexed. The door is open when it shouldn't be. What's happened? Perhaps Sarah has come home early. Maybe Dora couldn't do her chores on Tuesday, and has decided to make things up today. Or perhaps Harry, who is the regular user of the back door, forgot to close it when he left the house this morning. On the other hand, it could be a burglar.

We identify these four alternatives by the obvious names. Harry's candidate space is the set {**Sarah, Dora, Harry, Burglar**} .

7.1.2 Plausibility

The description continues:

> Harry entertains the candidate **Sarah**. No, he thinks, Sarah is at school. She never comes home before 5.30 and it's only about 4.40 now. Harry also rejects **Dora**. Dora hasn't missed a Tuesday in years and years, and, in any case, she would have mentioned it if she were planning to come today. As for **Harry**, Harry recognizes that he is sometimes forgetful, but he's never been forgetful about shutting he door when he leaves for his office. This leaves **Burglar**. 'Oh, oh', says Harry 'I think this might be a break-in!'

Harry has riffled through the candidate space and has rejected **Sarah, Dora** and **Harry**. The rejections are judgements of *implausibility*, each rooted in what appears to be a belief about what is generally the case. It is not impossible that Sarah is in the house, but it is implausible because she never is in the house at this time on a weekday. We note that generality-claims are not restricted to classes, but can apply with equal effect to individuals. Of course, like all such claims, these are generalizations that license defaults for Harry; each is made in at least implicit recognition that they needn't be disturbed by true negative instances. That being the case, the imputed generality is not one of universally quantified conditionality but something slighter. There are two candidates to consider. One is that the utterance in question expresses a generic claim. The other that it expresses a claim about what is usual. (We take up their difference below.) The inference that it isn't **Sarah** since Sarah never is at home at such times, is made in recognition of its requisite defeasibility. This suffices to make the rejection of **Dora** a rejection founded on implausibility. Accordingly, we might posit for Harry a plausibility logic. It suggests that Harry possesses an inference schema to the effect that

Proposition 7.3 (Implausibility) *If S is in an agent's candidate space with regard to an abduction problem A, and the agent holds a generic claim G concerning the subject matter of S, and G is incompatible with S, then X infers that the (propositional) implausibility of S's occurrence defeasibly disqualifies it at a solution of A.*

Having reasoned in the same way with regard to **Sarah** and **Harry** and provided that Harry's candidate space has taken in no additional members, Harry opts for **Burglar** (see just below). Speaking more realistically, he finds himself in the grip of **Burglar**. He not only "thinks" that it may be so; he is seized with apprehension that it may be so. (This is reminiscent of Peirce's insistence of an idea).

We may now say that Harry's plausibility logic contains an Eliminative Induction rule in the manner of Bacon, which he applies at his problem's *resolution point*:

Elim Induction Either **Sarah** or **Dora** or **Harry** or **Burglar**. Not **Sarah** and not **Dora** and not **Harry**. Therefore **Burglar**.

At a certain level of abstraction there is no harm in assuming an application of *Elim Induction.* Something like it, though different, is actually in play, since 'not X' here means 'X is not a plausible solution of...', and so does not strictly negate X.

Candidate spaces have the potential for *looping*. If **Burglar** is the only candidate not yet eliminated from Harry's space, then if Harry sufficiently dislikes **Burglar**, he may find that the candidate space has taken in a new member, e.g., 'The Sears man has come'. We put it that the structure of explanation will tolerate only low finite looping at most. Another possibility is *retraction.* Harry might dislike **Burglar** and find that he has revisited, e.g., **Sarah**. Here, too, an explanation will permit only low finite oscillation. In general it is realistic to allow for options in addition to the one-option or no-option approach. By and large, the failure to find an option which the abducer likes is a trigger of (cautious) investigative actions.

7.1.3 A Resolution Point

Harry's solution of his abduction problem involved the elimination of all candidates but one. **Sarah** was eliminated by the generalization, 'Sarah is never ever home early', ($G^{\mathbf{S}}$ let's call it), **Dora** by its corresponding $G^{\mathbf{D}}$, and **Harry** by $G^{\mathbf{H}}$. This leaves **Burglar**, or $\mathbb{B}$ for short. Harry plumped for $\mathbb{B}$. But why? $\mathbb{B}$ too is attended by its own $G^{\mathbb{B}}$, 'Burglaries don't occur in this neighbourhood'. Why didn't $G^{\mathbb{B}}$ cancel $\mathbb{B}$, just as $G^{\mathbf{C}}$ cancelled **C**, $G^{\mathbf{D}}$ cancelled **D**, and $G^{\mathbf{H}}$ cancelled **H**? It is evident that Harry's eliminations have left him with the following termination options:

TO_1 $G^{\mathbb{B}}$ cancels $\mathbb{B}$, and there is neither retraction nor looping. The abduction problem crashes.

TO_2 $G^{\mathbb{B}}$ cancels $\mathbb{B}$ and either retraction or looping occurs. The abduction problem is renewed short of resolution.

TO_3 Although $\mathbb{B}$ is a negative instance of $G^{\mathbb{B}}$, it does not cancel it (recall that $G^{\mathbb{B}}$ is a *generic* claim), and there is neither retraction nor looping. $\mathbb{B}$ solves the abduction problem.

7.1.4 How to Get Determinacy Out of Indeterminacy

It is necessary that the plausibility logic give abducers the means of selecting from multiples such as $\{TO_1, TO_2, TO_3\}$. How does an abducer know when to bet his confidence in the exhaustiveness of his candidate space against the fact that the solving candidate is a negative instance of generic claim in which the abducer also has confidence? Or how does he know when to bet in reverse; that is, when to bet his confidence that the last surviving member of the original candidate space is indeed cancelled by the generic claim of which it is a negative instance? Similarly, how does the abducer know when and when not to retract a prior decision of candidate cancellation in order to evade selection of a TO_i that involves a candidate which is not only a negative instance, but which he now thinks is cancelled by it?

It is useful to repeat that our present purpose is to model Harry's actual case. In real-life abductive situations, such as that of Harry, it is rarely helpful for Harry to ask himself why he went the TO_i-route rather then the TO_j-route. There is ample evidence that the routines of actual abductive practice are not much accessible to human introspection. So the question before us constitutes an abduction problem all of its own for the theorist who is trying to figure out Harry's abductive situation.

We put it that the plausibility logic favours the following inference schema which, we emphasize, can only have defeasible legitimacy.

> **The Auto Rule** To the extent possible, favour the option that has an element of *autoepistemic backing*. For example, in the case we are investigating, the generality claims about Sarah's never coming home early, is likely to be underwritten by two factors of autoepistemic significance. One is that if it were indeed true that Sarah never comes home early, this is something that Harry would know. And if today were to be an exception to that rule, this too is something that Harry may well have knowledge of.

Autoepistemic inferences are presumptive in character. Given that a candidate hypothesis is not known to be true, it is presumed to be untrue. As long as the degree of epistemic value of the presumption is less than the levels attained by

Harry's K-set, the *Auto Rule* preserves the abductive character of Harry's problem. **The Auto Rule** bids Harry to favour **Burglar** since, **Sarah, Dora** and **Harry** are subject to the requisite autoepistemic factors. So is **Burglar**. Had there been previous burglaries, Harry would have heard of them. But he hasn't. Why then do we say that **Burglar** is the best hypothesis to select if they all pass the autoepistemic test? The answer is that the autoepistemic test is supplementary to the characteristicness test, and it is the characteristicness test that **Burglar** doesn't do very well with. Rightly; for, short of locked gates and streets filled with security guards, how could it be in the nature of what constitutes Harry's neighbourhood is that burglaries not happen there.

Human reasoners are natural conservatives. They lean towards explanations that deviate from the normal as little as is consistent with getting an explanation. It is the sort of favouritism that Quine had in mind in proposing his *maxim of minimal mutilation* as a principle guiding scientific theories in handling observationally recalcitrant data. It is a maxim which favours adjustments to a theory as modest as get the the problem solved.

We assume such a principle to be in play in Harry's plausibility logic. We assume it is in the form of least possible deviation from the norm. Having it in play seems to count against Harry's selection of **Burglar**. For if there were indeed a burglary, this would be quite a deviation from the normal, whereas if Sarah had come home unexpectedly, or Dora had switched her day without telling anyone, or Harry for the first time in his life had forgotten to shut the door, these would be lesser deviations from the norm.

It falls to Harry's plausibility logic to hazard an answer to the following question.

> Given that one of $\mathbf{S}$, $\mathbf{D}$, $\mathbf{H}$ and $\mathbb{B}$ does obtain, is there a way of selecting one as the most plausible?

Whether the logic is capable of furnishing an answer in every case, it would appear that if there is a least plausible of the present lot it is $\mathbb{B}$. If so, isn't it the wrong answer for Harry?

We note that in finding thus for $\mathbb{B}$, the logic equates least plausiblity with greatest deviation from a contextually indicated norm. But this is not an answer to Harry's abduction problem. Harry's abduction problem is to determine what best explains *the open door*, which was Harry's trigger. The abductive task was not to determine who most plausibly was in the house. The door is wide open in late afternoon of an overcast day. Given that Sarah had come home early or that Dora had switched her day, how plausible is it that the door is left wide open on either of these accounts, when a burglar would have had no particular motivation to close the back door of the empty house that he had burgled without incident earlier in the day?

This leaves Harry with the task of deciding between **Burglar** and **Harry**. In opting for **Burglar** he was certainly not making a canonical choice. A different person in exactly the same situation except for this difference might well have opted for the explanation in which he did indeed forget for the first time in his life to close the door. But in Harry's actual situation, the autoepistemic factor is clinching:

> If I had left the door open, I would have remembered. But I don't, so I didn't.

7.1.5 Alternatives

Consider a second, hypothetical, case, somewhat like the Harry-case (which was a real situation). Everything is as before, except for factors we shall now detail. **Sarah** is in Harry's candidate space. As Harry begins the winnowing process, he recalls that Sarah is in a distant city assisting their daughter with a new baby. There is no factor of generality about this, nothing that would justify the tentativeness of a plausibility judgement. Sarah's being in that place is a *fact*. It is a fact incompatible with **Sarah**. End of story.

The present example presents us with two modelling options. One is to have the plain fact that Sarah is away bear on the model at the candidate space-*construction* stage. If it is a fact that Sarah is away and Harry knows it, then **Sarah** doesn't belong in the candidate space to begin with. On the other hand, since Harry has appeared briefly to have forgotten Sarah's whereabouts only to have recalled them after the candidate space was fixed, there is a supplementary mechanism at work at the stage of *winnowing*.

SupplMech: If S is in the candidate space of a subject X in relation to an abduction problem A, if F is a fact incompatible with S, if X did not know (recall) F during the construction of the candidate space, but X called F at a subsequent stage in the abduction process, then X categorically and summarily removes S from the candidate space.

How reasonable is it that a plausibility logic would embed a rule such as *SupplMech*? We think it inadvisable to press the matter over-much. Either *SupplMech* is in the plausibiltiy logic, and permits candidate exclusion on factual grounds as limiting cases of implausibility determinations; or *SupplMech* is an external supplementation of the plausibility logic. But either way, *SupplMech* emphasizes the importance of the fact that begins as a memory search, and that if such a search is successful, the abduction problem ends. In the limit case, it never becomes an abduction problem. *SupplMech* also presents us with a problem. It has to do with the natural tension between the requirements of nescience. The same may be said for the *The Auto Rule*. In the first case, it is simply assumed that memory trumps plausibility. If a proposition P recommends itself to a reasoner on grounds of its plausibility, remembering that (in effect) *not*-P is enough to eliminate P. Similarly, if, as in the second case, a possibility that P is not attended by the agent's memory that P, it too is excluded. The two cases bear on the nescience condition. An agent is met with an abduction problem only if his target T is not attainable with the resources contained in his knowledge-module K. Since ignorance is an invariant feature of the abductive process, the point at which draws the conclusion $C(H)$, is a point at which H must fail to attain the epistemic standard evinced by K. This provides that at that juncture H possesses either diminished epistemic virtue or none at all. Since any candidate for the role of H must be subject to like epistemic impairment, it would appear that a remembered fact to the contrary should trump it, and that a failure to have a memory that confirms it should likewise prevail over it. Taking our cases in reverse order: (1) How likely is it that Harry's failure to remember that he left the door open will *always* have an epistemic ranking higher that the H it putatively excludes? (2) How likely is it that Harry's recollection that Sarah is away carries the epistemic clout to trump any H to the contrary? There is no *à priori* method for answering these questions in a general way. *Phenomenologically*, the same can be said for the autoepistemic infererer, for whom the failure to have a confirming memory of H disconfirms it. But phenomenological indicators are not always reliable, as we shall soon see in somewhat greater detail. For the present, it suffices to sound an admonition. Both the *The Auto Rule* and *SupplMech* are defeasible instruments, and must be deployed with requisite caution.

Our second hypothetical case raises further questions about how Harry actually operated in the test case we are attempting to model. For example, why have we represented Harry's selection of **Burglar** as the result of the exhaustive elimination of the candidate alternatives on grounds of their implausibility? Why did we not instead represent his choice as having turned on his finding **Burglar** to have greater plausibility than its rivals? Why can't the processor of an abduction problem produce a solution by opting for the most plausible alterative in his candidate space, without the nuisance and the cost of running thorough a number of implausibility determinations? There is no reason in principle why this couldn't happen. (In fact, there is rather a lot of evidence from empirical investigations of learning that suggests that this is often what does happen.) When it does, his abduction exhibits the following structure:

Proposition 7.4 *If* $\{S_1, \ldots, S_n\}$ *is the candidate space of a subject* X *in relation to an abduction problem A, and* X *judges* S_k *to be more plausible than any other, then* S_k *gives the resolution point of A for* X. *If* S_k *is chosen in the manner above, then* X *is committed to finding the other* S_j *less plausible than* S_k *but not necessarily implausible.*

We see that the winnowing of candidate spaces can proceed by way of judgements of negative plausibility and judgements of positive plausibility. Seen the first way we are reminded that membership in an abducer's candidate space is not a matter of plausibility. It is a matter of *relevance*. However, seen in the second way, we are reminded that membership in a candidate space is not forbidden to relevant potential explainers whose occurrence is also implausible.

Then, too, there is the following kind of situation. Harry comes home to find thick black smoke pouring from the house. 'Good heavens', says Harry, 'the house is on fire!'. The occasion is backed by a generality-claim of which it is a positive instance, and it instantiates no more plausible a generality than it.

Abductions of the negative plausibility sort resemble belief-revision by way of backwards propagation [Gabbay, 2000]. Such structures have had a lot of attention from AI researchers, and are comparatively well-understood. For all their apparent economic advantages, it is not as easy to model abductions via judgements of positive plausibility; nor is it at all easy to show this kind of approach to economic advantage.

We close this section with two related matters. Why, we again ask, did we represent Harry as finding **Burglar** plausible just on the grounds that he found its rivals implausible? It is a question invited by a tempting confusion which we should try to discourage. So, to repeat: we do *not* argue that Harry's actual selection of **Burglar** hangs on a favourable judgement of plausibility. Harry's decision is captured by the following

Choice Rule: If a subject X has a candidate space $\{S_1, \ldots, S_n\}$ in relation to an explanationist abduction problem A, and if X has rejected all the S_i but S_j for their implausibility, it suffices for the selection of S_i that nothing *else* of explanatory potential has sufficient plausibility for solution. (Assuming solvability, no retraction and no looping).

7.2 The Piccadilly Line

We now come to the interesting question of the interplay of the remembered and the plausible in contexts of abductive reasoning. This, too is a real case. Names have been changed in the interest of privacy. Dave is an academic at a distinguished university in central London. Buzz is an academic at a distinguished university in the State of New Jersey. Dave and Buzz have known one another's work for years, and they see each other from time to time at conferences. Only two weeks previously they were keynote lecturers at a conference in Canada. It was pleasant for them to have renewed their acquaintance.

Dave lives in North London. Greater London is made up of five zones, with central London the inner dot and outermost London a large belt called zone five.

Dave lives at the juncture of zones three and four. Early each morning he boards a Piccadilly Line train, which gets him into central London by about 8.00.

Today Dave boards the train at the usual hour. As he takes his seat, he sees that directly opposite, immersed in a book, is Buzz, or someone who looks just like Buzz. Now Buzz is a well-travelled person. Like nearly everyone in his situation, a trip to London would be a trip to central London — to the University of London, the British Museum, the West End, and so on. Bearing in mind that the man who looks just like Buzz was already seated when Dave boarded, Buzz had to have originated his journey in zone four or five. But again, zones four and five hold no professorial, cultural or touristic interest for people like Buzz. So Dave concludes that this isn't Buzz after all. (Actually he came to *see* that it was not Buzz— *une force majeure* again.)

It bears on our discussion that Buzz has a physical appearance that distinguishes him almost as much as his professional work has done. He is suavely slender, with an elegant van Dyck beard and kind of benign charisma. Buzz's look-a-likes do not come thick on the ground. But look-a-like is what Dave thinks.

Dave's thinking so is the result of how he handled an abduction problem. The trigger was Buzz's appearance in the outer reaches of greater London. It is an appearance that contradicts the pertinent generalities. People like Buzz come to London to give lectures to the British Academy. People like Buzz come to London for the Royal Ballet in Covent Garden. People like Buzz enjoy shopping in Jermyn Street. People like Buzz prefer to stop at good hotels or clubs where they may have reciprocal privileges. What's more, although it is true that Buzz has lots of European relatives, they all live on the continent.

Still the man sitting opposite Dave is the spit-and-image of Buzz. Part of Dave's abduction problem is to entertain what might bring Buzz to North London at the break of day.

Dave could not have had this problem had he not been prepared to consider alternatives. Two things are noteworthy about Dave's situation. Although he is on friendly terms with Buzz, Dave does not greet him. Dave is not shy; so this is odd. Even so, Dave's problem is not to ascertain whether his fellow passenger is Buzz, but rather to figure out why Buzz *would be* in this wholly counterexpected situation. Dave's case contrasts with Harry's. Harry could bump into Sarah in the most shockingly improbable place on earth and not be in the least doubt that it is Sarah that he sees.

Dave solved his abduction problem by inferring that his fellow traveller was only a look-alike. In so doing, he overrode the best explanation of why this man looks exactly like Buzz, namely that he *is* Buzz. Bearing on this was the failure of Dave's predecessor abduction problem, which was to sort out why Buzz would be in this part of London.

In the absence of further information — of information we could have got just by asking — Dave had no answer to give. He had lots of *possible* answers, of course, including some relevant possibilities. Buzz may have been lost; perhaps he has friends in North London; perhaps he is an insomniac and is riding the trains to fill unwanted time.

But Dave had no basis for winnowing these possibilities via considerations of plausibility and implausibility.

As the train sped on, Dave found himself adjusting to his newly arrived-at belief in interesting ways. Buzz is quite slender, and this man is quite slender; but not, as Dave now sees, as slender as Buzz. Buzz has a beautifully sculptured van Dyck beard, as does the man opposite; but it is not as elegantly shaped as Buzz's beard. The facial resemblance is also striking, but not really all that close, as Dave now sees.

Buzz's look-a-like has been reading all this time. He now marks his page and looks up. 'Hi, Dave!'

Buzz and Dave then fall into an amiable conversation, during which it comes out that Buzz had attended a wedding in North London and stayed there with friends overnight. He was on his way to Heathrow to catch the first flight to Leeds where he would give a lecture that evening. After cheerful good-byes, Dave got off at Holborn, and Buzz resumed his reading.

Of course, Dave got it wrong. Abduction is a defeasible process; getting it wrong is not anything to be ashamed of as such. Dave was put wrong by the tug of what he took to be characteristic of (people like) Buzz. It seems that Dave was running something like a Buzz-*script*. It is also possible that he was also running a first-thing-in-the-morning-on-the-southbound-Piccadilly-line script. At no point at which they might intersect would we find Buzz or anyone like Buzz. (It would seem that not even Dave has a place in the Piccadilly line script). The pull of generality-rooted expectation is very strong. In Dave's case it was strong enough to override the evidence of his senses. This happened because Dave didn't have a plausible account to offer which would have overridden, in the reverse direction. He could not override, 'This is not the place for the likes of Buzz', because he couldn't figure out why it would in fact have been Buzz. And, of course, he didn't ask. Such is the tug of the characteristic.

We end this section on a cautionary note. We have made the point that although abducers seem not to solve abduction problems by constructing or even inspecting filtration structures, it does appear that the factors of relevance and plausibility (and, certainly, explanatoriness) sometimes and somehow guide their efforts.

This may well be so. But why would we suppose that because this is the case abductive agents are running relevance, plausibility and explanation *logics*? Logic, we said, is a principled description of aspects of what a cognitive agent does. If it is true that, in solving abduction problems, practical agents are drawn to relevant plausibilities with explanatory force, relevance, plausibility, etc., enter in a nontrivial way into what logical agents do. So it is appropriate to postulate logics of these concepts. What is problematic is the supposition that an agent's subscription to them might be virtual. The problem is mitigated by the *RWR* (representation without rules) approach to cognitive modelling. On this approach cognitive systems employ representational structures that admit of semantic interpretation, and yet there are no representation-level rules that govern the processing of these semantically interpretable representations, see Horgan and Tienson [1988; 1989; 1990; 1992; 1996; 1999b; 1999a]. Critics of *RWR* argue that it can't hold of connectionist systems [Aizawa, 1994; Aizawa, 2000]. Since we want to leave it open that some at least of the cognitive processing of practical agents, it matters whether this criterion is justified. We think not, although we lack the space to lay out our reservations completely. The nub of our answer to critics of the *RWR* approach is as follows.

1. Critics such as Aizawa point out that connectionist nets are describable by programmable representation level rules. They conclude from this that connectionist nets execute these rules [Aizawa, 1994, p. 468].
2. We accept that connectionist nets are describable by programmable representation-level rules. But we don't accept that it follows from this that connectionist nets should be seen as executing such rules.

As Guarini observes,

> The orbits of the planets are rule-describable, but the planets do not make use of or consult rules in determining how they will move. In other words, planetary motion may *conform* to rules even if no rules are *executed* by the planets [Guarini, 2001, p. 291].

A full development of this defence can be found in [Guarini, 2001]

What, we have been wondering, could a virtual logic be? We propose that a reasonable candidate is the requisite description of a cognitive system seen as a connectionist net that satisfies the condition of the *RWR* approach. It is a logic of semantic processing without rules. We return to this suggestion in chapter 9.

7.3 Plausibility Again

If one takes a filtration-structure approach to abduction, we see the abducer selecting the most plausible of a set of relevant possibilities. The sense of plausibility invoked by such cases is propositional plausibility. In other cases, such as radically instrumental abduction, any element of propositional plausibility that may happen to be be present counts, if for anything, *against* the selected hypothesis rather than for it. If there are factors of plausibility counting in favour of the selected hypothesis, it is in the strategic sense of plausibility that this is so. So there is a kind of duality between propositional and strategic plausibility. In those cases in which rival hypotheses are eliminated for their implausibility, the factor of plausibility may be present in either sense. In the foregoing sections we have conjectured a conceptual link between what is plausible and what is characteristic. This is (we say) true for propositional plausibility, not for strategic plausibility. The question of probativity arises for both senses of what is plausible. There is, it seems, no direct link between a proposition's plausibility in either sense and its truth. But accepting what is propositionally plausible is intrinsically a reasonable thing to do; and accepting proposition for its strategic plausibility is also a defensible thing to do, since by definition the strategically plausible exhibits other virtues that bear on the successful completion of a cognitive task. In each case, however, there are additional questions to ponder. One is, what is it that makes it reasonable to accept propositional plausibilities? The other is, what is is about these other strategic virtues that makes it reasonable to accept strategic plausibilities? In each case, there is room for an *abductive* answer, to the affect that propositional-plausibility scores well with truth, as do the other instrumental virtues. A rival, and more circumspect, answer is that they do well on the score of empirical adequacy, albeit defeasibly so.

Proposition 7.5 (The quasi-duality of plausibility) *That a proposition is propositionally implausible is no bar to its being taken as strategically plausible. In other words, strategic plausibility can trump propositional implausibility.*

7.3.1 Historical Note on Plausibility

We have postulated that the factor of plausibility enters into an abducer's reasoning in ways that may be largely implicit. We have conjectured that a virtual logic is a description of a cognitive agent that might represent (certain of) his or its cognitive processes as behaviour of a connexivist net that meets the *RWR* conditions. We also imagined that one of the ways in which a cognitive agent implements plausibility and relevance considerations as might be captured by such a logic. (We also, by the way, now have a serviceable specification of a virtual rule. A virtual rule is a law, soft law or tight correlation that a cognitive system can conform to without executing.)

Whether this approach to the virtual logic of plausibility and relevance is correct, it remains the case that either way, we owe the reader an account of what we take plausibility and relevance to be. To this end, plausibility will occupy us in what remains of the present chapter, and relevance in the next. In the pages that follow, it is for the most part propositional plausibility with which we shall be concerned. Contrary uses will be noted in the text.

In common usage plausibility is equated with reasonableness. The equivalence originates in the concept of the reasonable (*to eulogon*). It comprehends Aristotle's notion of *endoxa*, or opinions held by all or by most or by the wise. These are opinions endorsed by, as we would say, common knowledge, or by a received view or by the experts. *To eulogon* is discussed by the Skeptic Carneades in the last century B.C., in the context of the evidence of the senses and the testimony of experts (See here [Stough, 1969]). A related notion is the Greek *eikos* which means "to be expected with some assurance", and it may be translated as "plausible-or-probable". Rescher claims (and we agree) that the one meaning of *eikos* is captured in the idea of approximate truth or verisimilitude, which "ultimately gave rise to the calculus of probability", though this was not to be a Greek development [Rescher, 1976b, p. 38, n. 1]. Aristotle contrasts *eikos* with *apithanon*, which means "far-fetched" (*Poetics* 1460^a27); he also distinguishes it from what is true. In criticizing his rivals, Aristotle says, "While they speak plausibly(*eikatos*) they do not speak what is true (*alēthē*)" (*Metaphysics*, 1010^a4).

Rescher opines that the Greek identification of *eikos* with the probable anticipates the Latin *probabilis*, which means "worthy of approbation", and he approvingly quotes Edmund Byrne

> [Probability] refers to the authority of those who accept the given opinion; and from this point of view 'probability' suggests approbation with regard to the proposition accepted and probity with regard to the authorities who accept it [Byrne, 1968, p.188]:

For a discussion of the emergence and development of the mathematical conception of probability see [Hacking, 1975; Daston, 1988; Franklin, 2001].

It is well to note a present-day use of "plausible" in the Bayesian analysis of prior probabilities. This may or may not be a usage foreshadowed by Peirce's understanding of the plausible.

> By plausibility, I mean the degree to which a theory ought to recommend itself to our belief independently of any kind of evidence other than our instinct urging us to regard it favorable [Peirce, 1931–1958, p. 8.223].

Rescher suggests that Peirce's notion "seems closer to the idea of an *à priori* probability" than to the idea of being worthy of approbation [Rescher, 1976b, p. 39, n. 1]. In this we disagree.

Here is a further remark of Peirce:

> As we advance further and further into science, the aid we can derive from the natural light of reason becomes, no doubt less, and less;

> but still science will cease to progress if ever we shall reach the point where there is no longer an infinite saving of expense in experimentation to be effected by care that our hypotheses are such as naturally recommend themselves to the mind ... [Peirce, 1931–1958, 7.220].

This passage derives from about 1901. Ten years earlier Peirce was making a similar point. He took it that evolution has rendered the axioms of Euclidean geometry "expressions of our inborn conception of space". Even so, "that affords not the slightest reason for supposing them to be exact"; indeed space may actually be non-Euclidean [Peirce, 1931–1958, 6.29]. What Peirce is offering is an elucidation of propositional plausibility. He does so by way of the linked ideas of the innate instinct for guessing. Caught in the intersection of those pair of factors are propositions which recommend themselves to us with the "insistence of an idea". What is thus recommended is that they be looked upon favourably, that is be selected for testing and held provisionally meanwhile. It is reasonable to yield to such recommendations because, says Peirce, of our innate flair for guessing right. "Right" here means having a generally good track record in the propositions we send out for testing. One might dispute details of Peirce's analysis, but we take it as demonstrated that Peirce's plausibility is propositional-plausibility. In Peirce's account, there is little evidence of strategic plausibility having the same explanatory force. Peirce makes it a necessary condition of our looking favourably upon H that H have explanatory force with respect to some surprising event. But for Peirce, explanatory force is wholly non-probative. More than that, in the absence of the insistence of an idea, H's explanatory force is abductively inert. The best that we can say about H's strategic plausibility is that if guessing H is guessing right, then H meets a necessary but abductively inert condition on this being so.

7.3.2 Cut-to-the Chase Abduction

In preceding pages we have tried to make the following case. Whenever an explanationist abduction problem is successfully solved, there is a filtration structure in which the winning hypothesis has a determinate place. Filtration structures are transformations of spaces of possibilities into subsets of the best candidates for abductive selection. The transformations are wrought by successive filters that remove from the space of possibilities those that are irrelevant to the case in question and that, in turn, remove the implausibilities from the space of relevant possibilities. If there is a unique H that solves the abduction problem, $\{H\}$ will be that unit subset of the space of plausibilities such that H is the most plausible. Although every winning H has a corresponding filtration structure, it is not the case that, in selecting H from the vast space of possibilities, an abducer actually constructs the corresponding filtration structure. This would be a good point at which to raise the question of what conditions would have to be met to make it *false* that for every winning H there is a corresponding filtration structure, irrespective of whether the abductive agent actually constructs it or not. The answer is obvious. The claim would be false if a winning proposition H were, even so, either irrelevant or implausible or both; or if H were both relevant and plausible, but that being so makes no contribution to a finding for H rather than some other possibility. At times Peirce writes as a *cut-to-the-chase* abducer, in which the abducer goes directly to H, unmediated by reflections on relevance and plausibility. That this is the right way to read Peirce is suggested by the emphasis he gives to the winning H as that which presses for favourable regard. But, if this is a good account of how abductions are actually made, it does not falsify the filtration-structure hypothesis. It might still be the case that the object H of a Peirceian cut-to-the-chase abduction is one that has a determinate place in a filtration-structure. So until we examine, in the chapter to follow, the factor of relevance, our provisional claim is that cut-to-the-chase abduction is compatible with the filtration-structure hypothesis.

7.4 Characteristicness

As we mentioned earlier, the generality in question is either a generic notion or what can be called a normalcy notion. Normalcy claims are claims about what is usually the case. Generic claims are stronger versions of normalcy claims. They are claims about what is always the case. "Always" is apt to mislead. Readers of this book will associate it with universal quantification. For the rest of humanity, a certain helpful looseness is indicated. When Harry observes that ocelots are four-legged he is not minded either to abandon or qualify the claim upon discovery of just any exception. Generic claims are generalizations that tolerate exceptions. Normalcy claims are weaker; so they too tolerate exceptions. But there are differences. An exception to a generic truth can without stretch be considered a negative instance of it. An exception to a normalcy claim is not a negative instance of it; it is already catered for by the qualifier "usually". In greater strictness, exceptions to normalcy claims aren't even exceptions. That Harry hates ice-cream leaves it wholly undisturbed that people usually like ice-cream. It is different with generic claims. It is true that ocelots have four legs. It is also true that Ozzie, the ocelot, has only three legs. This is a true negative instance, and is in that very sense an exception. Exceptions to generic claims are true negative instances of them, whereas exceptions to normalcy claims are in no sense negations of them.

The idea of what is characteristic cuts loosely across the grain of the distinction between genericity and normalcy. This, as we will see, complicates the story we tell about characteristicness, but not by much. In our discussion of previous cases, we entertained a loose connection between propositional plausibility and characteristicness. The analysis of *The Open Door* found it natural to take judgements of characteristicness as a species of generic reasoning. This in turn suggests a tie between plausibility and genericity. It would be well to examine these connections more closely.

Consider a candidate space $\mathbb{R}$ for an abduction target T. When discussing The Burglary, we proposed that propositions are excluded from $\mathbb{R}$ where they are contra-indicated by what is characteristic. Thus although "Sarah has come home early" was a member of Harry's $\mathbb{R}$, it was excluded by the fact that it is characteristic of Sarah not to come home early on weekdays. This was the basis on which we then said that the **Sarah** conjecture was implausible.

Generic claims are also claims about what is characteristic. So it is natural to think of characteristicness as intrinsically generic. A little reflection appears to call the claim into doubt. Penguins don't fly. This is a generic fact about penguins, and saying so attributes what is characteristic of them. What, then, do we say about Harry's pet penguin, Hortense? Of course, Hortense doesn't fly either, and that would seem to be characteristic of her insofar as she is a penguin. Sarah is unlike this. It is characteristic of her not to come home early on weekdays; and Sarah is a woman, mother, teacher, and lots of other things. But it is certainly not characteristic of women or mothers, or even teachers, not to be home at 4:30 on weekdays. Sarah does not imbibe this characteristic from what is characteristic of any class or natural kind of which she is a member. If genericity is intrinsically a certain kind of generality, then it might be thought that what is characteristic of Sarah in *The Open Door* case is not generic to Sarah. On the other hand, even in cases of particular characteristicness, the factor of generality is not lost entirely. If it is characteristic of Sarah not to be home early, then it is true to say that Sarah doesn't generally come home early on weekdays. But the generality imputed by the use of the adverb "generally" is not the generality of all Sarah's, so to speak, but rather is all weekdays between September to July. All weekdays at 4:30 are such that Sarah is not home from school then. Alternatively, it is characteristic of Sarah's late-afternoon weekday doings that they not include her being home at 4:30. But characteristicness does not strictly imply genericity. Sometimes what's usual suffices for what is or is not in character. So we need not impute genericity to Sarah's not being home at 4:30 in claiming characteristicness for it.

What is characteristic of Sarah is what Sarah *would* or *would not* do. Something that Sarah would do is often what Sarah has a standing desire to do or preference for doing. What Sarah is *like* is often just a matter of what she *likes*. Sarah likes to do her chores after school lets out at 3:30. Sarah prefers to shop each day rather than weekly. Sarah likes to drop into Macabee's Books for coffee and to browse. She doesn't like to home too early; it is less interesting to be home schmoozing with Harry than drinking the excellent coffee at Macabee's.

Dora's situation presents a different fix on characteristicness. Dora is never at Sarah's house on Thursday, whether at 4:30 in the afternoon or any other time. It is not just that Dora wouldn't like to be there then (although in her actual circumstances this is certainly true). The point rather is that Dora *can't* be there on Thursdays. She has other clients to whom she is committed on Thursdays. Dora is a good professional. She knows that clients much prefer to have an assigned stable time at which Dora appears in their households. Except where strictly necessary (and always with appropriate notice), Dora never changes these commitments (a further characteristic which turns on what Dora is like, and in turn, in this instance, on what she likes). But the basic fact is that Dora never comes to Sarah's house on Thursdays because she cannot. So externalities, as economists call them, can determine what is characteristic for an agent to do.

What is characteristic of Dora is not characteristic of various of the class of which she is a member. It may be characteristic of professional cleaners that they don't deviate from fixed commitments, week in and week out. But this flows not from what it is to be a cleaner, but rather from what it is to have a good feel for customer relations. So it would appear that what is characteristic of Dora is something generic or normalic to some of the things she is, but not others.

It is widely supposed that "Birds fly" is a generic claim. For this to be true, it would have to be characteristic of birds to fly. That is, flying would be tied up with what it is to be a bird. Of course, this is not so. Penguins are birds, and penguins don't fly. Penguins never fly. Similarly for turkeys, and ostriches, and lots of other birds. What this suggests is that sentences such as "Birds fly" do not express a generic proposition, but rather are under-expressed instances of the quantified sentence "Most species of bird fly". (We note, in passing that "Most birds fly", while true, is not what "Birds fly" asserts.) Thus "Birds fly" is a normalcy-claim. Generic sentences, on the other hand, are not quantificational. "Crows fly" is a generic claim and true. It suggests, but does not assert, that most crows fly; but there are cases in which "FsG" is a true generic claim even though most Fs do *not* G [Carlson and Pelletier, 1995]. The sense in which generic claims are generalizations cannot be even the weak sense in which they imply inequivalent quantifications.

One of the attractions of this approach to genericity and normalcy is the light it throws on the concept of default. Sometimes a default is understood as any claim which might in fact be mistaken. This is an unwisely liberal use of the term. More soberly considered, a default is the propositional content of an inference from a generic or normalic claim. If we infer from "Ocelots are four-legged" that Ozzie the ocelot is four-legged, that is a default. Saying so is saying more than that it might be untrue. It is saying (or implying) that it might be untrue in a certain way. If Ozzie is not four-legged, then given that its negation is a default, it is essential that although Ozzie is not four-legged it remains true that ocelots are four-legged. Those that aren't are so in ways that does not damage this truth. They are non-four-legged adventitiously (a leg-trap casualty, for example) or congenitally (a birth-defect).

This is not to say that propositions such as "Ocelots are four-legged" can't be false, that they can't have genuine counterexamples. We might come upon a heretofore unknown species of five-legged ocelots. This would falsify "Ocelots are four-legged", but not "Most species of ocelot are four-legged". It would no longer be characteristic of ocelots that they are four-legged.

We now have the resources to define defaults.

Definition 7.6 (Defaults) *S is a default iff there is a generic or normalic claim G of which S is an instance; and S is such that if G is true, S's falsity would not necessitate the falsity of G.*

The proposition that Sarah is not now home at 4:30 is a default. It is an instance of the generic claim that Sarah doesn't come home early on weekdays. Of course, upon entering the house, Harry might see that Sarah has come home early. This might be explained in either of two ways. Sarah might be sick. Or she might tell Harry that she's grown weary of getting home so late, and has decided to come home straight from school from now on. Each of these gives a different way for the default to be false. The first way, it remains true that it is characteristic of Sarah not to come home early. The reach of the generic (or normalic) claim is undamaged. The other way is different. Sarah's early home-coming instantiates a new policy that retires the old habit.

There is a third possibility. Sarah's early homecoming is inexplicable. "I don't know", she says, "I just felt like it".

Let us briefly regroup. Our point of departure is *The Open Door*. Our present task is to say something useful about the property of *characteristicness*.

We see that characteristicness is bound up with the concept of genericity, normalcy and default. So, instead of attributing (as we did in 6.1) defaultness to generic and normalic claims, we propose that its more fruitful application is to instantiations of such claims. Thus, the generic claim

(1) Ocelots are four-legged

is fallible (though not a default), and its instance

(2) Ozzie is four-legged

is both fallible and a default.

Why? What's the difference between the fallible non-defaultness of (1) and the fallible defaultness of (2)? The answer is that (2) derives its plausibility from (1), and is subject to downfall in the face of a single true counter-instance. But (1) is not withdrawn for just any reason for which (2) is withdrawn. (1) remains true even though (2) is false. (1) and (2) have different logics, as we might say. They are both fallible, but in different ways. One could be mistaken about the four-leggedness of ocelots, but not every way of being mistaken about the four-leggedness of Ozzie is a way of being mistaken about the four-leggedness of ocelots. We require that there be away of marking this distinction. There is such a way: (1) is generic, and (2) is a default.

How does this play upon the abductively important notion of characteristicness? It appears that the generic-default distinction partitions the characteristicness property in the following way. If (1) is true, then it is characteristic of ocelots to be four-legged; but it isn't characteristic of Ozzie to be four-legged. Another way of saying this is that the truth of (1) confers characteristic four-leggedness on ocelots, and yet characteristicness is not closed under instantiation.

This makes *The Open Door* problematic. We said that (2) is not a statement of characteristicness. Yet (2) is a singular statement. This might lead us to suppose that singular statements can't be statements of characteristicness. This is belied by

> (3) Sarah is characteristically not at home early on weekdays.

(3) is a singular statement, and a statement not obviously an instance of any generic statement. So, it would appear that it can't be true that even where a singular statement is a statement of characteristicness, it is not so because it is a default, as we are presently defining it.

But, on the contrary, (3) is a statement of characteristicness and *is* a default, and yet does not qualify as a default by instantiation of a generic statement that attributes the same characteristicness that (3) itself plainly attribute.

Of course, we tried earlier to find a generic or normalic statement of which (3) can be an instantiation. This required a certain regimentation (which, as Quine has taught us, is tendentiousness in a good cause). We proposed that (3) be re-issued as

> (5) It is characteristic of this day that Sarah not be home early

which could then be taken to as derived from

> (4) Weekdays are such that Sarah is not home early on them.

It won't work. Compare (1) and (2) with (4) and (5). In the case of the first pair, (1) does all the work on behalf of characteristicness. Even if we held our noses and allowed that (2) also attributes characteristicness, it would be clear that its functioning thus is wholly parasitic on the characteristicness attribution embedded in (1).

With (4) and (5) it is the other way around. With (1) and (2) it is consistent to say

> (6) Even though (2) is not a statement of characteristicness, (1) nevertheless is a statement of characteristicness

But we can't say

> (7) Even though (5) is not a statement of characteristicness. (4) nevertheless is a statement of characteristicness

From this we may infer that:

> (8) That (4) is a statement of characteristicness does not derive from any generic or normalic claim of which it is an instantiation.
>
> (9) Whether a statement is a default depends on its relation to attributions of characteristicness. In particular, (2) is a default because it instantiates a statement of characteristicness; and (4) is a default not because it instantiates a statement of characteristicness, but rather because it is itself a statement of characteristicness.

What now of plausibility? *The Open Door* example suggests that

> (10) Defeasibly, possibilities that contradict (statements of) characteristicness are not plausible abductive candidates.

We have postulated a connection between the plausible and the characteristic. Beyond observing that sometimes a judgement of implausibility is based on contrary characteristicness, we haven't done much to elucidate this connection and to plumb its degree of reach. Neither have we reflected on whether the tie, such as it is, is affected by the distinction between propositional and strategic plausibility. We briefly turn to these matters now, in reverse order.

Planck conjectured quanta for their contribution to the unification of the laws of black body radiation. He did so notwithstanding the extreme propositional implausibility of the existence of quanta. Even so, Planck thought it reasonable to proceed against the grain of this implausibility. Quanta were nothing *like* anything then known to physics; so they were uncharacteristic of what physics quantified over in 1900. Planck's was a conjecture grounded in its instrumental yield. It was, we say, an strategically plausible conjecture to make. Why would this be so? It would be so, as we saw earlier, because it is characteristic of the laws of physics to admit of unification under the appropriate conditions. (This is why the failure at present to unify relativity theory and quantum mechemes is for many physicists a scandal.) Planck reasoned that black body radiation is such that it should be expected that it is subject to unified laws, and because such unifications are characteristic of physics, he made a conjecture that would achieve it. This suggests that in cases such as this the tie with characteristicness holds for both propositional and strategic plausibility. However, for other cases the reverse would appear to be true. Peirce holds that a tie between or among rival hypotheses is often broken by economic considerations. The theorist opts for H over H' because it is testing

H is comparatively affordable and testing H' is not, never mind that H' might possess greater propositional plausibility than H. So we ask, is it characteristic of experimental testing that hypotheses pass the tests to which they are submitted? Of course, the answer is "No", as we saw earlier; and with it a further question arises: How does this comport with Peirce's claim that selected hypotheses are the outputs of our innate flair for guessing right? The question embeds an undeniable tension between Peirce's views, but it is a tension that is moderated by his fallibilism. Fallibilism allows that the frequency of erroneous guesses might outrun the frequency of correct guesses even while it remains the case that we have the flair for guessing right. It is allowable provided that guessing right is guessing right *enough of the time* and about *enough of the right things*, enough, that is, to secure our survival and our prosperity.

It might be supposed that in those cases in which instrumental plausibility retains its tie with characteristicness, the requisite instrumental factors themselves display another characteristic attachment. It might be supposed that factors such as unifiability, simplicity and coherence are markers for cognitive success, and that there is an abductive explanation of this, to wit: that such factors are defeasibly probative. This will appeal to philosophical and scientific realists, needless to say; but it is also the view of common sense to which all of us naturally lean. If true, care needs to be taken. The purported link between what is simple (etc.) and what is true cannot be so tight as to make the selections of hypotheses that turn upon these instrumental virtues run foul of the requirements of nescience. Even so, it can hardly be doubted that when judgements of strategic plausibility turn or factors such as the affordability of scientific testing, the link with characteristicness will be at its most tenuous.

Are we now in a position to characterize the link more fully? To a large extent, the answer will turn on what we are prepared to say of plausibility independently of the tie with characteristicness. (This task will occupy us in the sections to follow.) But this is not to say that there is nothing more that can be said at the present juncture.

7.5 Common Knowledge

In the broadest strokes, what a cognitive agent finds to be plausible is something he sees it as reasonable to hold or to consider holding in the absence of confirming evidence. Reasonable assent in the absence of evidence is the hallmark of abductive reasoning. Plausibility is an integral feature of abduction, and plausibility evinces this same feature — reasoning without evidence. We will suggest that the natural habitat of the plausible is *common knowledge* (cf. the discussion of "wise crowds" in chapter 2 above.) Whether or not common knowledge is knowledge strictly speaking, it has a rather striking feature. When an agent X assents to P

on grounds that it is common knowledge that P, it cannot be the case that, then and there, P is knowledge *for* X. This contrasts with the case in which an agent X^* assents to P, P is a matter of common knowledge, but X's assent to P is independent of that fact. A further feature of the first case — i.e., of inferring P because and only because it is common knowledge — is that rarely does the agent in question have strong grounds even for the claim of common knowledge. Common knowledge loops recursively. Often something is taken by X as common knowledge when X takes it as common knowledge that P itself is a matter of common knowledge. It is something of a relief therefore that rarely when X draws upon common knowledge does he expressly say so. Saying so is usually reserved as an answer to a challenge; and in that context it is often not much of an answer. It is preferable that we view the contributions of common sense from the third person point of view, the view from which *we* try to discern what the *other* party is doing when it seems clear to us that he is exploiting what he takes to be common knowledge. In taking that view, recurring features come to light. What counts as common knowledge for X is a set CK of beliefs, or propositions to which he is disposed to assent, which he believes are widely shared — in effect that they satisfy the Aristotelian conditions on *endoxa*: they are beliefs held by all or by the many or by the wise. These propositions divide fairly naturally into two main groups. (1) They are instantiations of generic or normalic claims in CK. (2) They are propositions, including those generic or normalic claims, that come to X by hearsay.

In both cases, the factor of characteristicness is robustly present. The link between the general and the characteristic we have already discussed. The link between hearsay and characteristicness has yet to be made. Hearsay is like guessing. With guessing, we wanted to know how it could be that guessing is a cognitively virtuous endeavour, when it might be the case that incorrect guesses out number the correct. The answer was that it is characteristic of guessing that it serves us well even when it frequently goes wrong, instance by instance. Another way of saying this is that "Guessing is reliable" is a true generic or normalic claim even though the quantified sentence "Most guesses are right" might be false. This gives us some idea of the truth conditions for generic sentences of this particular type. "Guessing is reliable" is true if it is true that guessing is often enough right enough about the right things. Hearsay has something of this same character. Hearsay gets many things wrong, but it gets the right things right sufficiently often to have become a wholly entrenched and indispensable instrument of our largely successful negotiations with a hostile world. With embeddedness goes characteristicness. We may take it, then, that the elements of characteristicness are embedded in common knowledge. If, as we suggest, common knowledge is the principal source of what an agent takes to be plausible, the tie with characteristicness is reconfirmed. Apart from this, common knowledge passes the other tests for plausibility. The instantiation of a generic or normalic claim is a default, as is a proposition accepted on hearsay (given the generic or normalic truth that hearsay is reliable). In each case, the default is allowable on sufferance. In neither case are we talking about evidence.

Before bringing this section to a close, this would be a good place to return to the point that cases such as *The Open Door* invoke implausibility as a reason to exclude a candidate hypotheses. It is not intended that survival of this exclusion test confers plausibility on the hypotheses left standing. "P is not implausible" does not imply "P is plausible". This suggests a two-tiered structure for plausibility filters. In the first tier, the reasoner excludes candidate hypotheses on grounds of implausibility. In tier two (mindful of the possibility of countervailing considerations), the reasoner seeks the most plausible of the survivors and excludes the rest.

Assuming a case in which the reasoner finds that he is able to discern the most plausible of a not implausible lot, what would count as "countervailing considerations?" One is that the most plausible is not plausible enough to warrant its selection. Another is that the most plausible candidate makes a strong claim on abductive success, but, for instrumental reasons (including cost), this is not the hypothesis that is sent to trial. We take such cases as showing that there is a principled distinction, which Peirce sometimes blurs, between drawing a conclusion in the form $C(H)$, and reaching a decision to send H to trial. This is very much as it should be: "It is justified to conjecture that H" is not equivalent to "It is justified to send H to trial". It is possible to take a justified decision to send H to trial without it being justified to conjecture that H before the fact. It is possible that what justifies sending H to trial is that it has passed through the two-tiered plausibility filter. But not everything that is the most plausible of a bunch of not-implausibles is something whose conjecture would be justified. Accordingly,

Proposition 7.7 (Plausibility and trial decisions) *If an agent X is faced with an abduction problem in regard to T, and if (1) $R^{pres}(K(H), T)$ and (2) H is most plausible in a set of not implausible alternatives, then if, on the strength of (1) and (2), the reasoner solves his problem with a decision to send H to trial, he has* not *performed an abduction with regard to H.*

As we see, not every solution to an abduction problem is abductive. Abductive solutions pivot centrally on the factor of conjecture and discharge. Sending a most plausible H to trial need not involve the judgement that it is justified to conjecture that H. Neither need it embed a decision to release it for premissory work in future inference. The conclusion of an *abduction*, therefore, cannot be that H would be a good bet for testing. It is necessary to keep these distinctions in mind as we move to the next section on plausibility logics.

Considerations of plausibility lie at the heart of much of practical reasoning. It is also a dominant factor in theoretical reasoning. Its utter prominence is matched in inverse proportionality by the comparative meagerness of the research programme in the logic of plausibility. A notable exception is Rescher, who has some interesting suggestions to make, albeit at the level of propositional logic.

7.6 Rescher's Plausibility Logic

According to Nicholas Rescher (with whom we agree) a fundamental fact about plausibility is:

Proposition 7.8 (Plausibility negation) *The negation of a plausible proposition is not necessarily implausible.*

Corollary 7.8(a) *The negation of a plausible proposition might be as plausible as it.*

These things being so, we must modify an earlier suggestion, according to which, for large classes of cases, abduction is the engagement of the most plausible hypothesis.

Proposition 7.9 (Plausibility adjudication) *It is sometimes the case that the adjudication of rival plausibilities cannot be achieved by picking the most plausible of them.*

Proposition 7.9 provides for the following kind of case. Let H and H^* star be equally ad highly plausible hypotheses, and that H^* implies $\neg H$. It may appear that, since H and H^* contradict one another, the information that each has the highest plausibility can give the adjudicator no relevant guidance. Doubtless such will sometimes be the case. But we should not turn a blind eye to the possibility that they might *both* be conjectured and *both* be discharged. Provided that an appropriate degree of caution is practised (e.g., we want to avoid claims in the form, $C(H \cap H^*)$), except for dialetheic reasons. Each might be released, and each on sufferance, for premissory work in future inferences, except, again, for dialetheic contexts. Either way, the rivalry between H and H^* is unresolved. In the first kind of case, their equal plausibility proves adjudicatively paralyzing. In the second case, the discharge decisions ignore the rivalry.

There are further respects (as we shall shortly see) in which plausibility logics differ from standard probability logics. Since both are theories of ampliative inference, it is not unnatural to assume that they are rivals. But this is not the spirit in which Rescher develops his account of plausibility. Its motivation is considerably more circumscribed. Classical logic tells us that inconsistency is a disaster. Standard probability can no more handle inconsistency than a number can be divided by zero. Rescher seeks a principled answer to the question of how best to select consistent subsets of beliefs from inconsistent sets. Accordingly, Rescher wants a paraconsistent theory of belief-revision. Belief revision is also an essential feature of abduction. Twice-over, an account of the Rescher kind is grist for the mill of abductive logic. It says something about plausibility, and it does so in the context of saying something about belief revision.

Rescher sees plausibility as essentially tied to authoritative sources, of which expertise is an important instance. One of the more dramatic things about experts is that they disagree. How might these disagreements be resolved?

Rescher assumes that experts on a given subject form a poset under the partial order "has greater expertise than". Plausibility in turn can be construed in ways to be detailed below.

Suppose that an authority or group of authorities has made a number of pronouncements, and that the set of propositions vouched for by these authorities is collectively inconsistent. Suppose further that we can rate the plausibility or reliability of these authorities in some comparative way. Is there a rational way to deal with such a situation? One such procedure, developed by [Rescher, 1976b], is called *plausibility screening.* Rescher's method is to scan the maximally consistent subsets of the inconsistent totality and give preference to those that include the maximum number of highly plausible elements. This general process can also function in other ways. For example, for some purposes we might want to give preference to those sets that include as few low-plausiblity elements as possible. It depends on whether our goal is to maximize overall plausibility or minimize overall implausibility. The two policies are not identical.

Suppose that we are given a fragment of what appears to be a thirteenth-century manuscript on logic. It has been examined by three experts on historical manuscripts on the logic of this period, Professors, X, Y, and Z. Let us say that we can rate their respective pronouncements on a scale of one to ten as follows: X has a comparative reliability of 8, Y has a rating of 5 and Z a rating of 2. (Even though Z's rating is low, it must be emphasized that Z is a bona fide expert, and that is low rating is a *comparative* matter.) Professor Y ventures the opinion that the manuscript was authored by William of Sherwood, the thirteenth-century Oxford logician, or by William of Ockam, his near contemporary. Professor X asserts that if the document were authored by William of Sherwood then it would definitely make reference to Aristotle's doctrines on logic. But, he adds, no such reference to Aristotle is made. Professor Z points out that if the document was authored by William of Ockam, then from what we know of William of Ockam, it would include references to Aristotle's doctrines.

We are supposing, then, that these authorities vouch for the following propositions:

Authority X (who has a reliability value of 8): $A \supset B, \neg B$
Authority Y (who has a reliability value of 5): $A \wedge C$
Authority Z(who has a reliability value of 2): $C \supset B$,

here

$A =$ The manuscript was authored by William of Sherwood.
$B =$ The manuscript makes reference to Aristotle's doctrines on logic.
$C =$ the manuscript was authored by William of Ockam.

We now put it that the reliability rating of a *proposition* is the same as the reliability rating of the expert who vouches for it. The sets of given propositions, $\{A \supset B, \neg B, A \vee C, C \supset B\}$, is inconsistent, as a truth table will show, but it has four maximally consistent subsets as follows:

1. $\{A \vee C, A \supset B, C \supset B\}$, rejecting $\neg B$
2. $\{A \vee C, A \supset B, \neg B\}$, rejecting $C \supset B$
3. $\{A \vee C, C \supset B, \neg B\}$, rejecting $A \supset B$
4. $\{A \supset B, C \supset B, \neg B\}$, rejecting $A \vee C$.

Notice that (1) and (3) both reject one of the highly rated pronouncements of X. Therefore, given that we want to maximize plausibility, we can eliminate both (1) and (3) as candidate subsets. Looking at the remaining two subsets, we see that we have a choice between rejecting $C \supset B$ (which has reliability (2)) and $A \vee C$ (which has reliability (5)). Again, since our policy is to maximize plausibility, we will want to reject any alternative that excludes propositions of relatively high reliability. So the choice here is straightforward as well. We reject (4) because it excludes $A \vee C$, a proposition that is more reliable than $C \supset B$, the proposition excluded by (2). All told, then, the most plausible maximally consistent subset is (2). Thus, on this model, the rational way to react to inconsistency in this instance is to accept the pronouncements of X and Y and reject the opinion of Z. Note that the plausibility of (2) suggests that the manuscript was indeed authored by William of Ockam, since (2) logically implies C.

In our example, we selected the most plausible subset by pruning the original, inconsistent set of data. We identified the most plausible maximal consistent subset of S as that which excludes propositions of least reliability. The method of plausibility screening tells us even more, however. If we look at (2) and (4), the preferred subsets of our example, we see that the two propositions $A \supset B$ and $\neg B$ each appear in *both* (2) and (4). In other words, no matter which of (2) or (4) we decide to accept, we are going to accept $A \supset B$ and $\neg B$. These two propositions therefore constitute something akin to a "common denominator." We also see that in this case the pronouncements of X have a certain preferred status: no matter whether we reject the opinions of Z by rejecting (2), or reject the opinions of X. Plausibility screening also tells us that (2) is not consistent with (4), as a truth table will show, and thus that we must choose between (2) and (4). As we say, in this case it is preferable to select (2); but in choosing either (2) or (4) we will still be accepting the common subset $\{A \supset B, \neg B\}$.

The method of plausibility screening consists of several steps: First, we take the original set of pronouncements of our authorities and test them for consistency by constructing a truth table. If the set is inconsistent, we then determine all the maximally consistent subsets of this set. We then look over the alternatives and reject those sets that tend to exclude the most reliable propositions until only one set is left. In the event of a tie, we look to see if there is a common component among the tied sets. As well, we can try to discover whether there is a common component among any number of the maximally consistent subsets that tend to be preferable.

We could bend our example to form an illustration of a tie: suppose that Z is assigned a reliability value of 5. The (2) and (4) would be tied. Whichever set we rejected, we would be rejecting a proposition of value 5. Both (2) and (4) are preferable to (1) and (3), but we cannot narrow the field down to one proposition. In this context, the best we can say is that we will want to accept $\{A \supset B, \neg B\}$ because it is common to both (2) and (4).

Here is a second example, which we develop in a little more detail. Suppose we have the following pronouncements of three authorities, X, Y, and Z:

Authority X (who has a reliability value of 9): $A \supset B, \neg C$
Authority Y (who has a reliability value of 7): $B \supset C, \neg A$
Authority Z (who has a reliability value of 2): $A \vee B, \neg(A \wedge B)$.

First, we construct a truth table with a column representing the truth values for each proposition in the set of propositions stated by the experts:

	A	B	C	$A \supset B$	$\neg C$	$B \supset C$	$\neg A$	$A \vee B$	$\neg(A \vee B)$
(1)	T	T	T	T	F	T	F	T	F
(2)	T	T	F	T	T	F	F	T	F
(3)	T	F	T	F	F	T	F	T	T
(4)	T	F	F	F	T	T	F	T	T
(5)	F	T	T	T	F	T	T	T	T
(6)	F	T	F	T	T	F	T	T	T
(7)	F	F	T	T	F	T	T	F	T
(8)	F	F	F	T	T	T	T	F	T

Second, we can scan the truth table and highlight all combinations of true propositions, omitting only those that are proper subsets of others. Some rows, like (1), (2), (3), and (7), will have true propositions in them, but these patterns of true propositions will already be included in one or more of the other rows. For example, the true propositions in (1) form a subset of those in (5); the true propositions in (2) form a subset of those in (6); the true propositions in (3) form a subset of those in (4); and the true propositions in (7) form a subset of those in (8):

	A	B	C	$A \supset B$	$\neg C$	$B \supset C$	$\neg A$	$A \vee B$	$\neg(A \vee B)$
(1)	T	T	T	T	F	T	F	T	F
(2)	T	T	F	T	T	F	F	T	F
(3)	T	F	T	F	F	T	F	T	T
(4)	T	F	F	F	**T**	**T**	F	**T**	**T**
(5)	F	T	T	**T**	F	**T**	**T**	**T**	**T**
(6)	F	T	F	**T**	**T**	F	**T**	**T**	**T**
(7)	F	F	T	T	F	T	T	F	T
(8)	F	F	F	**T**	**T**	**T**	**T**	F	**T**

Third, we look at the truth table, one row at a time, and list those propositions that are true in each highlighted row. For example, in (8) we see that the following propositions are true: $A \supset B$, $\neg C$, $B \supset C$, $\neg A$, and $\neg(A \wedge B)$. Reading off the true propositions for the remaining rows that contain highlighted propositions, we find that we have a total of four maximally consistent subsets:

(4) $\{\neg C, B \supset C, A \vee B, \neg(A \wedge B)\}$
(5) $\{A \supset B, B \supset C, \neg A, A \vee B, \neg(A \wedge B)\}$
(6) $\{A \supset B, \neg C, \neg A, A \vee B, \neg(A \wedge B)\}$
(8) $\{A \supset B, \neg C, B \supset C, \neg A, \neg(A \wedge B)\}$

Fourth, for each such maximally consistent subset, we list at the right those propositions of the original set that are not included.

(4) $\{\neg C, B \supset C, A \vee B, \neg(A \wedge B)\}$, rejecting $A \supset B, \neg A$
(5) $\{A \supset B, B \supset C, \neg A, A \vee B, \neg(A \wedge B)\}$, rejecting $\neg C$
(6) $\{A \supset B, \neg C, \neg A, A \vee B, \neg(A \wedge B)\}$, rejecting $B \supset C$
(8) $\{A \supset B, \neg C, B \supset C, \neg A, \neg(A \wedge B)\}$, rejecting $A \vee B$.

Fifth, we scan the maximally consistent subsets in order to construct a preference ordering. The general rule here is that any set that rejects a highly reliable proposition should be eliminated. Clearly we can eliminate (4) because it rejects $A \supset B$, which has a value of 9. Likewise (5) must be eliminated because it rejects $\neg C$, which also has value 9. That leaves (6) and (8). Here the choice is also clear. Row (6) rejects $B \supset C$ (which has a value of 7), whereas (8) rejects only $A \vee B$ (which has a value of 2). On the policy that plausibility is to be maximized, we therefore eliminate (6) and accept (8).

Sixth, in the event of a tie, we look to see if there is a "common denominator" subset that should be accepted even if it is necessary to reject all of the maximally consistent subsets. Looking over the maximally consistent subsets, we see that the subset $\{A \supset B, \neg C, \neg A, \neg(A \wedge B)\}$ is common to both (6) and (8). Furthermore, we see that $\{\neg(A \wedge B)\}$ is common to all four maximally consistent subsets. These common components could serve as tiebreakers, although in this case this is not necessary — (8) stands out as the clear winner. Nonetheless, it is interesting to note that despite its low individual plausibility, $\neg(A \wedge B)$ is highly acceptable because it is "carried along" in (8) and in every other maximally consistent subset.

Finally, there may exist undecidable cases. If authority X says P and authority Y says $\neg P$, and if both authorities are equally reliable and this is all the information we are given, then plausibility screening will not tell us which proposition to accept. It is simply a stalemate, and we have to wait for more information before making our decision about whether to accept or reject P.

7.6.1 Reliability

Perhaps the first thing to notice about this approach to the adjudication of rival plausibilities is that the reliability values used in our examples are not intrinsic to the account. As formulated, the logic of plausibility screening may appear to allow for cases in which the reliability values of the experts are all so low as to making the winning maximal consistent set one of very low plausibility, intuitively speaking. But this is a misconception. By the construction of the screen, all the participants are genuine experts and each of their pronouncements is plausible to a degree that could warrant acceptance, however provisionally.

Critics of Bayesianism will no doubt be quick to see a similar weakness in the theory of plausibility screening. Just as there appears to be no general and direct way to assign prior probabilities, the same would appear to be the case for prior plausibilities, or what Rescher calls plausibility-indices (see the section to follow). It is true, of course, that in real-life we make these assignments when we are confident that they give approximate expression to our intuitive judgements or our commonsense estimates of which is the greater expert. But, it remains the case that there is no general method for plausibility indexing even for classes of human claimants to expertise. Beyond that, it matters that what we take as plausible ranges far beyond the reach of any human expert — a point to which we shall briefly return.

It might be said that the greatest difficulty with setting prior plausibilities lies in the claim that, in the absence of a general method, we do this intuitively or under the guidance of commonsense. What can "intuitively" mean here, if not "in ways that we find plausible"? And what can the guidance of commonsense be here, if not the counsel afforded by our judgements of characteristicness? Either way, we import the notion of plausibility into our account of how to set prior plausibilities. The ensuing circularity requires that we cannot embed these plain truths in our account of plausibility. This is significant. It leaves plausibility theory impotent to lay down principled conditions on plausibility indexing.

On the positive side, the theory of plausibility adjudicates conflicting expert testimony eucumenically and realistically. Instead of an all-or-nothing favouritism for the total testimony of the most qualified expert, the theory seeks to accept the biggest part of the combined testimony of all the experts that it consistently can at the lowest cost (i.e., rejecting claims of least plausibility). Even so, while the *idea* of saving as much of the joint testimony of all the experts that consistency and plausibility-cost will allow may be a good one, Rescher's specification of how this is achieved is problematic.

Consider two cases. Experts X and Y have been sworn in a criminal trial. Each gives psychiatric testimony concerning the question of diminished responsibility. X has an expert ranking of 8, and Y has a ranking of 7. In the first case, X and Y agree on everything each other says save for one proposition (e.g., that the defendant had at most diminished responsibility for his action). In that case, the nod goes to X's testimony. In a second case, everything is as before except that X and Y disagree on everything each other testifies to. Here the nod also goes to X. More generally, in all such cases, the nod goes to the most highly ranked expert except when one of them introduces testimony which is neither repeated nor denied by the other. Doubtless there are lots of real-life cases in which this is so. But as our examples show there are other cases in which the method of amalgamation is just the all-or-nothing strategy in favour the most highly ranked expert's total evidence. It is also necessary to note that in

Proposition 7.10 (Reliability) *In Rescher's model, plausibility is parasitic on reliability, which in turn is a matter of the authority of sources.*

7.6.2 Axioms for Plausibility

A central idea in Rescher's approach is *plausibility indexing*. The integer 1 denotes total reliability and gives effective certainty. Rescher rates disciplines such as logic and mathematics as having reliability-value 1. In all other cases, 1-n/n, n-1/n, n-2/n, ..., 1/n denote diminishing degrees of positive reliability. Thus even the least reliable of sources is reliable to *some* extent. 1/n denotes minimal positive reliability.

Corresponding to a reliability index for sources is that of a plausibility index for sets of propositions. Such sets Rescher calls *p-sets*; they are sets of propositions endorsed by sources of positive reliability. Indexing of p-sets **S** is subject to four axioms.

By Axiom 1 (*metrization*), every proposition in **S** has a logical value k ($0 \leq k \leq 1$). Axiom 2 (L-truth maximization) provides that every truth has plausibility value 1. By axiom 3 (compatibility), all propositions of plausibility 1 are mutually co-tenable. Axiom 5 (*consequence*) provides that for any consistent sets Σ of proposition in **S** and any proposition in **S** and any proposition in **S** entailed by it, the entailed proposition cannot have a lower plausibility value than any proposition in Σ. Axiom 5 permits that a proposition and its negation can have plausibility values as high as they come, short of 1. By Axiom 6, differentially plausible rival propositions are to be adjusted in favour of the higher (highest) plausibility value.

Rescher's axiomatization of an elementary propositional language for plausibility is attractive in a number of respects. But it also exemplifies the distorting tug of a theorist's formalism. All formal models occasion distortions. When the model is good, the distortion is tolerable, or even better; for it might correct a prior belief naively held. In the present case, we find some of the distortions to be regrettable. By the doctrine of plausibility indexing, Rescher is able to say that the truths of logic and mathematics have the highest plausibility. This is contradicted by the long history of logic and mathematics; in which one of the central roles of proof is to demonstrate the truth of the (even wildest) implausibilities. The concurrence in Rescher's account with the most true and the most plausible is occasioned in part by the need to engage the machinery of a smooth formalism. It is also explained in part that Rescher's main purpose is not to produce a conceptually wholesome account of plausibility, but rather to construct an elementarily useful model of belief revision in inconsistent contexts. But unless Rescher's plausibility bears some affinity to the real thing, the model of belief revision will do its work blindly. So we think it necessary to point out the actual non-concurrence of the most plausible with the most true. Accordingly,

Proposition 7.11 (L-truth maximization) *Rescher's Axiom 2 misstates the relationship between the most true and the most plausible.*[3]

Again we see in this the conflation of plausibility and reliability. Suppose that it could be said that logic and mathematics are the most reliable disciplines. Then, perhaps it wouldn't be too much of a stretch also to say that propositions determined to be true by the methods of logic or mathematics are, in this same sense, the most reliable of propositions. If we did agree to say these things, they would not give us the slightest reason also to say that the propositions of logic and mathematics are the most plausible propositions. Accordingly, a metrization that ranked disciplines (and their claims) according to their reliability could not in general be

[3] Similar reservations extend to Hailpern's account of plausibility measures[Hailpern, 2003, pp. 50–55: esp. 51].

supposed to constitute an acceptable ranking of their plausibility. This bears on Rescher's entire enterprise.

Proposition 7.12 (The plausible and the reliable) *Rescher's formalization of plausibility exhibits a systematic* confusion *between the reliable and the plausible.*

Corollary 7.12(a) *Rescher's theory of plausibility screening is better understood as a theory of reliability screening.*

Corollary 7.12(b) *The transformation of Rescher's accounts of plausibility into accounts of reliability is an attractive contribution to epistemology. Bearing in mind that abduction is reasoning that is epistemically subpar in essential ways, the factor of plausibility at the heart of abduction is not well-captured by an account of reliability.*

We close this subsection with a brief word about Rescherian plausibility-consequences. Axiom 5 honours the basic structure of standard consequence relations in logic: In a true statement of consequence, the antecedent cannot have a higher value than the consequent. It might be natural to think that plausibility is closed in this way under consequence. But it isn't. Consider a case. Harry is under investigation for the murder of Lou. At present, the case against him is inconclusive. A proof of motive is not required for a criminal conviction, but often police and prosecutors seek such a proof for the reasurrance it appears to give to juries. The investigating officer ruminates as follows.

1. It is quite plausible that Harry bore Lou ill-will.

2. It is fairly plausible that Harry would operationalize his ill-will in some way.

3. So there is some slight plausibility that Lou's death came about for these reasons.

Not only is the stronger plausibility of the premisses not preserved in the (wholly reasonable) inference, but the reasoner himself marks this fact by the placement of qualifications on the imputed plausibilities: quite plausible in (1), fairly plausible in (2), but only slightly plausible in the conclusion.

Proposition 7.13 (Plausibility and consequence) *Rescher's Axiom 5 fails for plausibility.*

Corollary7.13(a) *Axiom 5 is more reasonably construed as a reliability axiom.*

A number of theorems are provable from the Rescher axioms. Here are some of the more important ones. Proofs are easily reconstructed and are here omitted.

Theorem 1 If Q follows from P, then P's plausibility value cannot be greater than Q's.

Theorem 2 Interdeducible propositions have the same plausibility values.

Theorem 3 For any P, Q and $\ulcorner P \wedge Q \urcorner \epsilon$ **S**, the plausibility value of $\ulcorner P \wedge Q \urcorner$ = the lesser of those of P and Q. In symbols $\mid P \wedge Q \mid = min[\mid P \mid, \mid Q \mid]$

Theorem 4 For any P, Q, and $\ulcorner P \vee Q \urcorner \epsilon$ **S**, $\max[\mid P \mid, \mid Q \mid] \leq \mid P \vee Q \mid$

Theorem 5 For any P, Q, $\ulcorner P \wedge Q \urcorner$, $\ulcorner P \vee Q \urcorner \epsilon$ **S**, if $\mid P \mid = \mid Q \mid$, then $\mid P \wedge Q \mid = \mid P \mid = \mid Q \mid \leq \mid P \vee Q \mid$

Theorem 6 For all a, $\mid \forall x(Fx) \mid \leq \mid Fa \mid$; and $\mid Fa \mid \leq \mid \exists x(Fx) \mid$.

P-sets are not subject to deductive closure. However, when a p-set is consistent it is possible to extend the plausibility index of **S** to cover **S**'s deductive closure. Let P be a consequence of **S** (which may or may not occur in **S**). Then the sequences

$$\begin{array}{l} P_1^1, P_2^1, \ldots, P_{n_1}^1 \\ P_1^2, P_2^2, \ldots, P_{n_2}^2 \\ P_{1^i}, P_{2^i}, \ldots, P_{n_i}^i \end{array}$$

are the propositions in **S** that imply P in the various rows of a plausibility matrix. Then to determine the plausibility index of the deductive closure of P, select the maximum of the minima

$$min_j P_j^i$$

for each such sequence. Thus

$$\mid P \mid = max_i min_j P_j^i$$

The inconsistency of **S** is another matter. In the case of an inconsistent p-set **S**, "the automatic addition of logical consequences must be avoided." As we have said, plausibility screening is Rescher's most general method for handling inconsistent p-sets, although other possibilities are also discussed [Rescher, 1976b, ch. 5 and 6]. Since our main interest in this chapter is the explication of plausibility, we shall not expound these alternative possibilities.

Central to Rescher's conception, as with those of some of the ancients, is that the reasonable is not to be equated with the probable in its Bayesian sense. The difference between plausibility and probability is attested to in a number of ways, some of the more significant of which are their respective treatments of the logical operators. The calculus of plausibility (as Rescher calls it) is not closed under negation. If $\mid P \mid$ is known, this is not generally sufficient for the determination of $| \neg P |$. But the probability of $\ulcorner \neg P \urcorner = 1-$ the probability of P. The negation operator always raises or lowers the probability of the proposition negated; and probability always degrades conjunctive probability. Thus, where P and Q are independent, $Pr(P \wedge Q) = Pr(P) \times Pr(Q)$. But negation does not in general raise or depress plausibility nor does conjunction degrade plausibility. Nor is $Pr(P|Q)$ definable when Q is inconsistent, whereas inconsistent sets are perfectly open to plausibility indexing.

7.7 Plausibility and Presumption

In earlier sections, we considered the link between the plausible and the characteristic. In the present section we shall briefly examine a proposed link between plausibility and presumption. As it is developed in Walton [1992a], plausibility is subject to two restrictions, both of which we regard as inessential. One is that plausibility is a property instantiated only (or paradigmatically) in dialogues. The other is that such dialogues involve only (or paradigmatically) everyday matters. If plausible reasoning were intrinsically dialogical and instrinsically quotidian, then Walton's account would be of little value in the explication of abduction, which is neither. But as we say, we see nothing in Walton [1992b] that shows these constraints to be essential; consequently we decline to impose them. However, we agree that, just as abduction embodies a certain kind of enquiry, presumption can also usefully be relativized to enquiry. Thus whether something is reasonably presumed or not will in general be influenced by the kind of enquiry in process, and by the cognitive targets it embeds.

Walton's central idea is that plausible reasoning "needs to be understood as being based on and intimately connected to presumptive reasoning" [Walton, 1992b, p. 4] Presumptive reasoning, in turn, is said to contrast with both deductive and inductive reasoning. "Presumptive reasoning", says Walton, "is inherently subject to exceptions, and is based on a defeasible general premise of the form 'Typically (subject to exceptions), we can expect that if something has property F, it also has property G'" [Walton, 1992b, p. 4]. If it appears unclear as to why presumptive reasoning can be neither deductive nor inductive, Walton says that "deductive reasoning is based on a universal general premise of the form 'All F are G', and inductive reasoning is based on a probabilistic or statistical general premiss to the effect that most, or a certain percent [sic] of things that have properly F also have property G" [Walton, 1992b, p. 4]. We won't take the time to subject these caricatures to the derision that they so richly deserve. Let it suffice that, properly understood, the contrast class for presumption is neither deduction nor induction, and that, even as characterized by Walton, presumptive reasoning cuts across the grain of these distinctions. What makes this so is the role of presumptive *premisses* in arguments of all kinds. What Walton misses is that presumptive inference is two different things: (1) inference *to* a presumption, and (2) inference *from* a presumption.

Walton distinguishes among three kinds of presumptions — *required, reasonable* and *permissible.*

> A proposition A is a *required presumption in an inquiry* if and only if, (i) whether or not A is true is open to further argument or inquiry, and (ii) A may be inferred from what is presently known in the inquiry, and (iii) A must be inferred from what is presently known in every inquiry of this type, in the absence of special circumstances [Walton, 1992b, p. 53].

A proposition A is a reasonable presumption in an inquiry if and only if the first two conditions are met and "A may reasonably be inferred from what is usually expected in relation to what normally can be expected to be found in this type of enquiry, in the absence of exceptional circumstances" [Walton, 1992b, p. 53]. A is a permissible presumption in an inquiry when A "may be inferred from what is known, but does not have to be" [Walton, 1992b, p. 53]. We note in passing two unattractive features of these definitions. One is their imprecision. There is, for example, no discussion of the difference between "may be inferred" (condition (ii)) and "may reasonably be inferred ..." (condition (iii) on reasonable presumption). Also unexplained is the meaning of "must be inferred ..." (condition (iii) on required presumption.) Apart from this inexactitude there is also at least a hint of circularity. Thus something is a reasonable presumption when it is reasonable to infer it.

Perhaps the greatest difficulty with this account is making sense of the tie between presumptive reasoning and plausible reasoning. Though the connection is asserted, it is not explicated.[4] Even so, it may be possible to reconstruct the tie on the basis of what Walton says about presumption. Two points are particularly important. One is that the inquirer be undecided as to the truth or falsity of any proposition that he holds presumptively. The other is that the proposition "may be inferred" from what is known of situation in which the presumption arises. Jointly, presumptions are legitimate to hold even when they chance to be untrue. This legitimacy is off-set by the tentativeness with which a presumption is held or forwarded. This goes some way toward capturing Rescher's idea that "plausibility is a weaker counterpart to truth" [Rescher, 1995, p. 688]. Nowhere, however, does Walton speculate on the conditions under which our present state of knowledge makes it reasonable to hold defeasibly a proposition that might be false. This leaves his account of plausibility significantly underdescribed. A proposition is

[4] The great oddity of [Walton, 1992a] is that, notwithstanding its title, *Plausible Argument in Everyday Conversation*, it is a book with almost nothing to offer in the way of a positive account of plausibility.

plausible when it is an epistemically respectable possible falsehood. So is a probability statement in the unit interval [0, 1]; but Walton insists (rightly) that statements of probability are not statements of plausibility.

Plausibility, says Rescher, turns on a claim's credibility via the acceptance justifying backing that a duly weighted source (human, instrumental, or methodological) can provide. Thus if we think of informative sources as being graded by reliability, then the plausibility of a contention is determined by the best authority that speaks for it. A proposition's plausibility accordingly depends on its probative status rather than on its specific content in relation to alternatives [Rescher, 1995, p. 688].

Walton, too, acknowledges a role for expertise in presumptive reasoning (see, e.g., [Walton, 1992b, pp. 259–262]). The difference is that Rescher makes the factor of authoritative reliability intrinsic to plausibility; whereas for Walton (rightly, in our view) it is a contingent matter, only occasionally sufficient for plausibility.[5]

Plausibility, we said, involves something characteristic embedded in the phenomena and the possibilities under the abducer's consideration. What is characteristic of Sarah is that she is never home early on weekdays. Of course, on any given day she might in fact have come home early, but that is a possibility wholly compatable with what is characteristic of her. Harry presumes until he knows better that Sarah has yet to come home. And what he presumes on this occasion is plausible because the generic or normalic proposition that Sarah doesn't come home early on weekdays is true and because it licenses the default in question. These are interconnections that are not discussed by Walton.

Rescher's notion of plausibility makes authoritative reliability the pivot. In his conception, plausibility is also an inherently economic matter. It is cheaper to make do with the plausible than to hold out for the true. It is a conception tailor-made for our notion of the practical agent, who must prosecute his cognitive agendas under press of scant resources. The practical agent, we said, is someone who often is obliged to discharge his cognitive tasks on the cheap. If it is part and parcel of those cognitive economies to satisfice with regard to the plausible, it is reasonable to suppose that recognizing the plausible is something that beings like us are reasonably adept at. Certainly one of the economies practised by practical agents is reliance on the say so of others. Rescher recognizes three classes of such reliance: human ("Is this the way to Central Station?", "Modern science has discredited the idea that intelligence is accurately measured by IQ-tests); instrumental ("The thermometer registers 40 degrees C, so I've got a very bad fever"); and methodological ("Strings are the simpler hypothesis").

[5] We regret that this book went to the printer weks before we became aware of Walton's *Abductive Reasoning*, which was scheduled to appear in December 2004,[Walton, 2004].

Rescher also sees a link between plausibility and presumption. He cites

> *The Fundamental Rule of Presumption* A positive presumption always favors the most plausible contentions among the available alternatives. It must stand until set aside by something yet more plausible or by direct counter-evidence ([Rescher, 1976b, p. 55]; *cf* [Rescher, 1977, p. 38]).

Although the leanness of his exposition makes it difficult to determine with confidence, Walton appears to lodge his notion of plausibility in the conceptually prior idea of presumption. Thus a proposition is plausible to the extent that it has the appropriate presumptive *bona fides*. For Rescher, it is the other way round.

> For plausibility can serve as *the crucial determinant* of where presumption resides [Rescher, 1977, p. 37].

The Fundamental Rule of Presumption is a rule about propositional plausibility, which is at least a close approximation of the basic tie between presumption and plausibility. Yet its range is restricted to presumptions in the form $C(H)$. It does not hold for conclusions in the form "H is a good bet for testing". Even so, the rule can be adaptated for strategic plausibility. If H leads to the conclusion that $C(H)$ for strategic reasons, the presumption rule holds by virtue of the structure of conjecture. Accordingly,

♡ **Proposition 7.14 (Extending the Fundamental Rule of Presumption)** *A positive presumption always favours the most plausible contentions among the available alternatives. It must stand until set aside by something yet more plausible or by direct counter-evidence. This rule is indifferent to the distinction between propositional and strategic plausibility.*

There is much to approve in Rescher's account of plausibility, especially in the contrast he draws between plausibilities and probabilities. There is less to be said for how prior plausibilities are handled. Rescher is quite right to have accorded epistemic significance to authoritative sources. You turn on the radio and hear that Kabul has been bombed. So, you know that Kabul has been bombed, albeit defeasibly. You ask whether this is the way to Central Station, and you are told that it is. So now you know, albeit defeasibly. You consult the bruised horizontal slash in a prairie sky, and you know, albeit defeasibly, that both and wind and temperature will soon rise. You see the spots on your child, and report chicken pox to your pediatrician. Harry sees the open door, and assures himself that it wouldn't have been Sarah's doing; Harry's knowledge of Sarah's habits is authoritative. You hear the drumming on the roof and, looking out, you see that it is raining hard. But now your source is impersonal; it is the evidence of your senses. You want to know what your bank balance is. So you balance your chequebook. Here, too, your source is impersonal. It is the evidence produced by your calculations.

This is problematic. Impersonal epistemic sources are patches of evidence or arithmetic rules, or some such thing. It is widely agreed that evidence is good as opposed to bad, or better rather than worse, when it affects conditional probabilities in requisite ways. But Rescher wants his plausibilities to be different from probabilities. He owes us a nonprobabilitistic account of how such evidence gets to be authoritative. This has yet to materialize. In its absence, we can only conclude that the concept of authoritative source has too short a leash to tie down the more far-ranging concept of plausibility. Whether or not this is so on account of the indetermenency of the doctrine of evidential sources, it is certainly so in virtue of the limited goal Rescher has set for his plausibility logic. Probability logics cannot handle inconsistent inputs. The goal of Rescher's deployment of plausibility is to repair this omission. There is no other goal, hence no other standard against which to judge the success of Rescher's enterprise. Even so, there are some welcome byproducts of Rescher's logic. It would be unreasonable to hold the description of plausibility to no condition other than providing a method for determining what is reasonable to accept among a number of jointly inconsistent propositions. It lies in the nature of Rescher's task to get certain things right about plausibility itself, independently of the limited goal of his theory. Indirectly, he is contributing something to the logic of abduction.

Rescher's understanding of propositional plausibility overlaps with our own. Like him, we see a close connection between the assurances of hearsay and the plausible. In our account the principal source of hearsay is that which forms the relevant parts of common knowledge. In Rescher's hands, the idea of epistemically authoritative sources is raised to a not inconsiderable level of abstraction, with the result that whatever gives one reason to hold something provisionally (or conjecturally) is to be counted as an epistemically authoritative source. This requires Rescher to subject the notion of epistemic authorativeness (or source-evidence) to strikingly wide attributions of degree. We demur from this approach because it cuts across the grain of a distinction we take to be important. It is the distinction between having reasons to accept or (provisionally accept) a proposition and having evidence that it is true. Another difference between Rescher and ourselves is worth repeating. Rescher defines plausibility-consequence in such a way that whatever follows from a set of plausible propositions must have a plausibility value at least as high as the least plausible of those propositions. In the approach we favour, there are non-trivial cases in which the opposite is true. To take another example, we might imagine that the common sense generic that ocelots are four-legged has a quite high plausibility ranking (by Rescher's lights, not ours), whereas the default inference from that same proposition that since Ozzie is an ocelot, Ozzie is four-legged is (as it should be) of a lesser grade of plausibility. Not only is the inference reasonable, it is characteristic of a very large class of plausibility reckonings deeply embedded in actual cognitive practice.[6]

In Rescher's plausibility logic, plausibility is always a property of the propositional content of an hypothesis. But as Peirce has made us aware, it is perfectly possible to want to engage an hypothesis not for the plausibility of what it *says* but rather because of the plausibility of its fit with the the type of abduction problem at hand. In the oft-quoted example from Sherlock Holmes, after you have discarded all the rival alternatives, it is advisable to deploy the one remaining no matter how implausible. The distinction embedded in this advice is that between the plausibility of what a hypothesis says and the plausibility of entertaining and even engaging it.

7.8 Brief Concluding Remarks

The sublogic of engagement takes note of a natural affinity, but not a necessary condition, between explanationist abduction and propositional plausibility. Among a set of rival hypotheses, the logic sanctions the elimination of the implausible, and the selection of the most plausible of those that survive the cut, provided the most plausible is plausible enough. Implicit in these provisions is recognition of two ways in which an abduction might abort. We should put it that

Definition 7.15 (Abductive abortion) *A process of abduction aborts if either the most plausible of plausible candidates for hypothesis-engagement is not plausible enough, or the most plausible candidates, though plausible enough for engagement, cannot be rank ordered, and their number is too large.*

Corollary 7.15 *Small sets of unranked plausibilities may serve as non-unique abductive solutions, even when they are jointly inconsistent. Just as there is a natural link between an explanatory hypothesis and a requisite level of propositional plausibility, there is also an affinity between radically instrumental abduction and strategic plausibility.*

We shall also say that

Proposition 7.16 (Radically instrumental abduction and plausibility) *If H is a hypothesis selected in the making of a radically instrumental abduction, then H scores high with regard to strategic plausibility and low with regard to propositional plausibility.*

[6] Rescher adapts his consequence-relation from Theophrastus' rule that "the modality of the conclusion [of a valid argument without redundant premisses] must follow that of the weakest premiss" [Rescher, 1976b, pp. 13 and 23–25]. There is a difficulty with this. Theophrastus' rule is a satisfactory modal rule only on the assumption of logical omniscience [Cohen, 1979].

We have been suggesting a robust connection between the plausible and the characteristic, and, in turn, between the characteristic and the generic. The ties between the latter two is hardly one of straight equivalence. Even so, there appear to be ways in which singular statements of what is characteristic can be paraphrased so as to be derivable from properly generic claims. Preserving a role for genericity in the analysis of characteristicness turns out to be a welcome bonus. For, as we indicated in chapter 2, a ready capacity for generic judgement seems to be one of the dominant skills of practical agents in tight cognitive economies.

Plausibility then falls out as what conforms to what is characteristic, with the implausible contradicting the characteristic. In some respects this approach to plausibility fares no better than Rescher's. But perhaps two advantages might be claimed. One is that judgements of characteristicness may guide us without circularity in setting prior plausibilities. The other is, that since an account of characteristicness does not depend on a authority-based notion of epistemic legitmacy, we evade the problem of extending that notion to impersonal contexts.

Chapter 8

Relevance and Analogy

> "One ventures the assertion that 'presumption' is the slipperiest member of the family of legal terms, except its first cousin 'burden of proof'.
>
> *McCormick on Evidence*

> "Human cognition is relevance oriented."
>
> Dan Sperber and Deirdre Wilson

> "Structurally similar problems must receive correspondingly similar solutions".
>
> Bas van Fraassen

8.1 Relevance

In claiming that relevance is a constraint on abduction, we commit ourselves to the idea that, properly analyzed, abduction will disclose the presence of information that is somehow helpful to someone or something.

The agenda of an abductive agent is not only to hit his target T, but to do so in certain contextually circumscribed ways. An agent's abductive agenda is to make the best of a state of ignorance, to rise above it somehow. This involves his finding and engaging a hypothesis which, together with a knowledge-module K, will bear a certain interpretation of the $\looparrowright$-relation to a payoff sentence, in such a way that that target would be met if the hypothesis were true. Typically an abductive agenda subsumes a number of such agendas. This creates the possibility — indeed the likelihood — that information that proves helpful to a subagenda may not be helpful to a different subagenda. Thus we have it that information that is relevant to a subagenda is likewise relevant to its subsuming agenda, but that information that is relevant to an agenda need not be relevant to a subagenda. Accordingly

Proposition 8.1 (Compositional relevance) *Agenda-relevance is compositionally hereditary. Anything relevant to a subagenda of an agenda is relevant to it.*

But,

Proposition 8.2 (Divisional relevance) *Agenda-relevance is divisionally non-hereditary. Something relevant to an agenda is not necessarily relevant to all its subagendas.*

The relevance of information to an agenda is a compact, and not strictly accurate, way of speaking of the relevance of information for an agent in relation to his (or its) agenda. Seen this way, relevance is a set of triples $\langle I, X, A\rangle$ such that information I is relevant to agent X with regard to agenda A. We say that this condition is met when

Definition 8.3 (Agenda-relevance) *I is relevant for X with regard to A to the extent that in processing I, X is affected in ways that conduce toward the advancement or closure of A.*

Agendas are structured states of affairs terminating in an endpoint. At their most intuitive and informal, agendas are constituted by what the agent aims for (the endpoint) together with the counterfactual states of affairs sufficient for its realization. Information is relevant for an agent having an agenda when it acts on the agent in ways that occasion the realization of a constituting condition. Seen this way, neither agendas nor relevance is something that an agent need be aware of being or been influenced by, even though there are cases in which the opposite is certainly true.

8.1.1 Relevance as Cognitive

A conspicuous feature of agenda relevance is its definitional tie to information-processing. In a somewhat primitive way, this makes relevance a cognitive relation. In this, agenda relevance stands apart from the two most dominant of the alternative approaches. In all the standard systems of relevant logic, relevance is a relation on n-tuples of propositions [Anderson and Belnap, 1975; Dunn, 1994]. In the leading pragmatic account (to date), relevance is a deductive relation on propositions and contexts [Sperber and Wilson, 1986]. One disadvantage that these approaches share is that it is difficult to take a notion which is intrinsically a relation on propositions and ready it for service as a constraint on *thinking*. It is true that many of the properties and relations required by a theory of reasoning are dealt with extensively by logic. But it is routinely the case that, as they stand, they will not serve the requirements of a theory of reasoning, or not anyhow without considerable qualification (which often is neither easily nor credibly achieved). Thus a standard logic of consequence gets the closure conditions on belief wrong — and wrong in ways not sufficiently mitigated by talk of approximation to reasoning ideals. (See chapter 2 above.) The fix of standard logics' fix on premiss-inconsistency is wrong for belief revision. Propositional accounts of relevance are also wrong for the structure of relevant reasoning [Gabbay and Woods, 2003a, ch. 5 and 6].

From its inception, logic has sought to serve two masters. One is to specify and characterize sets of intuitively logical properties and relations that are definable for propositional structures or for these in relation to abstractively set-theoretic structures. Here the main goal is to get these target notions right, where the question of rightness is intimately bound up with the issue of rightness *for*. Accordingly, a logic gets consequence right if it is right for sets of sentences taken without reference to factors of speaker-use and other pragmatic considerations. Historically, logic's further purpose has been to give a rigorous characterization of the canons of strict reasoning. To this end, logic has tended to generate the inference-rules required for accounts of reasoning by extending or otherwise qualifying the truth conditions on logical relations such as consequence and relevance. Aristotle started this trend, and notwithstanding the momentous lurch towards the mathematical in the past century and a quarter, the feeling still persists among logicians that a logically adequate account of reasoning requires rules of inferences that are got by "inferentializing" the appropriate laws of logic. The great majority of the nonstandard systems of deductive logic during this same hundred and twenty-five years have attempted to facilitate this conversion of truth conditions into rules of reasoning. This process is typified by Aristotle's logic of the syllogism. Aristotle thought that this logic would be the theoretical core of a wholly general theory of argument. Since the Greeks tended to think of inference as *interior* argument, the same would apply to a wholly general theory of inference. To achieve this end, Aristotle laid several constraints on what we recognize as classical validity (he called it "necessitation"). The result was the syllogism. A syllogism is a valid argument whose premisses are non-redundant, whose conclusion repeats no premiss, and whose conclusion is non-multiple. Like most logicians, then and now, Aristotle saw the reasoning he sought to characterize as the drawing of consequences from sets of propositions. Since what a proposition's consequences are is always relative to the consequence relation definable for such, Aristotle defined the relation of syllogistic consequence with his reasoning goal in mind. So syllogistic consequence just is classical consequence cut down by these constraints.

Reasoning by strict (or truth-preserving) consequence is always a matter of evacuating information already present in premisses. For this reason, reasoning by strict consequence can be called *archeological reasoning*. Under Aristotle's constraints every piece of archeological reasoning evacuates its premisses of their total syllogistic information, which is then repackaged in the single proposition that serves as the reasoning's conclusion. The logic that produces this result, the logic that Aristotle called the logic of syllogisms, is the first relevant (hence paraconsistent) non-monotonic, intuitionistic logic ever produced [Woods, 2001].

We liken the two longest serving approaches to relevance to classical consequence in Aristotle's theory. Like it, these concepts of relevance are too coarsely grained for deployment in psychologically realistic theories of reasoning. Something like Aristotle's inferentializing process is required to transform relevance from a strictly propositional relation to something that can serve the more complex and more fluid requirements of human (-like) reasoning in real time. (We note in passing that although Sperber and Wilson introduce the much needed factor of context into their account, contexts as developed there turn out to be too abstractly propositional for these further purposes. Even so, contexts bear some affinity to agendas [Gabbay and Woods, 2003a, ch. 6].)

The account developed in *Agenda Relevance* tried to keep these points in mind. To the extent possible, we would attempt to get the principles of relevance by inferentializing prior results got at the merely propositional level. Beyond that, we would feel at liberty to improvise.

This brings us back to the point that prompted the present digression. Agendas for us are structures whose closure can be attained as a result of information-processing. They are in that sense either overtly cognitive tasks or predispositions to cognitively realizable ends. Thus Harry may wish to know whether it snowed overnight. This is his agenda. He goes to the window and looks down into the street. On seeing the cars newly piled up with drifts, Harry is made to realize that it did in fact snow overnight. This closes his agenda. The information that the cars below were covered in snow was relevant for Harry with regard to that agenda. It got him to know what he wanted to know. In other cases, agendas can be more tacit, with endpoints the agent has no consciousness of. Even so, information may put Harry in a cognitive state which, had he thought about it, he would expressly have wished to be in, and, now that he is in it, he may recognize a prior tacit interest. When Sally is told that her married daughter has won the lottery, this may have caught her, and everyone else, wholly unawares. We may assume that Sarah had never entertained the idea of her daughter winning the lottery, hence that she lacked occasion to set herself the cognitive task of finding out whether the lottery had in fact been won. But Sarah has a standing interest in her daughter's welfare, hence an abiding (and largely tacit) agenda to be in cognitive states that keep her appropriately apprised.

The cognitive character of agendas spill over into agenda relevance. We may say that the fundamental fact about relevance is that it is cognitively helpful information. Such a view of relevance is tailor-made for abduction. Although a conjectured H need not score at all well by presently available epistemic standards, there are things that a would-be conjecturer of H should want to know before a decision on H is made. We see in this elementary fact an initial contrast between information that is relevant to H and information that is relevant to whether H should be conjectured and discharged (or, for that matter, sent straight to trial).

Having taken pains to explain why propositional relevance is typically too coarse-grained for use as a relevance-rule in theories of reasoning, it is necessary to emphasize that, even so, propositional relevance can play a subsidiary role in theories of reasoning. In the standard approaches to relevant logic, there are two main conceptions of relevance. In what can be called *topical relevance*, one proposition is relevant to another when the two share at least one atomic component. In what may be called *full-use relevance*, a proof of a proposition from a certain set of hypotheses is a relevant proof if it employs all the hypotheses in that set.

8.1.2 Topical Relevance

In its original setting, topical relevance is an elementary syntactic constraint. More complex treatments can be found in works such as [Walton, 1992b] and especially [Cuppens and Demolombe, 1988; Cuppens and Demolombe, 1989; Demolombe and Jones, 1999]. Topical relevance is a factor of central importance in the construction of inventory-search technologies and other kinds of taxonomic devices, illustrated by the basic idea of *search-by-subject*. When an abductive agent considers a range of candidates for possible engagement, two things may reasonably be assumed, as we have already said. If the abducer is on the right track, then the set of candidates is a proper subset of a larger set of possibilities; and the remaining H, if there is one, will itself occupy a proper subset of this subset. Here again two further questions present themselves. One is whether these subsets are constrained by a relation of topical relevance. The other is, assuming an affirmative answer to the first, whether in considering these candidates for possible engagement, the abducer makes a prior or concurrent judgement of topical relevance. Of course, we are asking these questions of topical relevance. Topical relevance is not, and does not generalize to, agenda relevance. We are espousers of agenda relevance; but it is not our view that agenda relevance is all there is to relevance. Our view is that agenda relevance is the most comprehensive of the going conceptions of relevance, concerning which topical relevance is a case in point. Although topical relevance is hardly the same as agenda relevance, it might reasonably be expected that agendas are sometimes such that topically relevant information is also relevant to their advancement or closure. It is well to note, however, that this is not what our first question asks. It asks whether candidates for possible engagement are carved out from larger possibility spaces by considerations of topical relevance. Even if they are, the question does not ask whether that abducer has *information* to this effect or whether having it conduces toward the advancement or closure of his agenda. Another way of asking our question is this: Filtration structures have been posited to exist for any H or small set $\{H_1, \ldots, H_n\}$that a successful abduction engages. Is this posit correct as regards the assumption of a relevance filter if the filter is topical relevance? The question is stacked with qualifications. *If* filtration structures always attend a winning conjecture, and *if* filtration structures invariably contain relevance filters, is it invariably the case that the filter is topical relevance?

We see no way of making the case either that a winning H is always topically relevant to the phenomenon that generated the abducer's target T or that rival hypotheses are invariably topically relevant to one another. Perhaps the greatest impediment to such claims of universality is the existence of out-of-the-box thinking in the context of what Peirce calls "originary" abduction. Whether this is strictly true will turn on details of the purported topicality. But intuitively there are successful abductions in which the abduced hypothesis appears to have had "nothing to do" with the target T or with alterative conjectural possibilities.

A safer conjecture is one that takes it into account that topical relevance is a natural guide in searches of possibility spaces. Accordingly,

♡ **Proposition 8.4 (Topical relevance)** *There is a significant likelihood that, for any H having a determinate place in a filtration-structure, H will be a member of a set of possible conjectures, each of which is topically relevant to T and to each other.*

Corollary 8.4(a) *There is no empirical evidence that in engaging such H's, abductive agents actually make these judgements of topical relevance.*

Corollary 8.4(b) *It is possible that such judgements are made tacitly, at the level of a logic of down below.*

Of course the devil is in the details. There are treatments of topical relevance (such as propositional variable-sharing) for which Proposition 8.4 is not very plausible in general, and certainly false in various specific cases. But the logic of topical relevance is an open research programme, which extends well beyond what we have space for here. Suffice it that we pass on the details of the present suggestion for further investigation. What is needed is an articulation of topical relevance for which Proposition 8.4 stands a decent chance of being true.

A natural question is whether the topical relevance that attends a winning H is itself agenda relevant with regard to the agenda that such an H be found. This turns out to be a question of rather striking difficulty. The stumbling block is the information parameter. Agenda relevance is defined for information that an agent processes. The theory of agenda relevance extends considerable latitude to this notion. (For a good over-view, see [van Benthem and van Rooy, 2003, pp.375–379].) So it is perhaps not obvious as to why the information-processing condition of relevance would pose a problem. One of the liberties the theory extends to the idea of information-processing concerns the processing part. The theory acknowledges that a great deal of the information that a cognitive system takes in is processed tacitly. A second liberty extends to the idea of information itself. Because the distinction between energy-to-energy transductions and energy-to-information transformations is neither exact nor well understood, we have postulated a conception of information which, intuitively speaking, might be thought to intrude too much into the energy-to-energy transduction side of this distinction. The problem is that a winning H could satisfy the projected constraints on topical relevance without it being the case that any information to this effect — even the extended sense of information proposed in *Agenda Relevance* — becomes available to H's abducer. This is not just to say that H might be topically relevant in the appropriate ways without the agent judging that this is so, but rather that it might be relevant without the abducer having the slightest information to that effect, tacit or otherwise. This being so, such information, if it exists, is not *actually* of relevance for the abducer's agenda. If he doesn't process it, it cannot be the case that in processing it, he was affected in ways that conduce to the advancement on closure of his agenda. This leaves the question of the subjunctive relevance of that information [Gabbay and Woods, 2003a, pp. 184–188]. Is it the case that were the abducer to have processed it, he would have been affected in ways that brought his agenda (closer) to completion? It would seem that the answer is "Not very likely". For given that the topical relevance of candidate hypotheses to the propositional content of the abductive target T substantially undetermines T's attainment,

Proposition 8.5 (Topicality and direct relevance) *Information that candidate hypotheses are about some same topic as the state of affairs that occasioned an abductive target T tends not to be* directly *relevant in advancing or closing the agenda of finding an H that attains T.*

Topical irrelevance is another matter however. Just as awareness of implausibility serves as an excluder of candidate hypotheses, so too does the awareness of irrelevance, provided some conditions are met. One is that the exclusion itself is defeasible. The other is that the excluder has some grasp of the fact that there is a significant positive correlation between what candidate hypotheses are about and what an abductive trigger is about. Subject to these constraints, it would seem that we could put it that

Proposition 8.6 (Topicality and irrelevance) *For information I, to the effect that a candidate hypothesis H is topically irrelevant to an abductive trigger T, and information I^*, to the effect that there is a significant correlation between what* *the candidate hypothesis is about and what an abductive trigger is about, if an abductive agent X were to process I and I^* (however tacitly), there would be some significant likelihood that H would not be selected by X as a possibility to be considered for abductive engagement.*

With plausibility and relevance alike, it is the negative form that wears the trousers. In both cases, the abducer has a large stake in considering low finite numbers of options. Plausibility is a weak marker for conjectural felicity, whereas implausibility is a stronger marker for conjectural infelicity. Topical relevance, likewise, is a weak marker for conjectural felicity, whereas topical irrelevance is a stronger marker for conjectural infelicity.

Proposition 8.5 holds for what might be called the direct relevance of topical relevance to an abduction's main agenda. It claims that information that a candidate for possible conjecture and the abductively triggering phenomenon are about some same topic would tend not, in being processed by an abductive agent, be of help to him in picking the H he should actually engage. But subagendas may be another matter. From what we know of the cognitive make-up of human individuals, it is advantageous to operate with small option spaces rather than large. Option spaces linked by considerations of topical relevance, both to their contained members and to the requisite abductive trigger, are smaller than unstructured spaces of possibilities and, even if nevertheless large in their own right, are compact rather than scattered, thus making access more efficient in principle. We may take it, then, that abductive agents have an agenda for the slimming down of option spaces, albeit typically this will be a tacit agenda. Even so, information to the effect that such and so hypotheses are linked to one another and to the triggering phenomenon by topical relevance may be assumed to be such that if it were processed by an agent would affect him in such ways as to direct the resources of his cognitive wherewithal to the smaller option space, carved out by the topical relevance constraint. We may sum this up by suggesting that

♡ **Proposition 8.7 (Topical relevance and subagendas)** *If an abductive agent processes information to the effect that a set of possible conjectures is topically intra-relevant and topically relevant to an abductive trigger, it is likely that in processing it the agent will be affected in ways that guide the direction of his cognitive devices to that set, rather than its opposite number. This would advance the agent's subagenda to work with smaller rather than larger option-spaces.*

Corollary8.7(a) *Again, there is no empirical reason to think that it is typical of such cases that the information relating to topical relevance is consciously processed or that it eventuates in the judgement that the set in question is a topically relevant subset, still less the judgement that it is likely that the winning H is to be found in this set.*

8.1.3 Contextual Effects

Of the two central conceptions of relevance dealt with by relevant logic, full-use-relevance more nearly resembles agenda relevance, which can be seen as a generalization of it. The same can be said for Sperber and Wilson's contextual effects notion of relevance. In the first case, a proof from hypotheses is relevant when all the hypotheses are actually used. Since all are used, all conspire to achieve the proof's conclusion. Thus a relevant proof from hypotheses can be seen as a disembodied agenda to achieve the proof's intended conclusion; and its preceding lines can be seen as agenda relevant in as much as each helps in the goal's attainment. In the second case, a proposition P is said to be relevant in the sense of Sperber and Wilson (or SW-relevant) with regard to a set of propositions C called a "context", if some Q not deducible from either P or C alone is deducible from them together, or if P either contradicts some proposition in C or strengthens (or weakens) the probability of some proposition in C. What is striking about this contextual-effects relevance is its structural similarity to the basic set-up of an abduction problem. This can be seen if we liken C to an abducer's knowledge-set K, P to his (or its) hypothesis H and Q to some target proposition. The similarity is not exact, but it is welcome all the same. For if one's target were to deduce some particular Q (or to strengthen or weaken some particular S) in C, that this cannot be achieved with the resources at hand in C means that, if it is to be achieved at all, new assumptions P will have to be deployed. On the theory of Sperber and Wilson, any P that fills this bill is relevant. But, as we have seen, not everything that hits an abductive target will count as abductively adequate. So the comparison stops here. Although Sperber and Wilson always refer to these Ps as assumptions, there is no formal requirement that they be forwarded only as conjectures. Neither does their account allow for consequence relations other than (a qualified) deductive relation or an inductive relation of conditional probability. Even so,

Proposition 8.8 (Abduction and contextual effects relevance) *Any abduction problem whose $\rightsquigarrow$-relation (where applicable) is interpreted in the appropriate deductive or probabilistic ways is such that any H that solves it is SW-relevant in K with regard to Q.*

Corollary 8.8(a) *Since the account of Sperber and Wilson easily extends without formal or conceptual loss to systems in which their original consequence relation is given the interpretational range of a standard abduction problem, then the qualifications in Proposition 8.8 are not essential. Thus any H that solves an abduction problem in relation to K, for some target Q, is SW-relevant in K with regard to Q.*

Sperber and Wilson's third way for a proposition P to be relevant in a context C is for P to contradict some Q already in C, thus occasioning the necessity to revise

C. It will not have escaped notice that this form of SW-relevance instantiates the kind of abductive trigger which Alesida calls "anomolous" [1997]. Accordingly

Proposition 8.9 (Relevance as trigger) *When P is SW- relevant by way of its contradiction of a Q in C, then P constitutes an abductive trigger for any cognitive agent holding the beliefs in C.*

In *Agenda Relevance*, we noted as a limiting case information which not only closes an agenda, but does so in a way that creates a new one. Consider a case [Gabbay and Woods, 2003a, pp. 188–189]. Sarah and Harry are trying to decide where to spend the long weekend in Banff. Sarah is scanning the paper for hotel prices. She turns a page, and her eye falls on an obituary headline. "Oh no!", she exclaims. "Freddie has died. We'll have to check tomorrow's flights to London". In processing the information contained in the headline, Sarah's former agenda is aborted, and a new one was created. The new agenda is to discover the times of flights to London tomorrow. In this limiting sense, new information was agenda relevant. Accordingly, we may say that anomoly-triggers have something of the structure of agenda relevance. The presentation of P creates an agenda for any holder of the beliefs in C. The agenda is to restore consistency. So

Proposition 8.10 (C-contradiction and agenda relevance) *Relevance by way of C-contradiction instantiates the basic form of an anomolous abductive trigger, which in turn instantiates the structure of agenda*-creating *relevance.*

8.2 Irredundancy Relevance

We have sketched the role that topical relevance might reasonably be expected to play in abductive contexts. In this section we propose to put our earlier suggestions to the test. We do so by considering the following possibility. Given that the principal function we have ascribed to topical relevance in abductive contexts is to shrink possibility spaces, could it be the case that other conceptions of relevance exist which perform in like manner, but more efficiently? One possibility is a conception of relevance which is a restriction of the full-use conception (see [Gabbay and Woods, 2003a, p. 297]). According to the full-use conception, premisses (or hypotheses) are relevant in a proof if and only if each premiss (or hypothesis) is used in the proof. It is a restriction of this view, originating with Aristotle, that no premiss (or hypothesis) be redundant. It is easily seen that premiss-irredundancy implies full-use relevance, but not conversely. The premiss irredundancy-conception of relevance resembles, but is not a case of, linear relevance; neither is linear relevance a case of it. A proof is linearly relevant if and only if all its premisses (or hypotheses) are used once and only once. Irredundancy-relevance allows for the multiple use of premisses, provided that for each such use its omission would cripple the proof. However in proofs by syllogistic inference, it is also clear that premiss-irredundancy and linear relevance coincide. Aristotle never speaks of the premiss-irredundancy condition on syllogisms as a relevance condition. But given its kinship with full-use relevance, there is no harm in speaking of *irredundancy-relevance.* The question now before us is whether the role that relevance intuitively appears to play in abduction might more efficiently be discharged by irredundancy-relevance than by topical relevance.

If a consequentialist abduction problem has a solution, then for some H, K, V and T, there is an interpretation of the $\looparrowright$-relation such that $K(H) \looparrowright V$ and that the truth of this conditional is such that $K(H)$ may be said to hit the abductive target T. In the early stages of this study, we required K to be the least class of what the abductive agent then knows for which these conditions hold. It would appear that the same constraint is also justified for H. Consider now the domain D of the $\looparrowright$-relation for which the present conditions hold. It is easy to see that the least class D^* of elements in D for which these same conditions obtain satisfies the constraints on irredundancy. It is also easy to see that the least class D^* of elements in D for which these same conditions obtain satisfies the constraint on irredundancy as adapted to the antecedents of conditionals. We have it that all such conditionals whose antecedents are drawn from D^* are conditionals with irredundant antecedents. Let N be the set of all such conditionals. We may put it that whenever an abduction problem has a solution, the required conditional exists in N. Call this a N-conditional.

The structure of N-conditionals places considerable pressure on the admissibility of Hs. No H can enter the antecedent of N-conditional unless it in conjunction with a subset of K constitute the least antecedent for which it is true that $K(H) \looparrowright V$. These are relevance constraints. The question is, do they serve the two principal functions played by topical relevance? If so, do they serve them equivalently? And do they serve them as efficiently?

Intuitively, it is reasonable to suppose that winning H will meet some kind of plausibility condition and some kind of relevance condition. But far and away the more basic and irreducible requirement is that H deliver the goods in the antecedent of a conditional in the form $K(H) \looparrowright V$. Suppose now that this is so for some interpretation of the parameters of an abduction problem. Suppose further that as it there occurs, H is not topically relevant to the state of affairs that triggered the abduction problem, but that there is some other H^* for which $K \cup \{H^*\} \looparrowright Q$ holds and its antecedent is redundant. Do we have a reason to favour consideration of the topically relevant but redundancy-tainted H^*s over the topically irrelevant but irredundant Hs? On the question of smallness of candidacy-spaces, the nod would seem to go to the irredundant Hs. Score a point for irredundancy-relevance. On the other hand, these very Hs lose whatever competitive edge that topical relevance may confer. In making our earlier case for a role for topical relevance in the solution of abduction problems, we helped ourselves to a necessary assumption. The assumption was that there is some significant likelihood that candidates for conjecture will be topically relevant to the triggering phenomenon; that there is nontrivially something or other that they are both about. It was an assumption tendentiously made, perhaps. But if we made it in that earlier context we can hardly not also make it here. This turns out to be consequential. It provides that topical relevance and irredundancy relevance will not in general converge. But this leaves it open that for H's that win the contest for selection in actual abductive practice, there is a positive likelihood of positive conversion.

It is a fundamental requirement, carrying a cost that abducers must pay, that in solving an abduction problem the requisite conditional hold. Truth conditions on conditionals are not always an easy thing to specify, but no matter what our preferences for relevance may run to, they must in any case be negotiated by the abducer. It may be suggested that, troublesome as they sometimes are to produce, truth conditions on conditionals are in general easier to specify than truth conditions on topical relevance. Whether this is actually so, it is unarguable that conditions on the irredundancy of the antecedent of such conditionals is a much easier "go" than producing as theory of topical relevance. So we conclude that for the general range of cases in which antecedent-irredundancy and topical relevance co-occur, it is more economical to form candidate spaces on the basis of admissibility to the role of irredundant antecedents. This leaves the residue to be decided. These are the cases in which topically relevant H^*s that occur only in requisite conditionals with redundant antecedents are pitted against topically irrelevant Hs that occur only in irredundant antecedents. The matter is settled as follows. Take any such H^*, and the conditional in which it occurs. Now form the smallest part of H^* which is itself an H for which the conditional holds. By the construction of the example, the residue of this contraction (a) is the source of H^*'s topical relevance and (b) makes no contribution to the fact that the antecedent in question yields the requisite consequent. So we conclude that when they do not converge irredundancy relevance trumps topical relevance.

There is, however, a further consideration that requires a certain emphasis. As we saw from our discussion in chapter 3 of the minimality constraint of the AKM-model, too much redundancy is a liability, but some redundancy is an aid to inferential flow. Accordingly, whereas an N always exists when an abduction problem has a solution, it is not a condition on the agent that he actually embrace N.

8.3 Relevance and Cutting to the Chase

In finding a place in the structure of abduction for both topical and full-use relevance, we have not lost sight of the claim to which we gave considerable emphasis in *Agenda Relevance* and which we have re-asserted here. That is the claim that of all the notions of relevance to date, for which at least minimal levels of theoretical articulation exist, agenda relevance is at once the most fundamental and the most inclusive of these. As we have seen, that topical, full-use and irredundancy relevance often co-occur with agenda relevance. Their concurrence typically arises in situations in which information that is relevant in either of the first three senses is information that conduces to the advancement or closure of an agenda. Upon reflection, these interconnections are often tighter than anything suggested by mere concurrence. True, it is sometimes the case that information that is, say, topically relevant is information helpful to the advancement of an agenda. But as our examples also show, it is sometimes the case that what makes information that chances to be topically relevant helpful to a cognitive agent is precisely that it is topically relevant.

The entering wedge of our discussion is what we have been calling cut-to-the chase abduction. By now enough has been said to make us see that, with the possible exception of outside-the-box abduction, every solvable abduction problem has an H (or a small set of Hs) that has a determinate place in a filtration structure. Thus for every abductively successful outcome, these are sets of conditions — conditions as possibility, conditions on relevance, conditions as plausibility — which the outcome satisfies. We may say, then, that

Definition 8.11 (Cut-to-the-chase) *In solving an abduction problem, a cognitive agent has performed a cut-to-the-chase abduction if and only if the abduction was performed without behavioral or introspective trace of negotiation with the conditions reflected in the problem's concomitant filtration-structure.*

Proposition 8.12 (Abductive speed) *It is typical of, but not essential to, cut-to-the-chase abductions that they are performed quickly.*

Proposition 8.13 (Cut-to-the-chase dominance) *What is known empirically of human cognitive agency suggests that the frequency of cut-to-the-chase abductions relative to abductions performed is statistically high.*

It is well to note that not every cut-to-the-chase abduction is speedily achieved. The definition of this class of abductions leaves room for ranges of exceptions, also empirically attested to. It sometimes happens that, when presented with an abduction problem, the agent is flummoxed by it. He has no idea of how to proceed. Days and weeks might go by during which the agent's only discernible behaviour in regard to his problem is his fretting about not knowing how to get on with it. Then suddenly, he awakens at three in the morning. "I've got it!", he thinks. And he might well be right.

Cut-to-the-chase abduction calls to mind Peirce's notion that successful abductions are produced by our innate flair for guessing right (that is, right enough about enough of the right things, enough of the time). The guessing that Peirce is here invoking is not of the jokey or slap-dash variety with which we might associate the name of Popper. Peircean guessing is a matter of surrendering to the insistence of an idea. The very fact of cut-to-the-chase abduction indicates that Peirce is onto something important, notwithstanding the perfectly correct observation of Thagard and others that Peirce has nothing to say about the structure of this instinct. It is hardly surprising. Our instinctual mechanisms do not expose their secrets in empirically lavish ways. Their structure and and mode of operation are themselves occasion of abductive reflection. In other words,

Proposition 8.14 (Cut-to-the chase triggers) *The existence and typical speed of cut-to-the-chase abductions are themselves an abductive trigger.*

Our present task is to lend encouragement to the claim that agenda relevance is relevance's most fundamental conception. We are suggesting that if this claim is true, there should be some indication of its truth in those contexts in which relevance leaves a large footprint. Abduction is such a context; and cut-to-the-chase abduction is our particular focus. We abduce as follows.

In one of the epigraphs to this chapter, we find Sperber and Wilson's insightful remark that human cognition is relevance-oriented. It is an orientation in which irrelevance would seem to wear the trousers, as we have said. At any given juncture, the human agent is awash in oceans of information, most of which, given the particularities of his situation then and there, is noise. A striking feature of the overall competence of the human agent is his aptitude for the suppression of noise. In suppressing waves of noise, the human agent is suppressing information that is irrelevant to how he is presently situated. The human agent is an evader of massive irrelevance. It is easy to see that what the agent manages to evade is information that would be useless (or worse) in the prosecution of his cognitive agendas.

The aptitude for irrelevance-evasion is accompanied by a second, equally striking, capacity. In lots of cases — certainly typically — the human agent is able to achieve these exclusions very quickly. The net effect of these aptitudes is that at endless turns in an hour of human life agenst are in a situation in which an abundance of agenda irrelevance has been suppressed in the blink of an eye. This suggests a twofold connection with Peirce's guessing instinct. Irrelevancy evasion would itself appear to be instinctual, and it would also appear to be integral to the structure of guessing. Here is why. To the extent that an agent is competent in the suppression of agenda irrelevancy, what remains of any given episode of suppression is information which is to some non-zero degree relevant. The whole thrust of the aptitude for the suppression of irrelevance is that impediments to the achievement of cognitive ends be removed. This being so, it may be said that if irrelevance evasion serves its role competently, the net effect will be stocks of information which, at a minimum, do not impede agenda advancement and which, in large ranges of cases, actually facilitates it. If this were not the actual outcome of irrelevance-evasion in the general case, we would be forced to explain the comparative accuracy and speed of our cognitive achievements in the light of information that not only didn't facilitate our cognitive tasks, but actually impeded them. The better explanation of cognitive success is that it is attended by helpful rather than unhelpful information.

The phenomenon of cut-to-the-chase abduction is a particularly vivid representative sample of our cognitive speed. What makes it stand out is that, when successful, it owes nothing of its success to confirming evidence for H. Abduction is inference without the benefits of evidence; it is inference achieved in a condition of ignorance. This places a special premium on a particular form of irrelevance evasion. It makes it hugely important that the residues of irrelevance evasions be helpful to the cognitive task at hand. Given the construction of the human agent this could not happen unless the residues were small, as well as otherwise helpful.

In this we see the inapplicability of other conceptions of relevance we have been discussing in this chapter. If the residue of irrelevance suppression were held to the standard of topical relevance, and it only, the residue would not in general be either small or otherwise helpful. If residues were required to be full-use relevant and nothing else, they could be arbitrarily large and heftily useless otherwise.

Likewise, if residue were SW-relevant and nothing else, it would be ill-suited for the task at hand. Recall that SW-relevance is instantiated in three ways. In the first, adapting to the consequentialist abductive context, H is relevant to K, when for the payoff-proposition V, $K(H) \looparrowright V$. In the absence of further constraints on K and H, H could be both large and abductively useless. The latter would be so when deductive consequence is not the consequence relation that the abduction problem requires to be instantiated. The second mode in which a proposition P is SW-relevant with regard to a context K is when, for some proposition V in K, its probability is raised or lowered on condition that P. This has little to do with abduction. Even if we ignore the fact that, being in K, V cannot itself be an abductive target, that P raises the conditional probability of V, supposing that V is a payoff for the target, lends no support to the proposition that V should be conjectured. In fact, the opposite is true. The higher the probability of V on P, the greater the evidence for it, and the greater the evidence for V the less it qualifies as a candidate for abductive conjecture. Finally, the third way in which a proposition P achieves SW-relevance in a context K is by the inconsistency between P and some proposition V in K. As we have seen, this can constitute an anamolous abductive trigger, but it is not part of the problem's solution.

The exception is irredundancy-relevance. If the residue created by irrelevance-suppression satisfied the conditions on irredundancy-relevance, then the residue would be the smallest set of information of help to the agent in his cognitive endeavours. And that itself is a virtue. It is a manifestation of agenda relevance.

What the operation of irrelevance evasion produces is agenda relevance. It produces agenda relevance at its most relevant when the residue it creates is also irrendundantly relevant. Irrendundant subsets of helpful information not only lose nothing of this original helpfulness, but given the virtues of smallness, they represent a gain in helpfulness. We shall say that

♡ **Proposition 8.15 (Convergence on irredundancy)** *The operation of irrelevance evasion tends to converge on irredundancy.*

This achieves the objectives of the present section. We have been able to show the dominance of agenda relevance as an abductive constraint. And we now have some better idea of the structure and mode of operation of the Peircean notion of insistent ideas. Why, then, in cut-to-the-chase abductions does the agent make the abductive choices he actually does make? Such choices are in the residues of those operations of irrelevance-suppression which tend to converge in irredundancy? Why are such choices "insistent"? Because they are *purely* helpful. Why is cut-to-the-chase abduction performed so quickly? Because irrelevance evasion is also performed quickly.

As we now have it, agenda relevance is not only the dominant mode of relevance in solutions of abductive contexts, it also comprehends other factors, such as plausibility. We earlier proposed a significant tie between abductive winners and plausibilities. If this is so, plausibility is a condition, albeit defeasibly, on large groups of abductive solutions. And if this is so, propositions that failed the plausibility test would not be helpful information. Accordingly, plausible information is typically agenda relevant; and the residues of irrelevance supression will therefore meet all applicable requirements of plausibility. It is in this sense that

Proposition 8.16 (The dominance of relevance) *Relevance dominates over plausibility.*

8.4 Legal Relevance

In Anglo-American jurisprudence, the relevance of a claim is that which increases or decreases the probability of some other claim [Cross and Wilkins, 1964, p. 148]. Legal relevance is therefore a straightforward case of probabalistic relevance (concerning which, see [Gabbay and Woods, 2003a, p. 92–101]). A cursory examination of standard textbooks on the law of evidence will show — [Cross and Wilkins, 1964; Murphy, 2000], for example — that the dominant judicial approach to relevance has to do with grounds for the admittance or exclusion of testimony.[1] This is especially the case in matters having to do with the accused's character. It

[1] See here [Klotter and Ingram, 2003, p. 71]: "The rules of evidence, for all practical purposes, are rules of exclusion. All evidence is admissible unless it is excluded upon objection".

is frequently the case that such decisions, for the admittance or exclusion evidence, are decisions taken on grounds of relevance or irrelevance. Equally, it frequently happens that these juridical determinations are either taken on grounds of, or even comport with, the definition of relevance. What a judge is required to do is to determine whether such evidence would, if heard, prejudice the jury, or induce it to give it more weight than it should. Think of a case in which the accused is charged with paedophilia, and the evidence on which the judge must rule is a prior history of violent sexual predation (but not paedophilia). The law of evidence requires the judge to refuse to hear this testimony if he determines that the jury will make more of it than it should. The judge is not in general required to ascertain whether this evidence increases the likelihood of the accused's guilt; rather he is required to fine whether this evidence — even though it did increase the probability of guilt — would violate the special protections the criminal law has evolved for person's indicted for serious offences. One such protection is jury impartiality and freedom from bias. Another is the standard of proof for conviction, underwritten by the law's tactical skepticism concerning what would suffice to demonstrate guilt in ways that English lawyers call *safe*. When a judge finds that these protections would likely be compromised, he enters a finding of *irrelevance*, and does so irrespective of whether the evidence in question would fail to affect the probability of the proposition billed in the indictment. In its day-to-day operations, the law embodies a notion of relevance which is orthogonal to the relevance it formally defines. The embodied notion of relevance is a matter of what bears on the court's first obligation which is to avoid wrongful conviction. (See [Cross and Wilkins, 1964, pp. 148–149, 153–156], and [Murphy, 2000, pp. 8–9, 132–149, 162–167, 178–179, 216–219, 360–365].)

This is a textbook case of agenda relevance if ever there were one. The court's agenda is to avoid wrongful conviction. Embedded is an important subagenda, conditional upon the first. It is that if producing a result that meets the highest standards of accuracy compatible with an assured realization of the primary objective. The primary objective could be called *Cliffordian*. Its objective is error avoidance even at high epistemic costs. The endpoint of the subagenda could be called (limitedly) *Jamesian*. Its goal is the attainment of truth, but not at the cost of violating Cliffordian strictures. In standard scientific practice, the emphasis is reversed. Robust science aims for truth, and does so in ways designed to keep errors to a manageable level. In most of what an individual cognitive agent does, this Jamesian dominance is also present. The law's departure from this dominance is therefore an important deviation from standard cognitive practice. Even more striking, is the sheer extent of the law's favouritism toward Cliffordian restraint. The flipside of the law's hostility to epistemically wrongful conviction is its tolerance of epistemically wrongful acquittal. In its determination to avoid wrongful conviction, the law sets the standard of proof artificially high. (This anyhow is the received view. See below). This means that by epistemic standards that are *not* artificially high, acquittal may be known to be epistemically unjustified even though sanctioned by canons of legal correctness. The law's toleration of acquittals known to be unjustified by reasonable though not artificially high standards, is the fundamental operational expression of a fact of central importance to a system of justice. It is that justice trumps truth; more carefully, it is that the avoidance of injustice takes precedence over the attainment of truth.

8.4.1 Ideology

For these reasons, it is easily seen that legal systems such as those that evolved in Anglo-American jurisprudence have something of the same basic structure as ideologies and dogmatic religions. All are systems that impose prior constraints on what the evidence is allowed to show. In purely operational terms, the best way of disarming such evidence is to refuse to hear it. This, too, is often the counsel of the religious leader or the ideologue: Stay away from considerations that may tend to discredit the requirements of orthodox belief. Thus some Christians advise against a secular education as an "occasion of sin"; and some ideologues recommend the prosecution or expulsion of those who consort with non-believers. In all three cases, there is known to be considerable potential for epistemic distortion, yet in only one is this knowledge given much formal recognition. It can safely be said, however, that of the three, the law's epistemic triflings have a non-epistemically coherent motivation. It is to avoid the injustice that inheres in the committing of a certain kind of mistake. In the other two cases, it is easier to find a presumably coherent epistemic motivation for the selective suppression of undermining evidence. This flows from the fact that the fulcra on which these constraints operate are themselves taken to enjoy the requisite sort of epistemic privilege. The ideologue will hear no evidence against P since P is known to be true. But when a judge refuses to hear evidence against the accused's presumption of innocence, it is not that that would disturb a fact known to be true, but rather because in hearing it the accused's artificial protections might be compromised.

The legal indifference to what might be called the natural epistemic weight of procedurally inadmissible evidence takes to an extreme the logic of epistemic deficits that are integral to the existence and to the solution of abduction problems. In the abductive context, the problem is that you have a target that exceeds your epistemic resources; and the solution of the problem requires that this epistemic short-fall remain in place throughout. Abductive reasoning is reasoning in conditions of ignorance. It is reasoning in which a proposition H is conjectured in the absence of evidence that H. The law goes this one better. It is a set of procedures eventuating in a finding F which may be known to be false. This anti-epistemic character of the agendas of Anglo-American jurisprudence spill onto the character of what is helpful to their closure. Where jurisprudential agendas are in play the information that advances them may be highly incomplete and may be known by officials (and sometimes by jurors) to be contradicted by evidence that will never be (officially) heard. Information of the first sort may be of very high levels of agenda relevance, where that of the second sort reach standards of corresponding unhelpfulness.[2]

Proposition 8.17 (Non-epistemic relevance) *Information that advances agendas embedding non-epistemic constraints may exhibit up to high degrees of non-epistemic agenda relevance.*

Corollary 8.17(a) *Proposition 8.17 reflects the fact that the law is an epistemic enterprise governed in part by non-epistemic constraints.*

We remark in passing that the epistemic divide that we claim to exist between science and the law is somewhat idealized. There is sizeable literature on the sociology of knowledge and related subjects which investigates the structures of knowledge-producing organizations and the extent to which they, too, are not only subject to non-epistemic constraints, but are so in ways that may eventuate in epistemically subpar outputs. Although it is easy enough to draw from these observations horribly overblown sceptical conclusions about the epistemic purity of science, there isn't the slightest doubt that these constraints are constantly at work in even the best science, as well as everyday belief-maintenance. It is a reaction that seizes upon the fact that not everything one holds can be subject to challenge at once. To do so is psychologically and economically impossible. Any sensible doctrine of fallibilism, whether in science or everyday life, carries a disposition to reconsider any belief currently held, at least in principle. But it would be fallibilism run amok if the operations of challenge and review were exercised either with undue reach or without regard to the presumed epistemic dependencies among what is currently held for true [Kuhn, 1962]. Thus by fallibilism's own lights, both science and everyday beliefs are, at any given turn, held on sufferance and in the absence of efforts to reconfirm. Since this pattern of re-investigative restraint is applied in the absence of prior epistemic assurance, but rather more on whether the project is affordable and whether there exist particular considerations which call it into question, there is a sense in which re-investigation decisions are taken non-epistemically. But, *faute de mieux*, this is nothing in which assertions of wild scientific relativism can justifiably be rooted.

The point of the present digression has been to draw attention to an important fact about agenda relevance. In the theory of Gabbay and Woods [2003a], relevance is defined for agents in regard to their cognitive agendas. There is a highly

[2] Not all testimony of this second kind is excluded in criminal trials (if it were, its agenda relevance would always be nil). Sometimes testimony will be allowed but judges will instruct juries to give it little weight. Such testimony may be helpful, but not much.

attenuated sense in which this was so. Cognitive agendas are understood to be agendas that advance or close by way of information processed by the agenda's owner. In so latitudinarian a conception of the cognitive, we see ample room for the operation of either non-epistemic considerations or factors of subpar epistemic standing. (Let us not forget that abduction itself is reasoning in epistemically subpar conditions). One way in which the non-epistemic may enter into the closure of a cognitive agenda is by being attended by an agenda that can be advanced by the agent's cognitive wherewithal even though the desired endpoint is not itself an epistemic state (as is the case with some *decisional* agendas, for example). The other standard way is one in which an agent's targeted endpoint is an epistemic state, but subagendas of the process have non-epistemic (again, e.g., decisional targets). In any event, what the cases recently discussed point out is that non-epistemic endpoints, such as the toleration of epistemically wrongful acquitted, can be abetted by processes that are generically cognitive in the wide information-theoretic sense of the term. This will be a fact reflected in the structure of agenda relevance itself.

8.5 Legal Presumption

In the previous chapter, we said something about the linkages between presumption and plausibility. Rescher's Fundamental Rule has it that presumptive reasoning will always (typically?) defer to the most plausible of the available alternatives. We also had occasion to remark that a natural habitat for presumptive inference is common knowledge, and we directed particular attention to the instantiation of non-universal generalities, i.e., those propositions, such as the generic or the normalic, that don't embed universal quantifications. This gives us a way of preserving the Rule's principal insight. Consider a case. What is common knowledge for Harry includes the normalic claim that birds fly and the generic claim that crows fly. On becoming aware that Jasper is a crow, he infers the default that Jasper flies from the generic proposition that crows fly, rather than the normalic proposition that birds fly. We can see that the inference that Jasper flies is safer when made from the generic claim rather than the normalic. This might lead us to suppose that "Crows fly" is more plausible than "Birds fly". But there is reason to doubt it. For one thing, both these non-universal generalizations are known to be true, and in our scheme of things, knowing that P is true precludes a finding of plausibility for it. A further consideration is that differences between these two claims lies in the structure of the generalities that they attribute. "Birds fly" expresses the truth that for the most part, or usually, birds are flyers. "Crows fly" asserts that it is characteristic of crows that they are fliers. It is not tied up with what it is to be a bird to be a flier; rather, birds *usually* are the sorts of thing that fly. Crows are different. Crows fly. It *is* tied up with what it is to be a crow to be a flier. Of course, sometimes birds don't fly. Penguins don't fly — any of them. Sometimes crows won't be able to fly either. This happens when it is characteristic of crows to fly, but this crow is contingently disabled by accident, illness or genetic defect. It bears repeating that, in being known to be true, these respectively different non-universal generalities possess an epistemic standing incompatible with plausibility. (Reliability, however, is another matter.) So if plausibility is indeed a factor here, we must find a different place for it to operate. Could this be the default that is inferrable from each of the generalities at hand? No. It is the same proposition in each case. Why would it be the case that "Jasper flies" has a greater degree of *propositional* plausibility when inferred from "Crows fly" than when inferred from "Birds fly"? That Jasper flies is not itself the natural home of these intuitive differences in plausibility value. Better that we repose the difference in the inference itself, i.e., in the interpretation we give to the ∴-operator. It is the *inference* of "Jasper flies" that is more plausible when drawn from a generic truth rather than a normalic truth.

This echoes a point we made in passing against Walton's account of plausibility. We said that it overlooks the intuitive difference between inferring *from* something plausible and inferring *to* something possible. As we now see, there is a third way in which plausibility enters the structure of ampliative reasoning. Plausibility is also a marker of conclusional force. We now find ourselves at a juncture at which it would be advisable to try to determine whether we have been following our own advice. We have said quite a bit about what might be called *instantial defaulting*, in which one draws an instance of a non-universal generality. This very large class can easily be said to capture two parts of our threefold distinction about presumption. In inferring "Jasper flies" from "Crows fly" (or "Birds fly") one presumes that Jasper flies precisely because of the organization of the inference's plausibility-structure. Although the premisses in each case are true, the strongest inference they will support is a plausible inference. It is easy to see why such an inference would also be called presumptive. Equally, a plausible inference from a true generality confers nothing but propositional plausibility upon the inferred default, in the absence of knowing better. This confirms the point that, in plausible inference, conclusions quite routinely have lower epistemic values than premisses.

We also see that when an instantial default is presumptively inferred from a true non-universal generality, a judgment of presumptiveness may also be made of the conclusion itself. We presume that Jasper is a bird (in the absence of knowing better). This covers two cases — inferring plausibly, and inferring to a plausibility. There is more to inference to a plausibility than instantial defaulting. The inferential flow goes in the other direction as well. It is typified by hasty generalization, about which something was said in chapter 2. Suffice it here to say that, prior to confirmation, hasty generalization is an inference from a sample to a generalization of one or other of our three types: universally quantified conditionals, generic propositions, and normalic propositions. The plausibility of the generalization varies inversely with the strength of the general premiss. Equally, the propositional plausibility that such inferences confer varies inversely with the strength of the generality inferred.

This leaves our third case to consider. It need not detain us long to get the basic picture. Again, consider a case. On the basis of a sample, Harry hastily generalizes to, say, the generic proposition that Fs are G. His sample is that Hortense is an F. His grasp on "Fs are G" is presumptive — giving it a certain degree of propositional plausibility. His other premiss — that Hortense is an F we may assume he knows to be true. Harry infers from these premisses that Hortense is also a G. Harry's inference has a plausibility compounded, and depressed, by the fact that it is no better than a plausible inferences from premisses one of which is itself only plausible. This is the situation for as long as it remains the case that the best that Harry can say for "Fs are G" is that *these* Fs are all G. The diminished plausibility is passed on. That Hortense is G, in the absence of knowing better, can have no greater propositional plausibility that attaches to "Fs are G or to the inference from it to "Hortense is G.

What this reprise clearly confirms is Rescher's insight that the plausible and the presumptive are intimately connected. It also shows us that

Proposition 8.18 (Undetermination of the presumption rule) *Whether in its original form or in the extension of it provided by Proposition 7.11, Rescher's Fundamental Rule of Presumption significantly understates the tie between the plausible and the presumptive.*

8.5.1 Types of Presumption

It is time that we trained these results on the role of presumption in the law. In the Anglo-American tradition, there are two main loci of the idea of presumption [Uglow, 1997, p. 686–702] and [Dennis, 1999, p. 387–391]. One is the doctrine of legal presumption. The other is the doctrine of the reasonable man (or as is now more commonly said, the reasonable person). In the latter, important as it is, the factors of presumption are rarely expounded, never mind given theoretical articulation.[3] It will be enough for our purposes here to schematize the reasonable person theory in the following way. The finders of fact in a trial are required to form their beliefs and draw their inferences on the basis of what in the same circumstances a randomly selected ordinary person would believe and infer, using only those cognitive resources intrinsic to such rationality as he possesses as an

[3] The reasonable person doctrine itself — not just how it links to presumption — is largely ignored in the standard legal texts. See, for example, [MacCormick, 1994; Hannibal, 2002] whose indexes contain no mention of it.

ordinary person. The ordinary person here is someone who is untutored both in the law and in the technicalities of the issue he is required to judge. (Think, for example, of the highly complex cases of fraud that attracted such attention to the early 2000s.) It is generally assumed that the best way to determine how the ordinary person would operate is to be such a person oneself. It is for this reason that person's with expert knowledge about the issues before the court are disqualified from jury duty. Thus it is anticipated that a juror will derive some reassurance in what he believes and infers from the fact that he is — for the purposes of the trial — an ordinary person, together with the fact (when it is one) that this is what he is disposed to infer. We may take it as given that the doctrine of the reasonable man is deeply presumptive and plausibilistic. This is so notwithstanding high standard of proof that that is said to attend a juror's ultimate determination. In reaching a verdict, jurors must, on the face of it, do their best to minimize the element of presumptiveness and to aim higher than even quite high propositional plausibility. But when they are concerned with the business of interpreting evidence and sizing up the credibility of witnesses, and the like, they are not held to this high standard. This anyhow is the received view.

The place in Anglo-American law in which the conception of presumption is given detailed express consideration is in the doctrine of legal presumption, the paradigmatic case of which is the presumption of innocence. It is appropriate that we pause to say a few words about this doctrine. The value in doing so lies in the fact that legal presumptions have virtually nothing to do with what the factors discussed in chapter 6 and in earlier parts of the present chapter. It is important enough to discover that legal presumptions are quite different from the general range of presumptions. It is even more telling that informal logicians and argumentation theorists are so drawn to this wholly inappropriate paradigm in attempts at crafting what they regard as general theories of the presumptive. (See here [Hansen and Kauffeld, 2005] and [Walton, 1992b].) To investigate this further, we need a distinction between the *legally* presumptive and the *standardly* presumptive.

The presumption of guilt is a position mandated by the requirement of justice. Anyone familiar with the operation of the legal system will know that the police don't bring weak cases to prosecutors, that prosecutors don't bring weak cases to trial, and that often judges won't permit weak cases to proceed. In the absence of evidence that the criminal justice system is massively corrupt and incompetent, the reasonable *standard* presumption is that the accused is guilty. The inference to this effect from generalities such as these carries a positive degree of conclusional plausibility, and the proposition that he is guilty carries a positive degree of propositional plausibility. Given the linkages we have charted between the plausible and the (standardly) presumptive, it is also necessary to say that the presumption of innocence violates the general conditions on (standardly) presumptive adequacy.

Given those general standards, the proposition that the accused is innocent is a bad presumptive bet. That is, it is to some degree or other epistemically disreputable.

Another embedded misconception is that presumptions distribute the burden of proof.[4] In fact, however, this is not true in criminal law, and it is not true outside it. The common law places the burden of proof wholly upon the prosecution. The common law mandates the presumption of innocence. It may be said that these two requirements are the cornerstone of Anglo-American justice. But even there, they bear no intrinsic link. Had the law evolved in such a way that the person protected by the presumption of innocence was himself obliged to prove the presumption on pain of losing its protection, it would still have been true that the accused entered the proceedings with that protection, and that he retained it throughout until proof had been adduced inadequate for its further retention. In the system that has actually come down to us, everything remains the same, except the proof that is designed to cancel the protection must be wrought by the accuser. Accordingly,

Proposition 8.19 (Presumption and the burden of proof) *The legal presumption of innocence carries no intrinsic favoritism as to where the burden of proof should fall — whether on the accused or the accusor.*

It is quite reasonable to say that a system of criminal law better protects against unsafe convictions if it is undergirded by these two protections, rather than by the protection afforded by the presumption of innocence alone. But this does nothing to change the fact that they are independent provisions.

The same is true of the standardly presumptive, even in those contexts in which a presumption is shared by two parties. It is frequently noted (and rightly) that if someone (Harry, say) challenges a presumption held by another party (Sarah, say), that it falls to Harry to make the case against that presumption. There are two reasons in particular for doubting this claim.

1. Suppose that Harry's move against Sarah's presumption is indeed attended by the requirement that Harry make good his case. Even so, to the extent to which this is true, Harry's burden inheres not in Sarah's presumption but rather than in Harry's *challenge*. Harry would have the same burden had Sarah *asserted* what she now presumes.

2. Even so, the burden of proof does *not* always lie with the challenger. "Presumably, Freddie was a Soviet spy", says Harry.

[4] "The discussion of presumptions is directly bound up with questions of the burden of proof" [Uglow, 1997, p. 686]. "Thus the 'presumption of innocence' is another way of stating the rule that the legal burden of proving guilt rests on the prosecution in criminal proceedings ..."[Dennis, 1999, p. 387].

> "Why would you say a thing like that?" Sarah replies. How likely is it that we would accept as Harry's next move: "Oh, no, it's up to you to show that he wasn't"?

The common law acknowledges kinds of presumption other than that of innocence. *Provisional assumptions* — sometimes also called presumptions of fact — are conclusions a juror may draw but need not, and once drawn may use as a fact unless successfully challenged by opposing counsel. The stock example is the presumption of intendedness attaching to an act which a party has been shown to have, or admits to having, committed. It is commonly said that the presumption of innocence creates a contrary proof burden for the defence [Dennis, 1999, p. 389]. The claim rests upon a confusion. The link is there, rightly enough, but it inheres not in the presumptiveness of the presumed intendedness but rather in the person to whom the intendedness is ascribed. Certainly the Crown has no interest in showing that the accused lacked the intention required to make his act a crime; it would serve only the interest of the defence to show this. But this would be so irrespective of whether the claim of intendedness were anchored in direct evidence led by the prosecution or in a provisional presumption.

Evidential presumptions — also called rebuttable presumptions of law — are conclusions a jury must draw upon proof of the basic fact in which the presumption is rooted, in the absence of contrary indications. If, for example, it can be established that a testator has executed an apparently rational will, it must be presumed, in the absence of evidence to the contrary, that he was sane when he executed it. Here, too, the burden of proving contrary indications falls to the party that invokes them. But, again, the burden inheres not in the fact that sanity was presumed, rather than established on directly led evidence, but in the fact that doing so is required by the prover's theory of the case.

Persuasive presumptions — also classified as rebuttable presumptions of law — are conclusions a juror must, in the absence of contrary indication, draw once the basic fact is proved. If, for example, a child appears during its parents' fertile years, it must be taken that the child is legitimate, except when the opposite can be established. Similarly, if it is established as a basic fact that no evidence that a person is alive has been forthcoming for a seven year period, it must be taken that the person is in fact dead. Being a rebuttable fact the onus rests with the would-be rebuttor, and has nothing intrinsic to do with its presumptive character. Finally, *conclusive presumptions* — also known as irrebutable presumptions of law — are conclusions that must be drawn upon the establishment of the basic fact. For example, it used to be the case in English law that a boy under fourteen years required the presumption that he was incapable of sexual intercourse. Until the rule was abolished by the Sexual Offences Act 1993,s. 1, this was an irrebuttable presumption. The gap between juridical relevance and *all* the going conceptions of it save one (including the law's own definition of relevance) represents a nontrivial deviation. The gap between juridical presumption and standard presumption is even wider, and may without exaggeration be said to represent the law's epistemic distortion at its most intense. The exception lies with agenda relevance, which allows for the prosecution of epistemically compromised agendas with cognitively commonplace resources.

8.5.2 The Reasonable Person

Wide as these gaps may be, and important as they surely are for any theory of practical reasoning, the distortions they give rise to are considerably mitigated in actual practice. Take the particular case of the juror, although much the same can also be said for the three other main protagonists in a criminal proceeding — the investigating officers, those who give sworn testimony, and the judge himself. Jurors have a twofold task. They must determine whether the prosection's theory of the case meets the standard of proof imposed by the criminal law. They must also interpret the evidence, weigh the credibility of whose who testify, try to reconcile testimonial conflicts, and so on. In performing these tasks the jury is *not* bound by the standard of proof borne by the prosecution in regard to the matter of the accused's guilt (nor is any other of the parties). The jury is free to reason in the ordinary way about these things and to reach decisions about the sundry details they throw up for consideration, also in the ordinary way. In vigorously contested cases, especially those based upon circumstantial evidence, it is commonplace for the prosecution's theory of the case to be a purported solution to an abduction problem, in which it is argued that the accused's guilt is the best explanation of the known facts. In like manner, the job of the jury is to try to piece together its own theory of the case, and here, too, it often happens that the theory will be an exercise in abduction. What shows this to be so is the standard definition of circumstantiality: "Direct evidence proves a fact without inference ... Circumstantial evidence is evidence from which a fact is reasonably inferred but not directly proven" [Klotter, 1992, pp. 67–68].

It is not foreclosed that, by standard epistemic standards, a jury can be wholly justified in its view of the case as the correct one. Nothing in the remit of a juror requires that he be an agnostic about the events in question until the point at which he enters his verdict. What is required is only that the juror not *decide* the case until he has heard it all. This implies a clear distinction between a juror's duties. On the one hand, he must form an understanding of the matter before him. On the other, he must judge it by the standard of guilt beyond a reasonable doubt. The two tasks are logically and procedurally disjoint. It is perfectly open to a juror to solve his own abduction problem in favour of the prosecution, but to vote to acquit in recognition that his abductive solution doesn't rise to the required juridical standard.[5] But neither must it be thought that a solution of an abduction problem can never meet the required standard of proof, owing to the intrinsically subpar epistemic factors that inhere in abductive reasoning. What shows this to be so is the sheer fact of criminal convictions based wholly upon circumstantial evidence.[6]

It is here, perhaps more than in any other context, that the requirements of ignorance are called into question. For how can it be countenanced that a solution to a criminal abduction problem could meet the high standard of proof imposed by law, if abductive solutions are epistemically subpar?

8.5.3 Reasonable Doubt

It bears on this question that the meaning of the reasonable doubt provision is not well-explained either in case law or in legal textbooks. As a prominent American textbook points out, "Reasonable doubt is a term in common use as familiar to jurors as to lawyers. As one judge has said it needs a skillful definer to make it plainer by multiplication of words ..." [Strong, 1999, p. 517]. It is sometimes supposed that it is the legal counterpart of the high standard of proof that one finds in science and mathematics, where, in all three cases, the standard is at the top of the epistemic scale. Whatever may be the case with science and mathematics, it cannot be so with convictions won on circumstantial evidence. The meaning of "beyond reasonable doubt" must preserve this fact. Cases in which a verdict of guilty is secured by circumstantial evidence are often those in which the link between evidence and verdict is understood probabilistically. There have been efforts of late to capture the structure of such reasoning in more or less stock models of Bayesian inference [Tillers and Green, 1988]. We ourselves are doubtful of the overall adequacy of this approach, even in civil cases in which the standard is "proven on a balance of provabilities". Inspection of the actual empirical record of such cases reveals the more dominant presence of abductive considerations. On the face of it, however, this cannot be right. For if it were right, we would have it that when a conviction is won on circumstantial evidence, the verdict is mired in nothing stronger than a conjecture. But surely not even the most confident conjecture of guilt meets the standard of proof beyond a reasonable doubt. Accordingly

Proposition 8.20 (The circumstantial conviction dilemma) *At first appearance, either circumstantial conviction cannot meet the required standard of proof, or it is not abductively grounded.*

[5] Strictly speaking, a juror is not required to form a theory of the case in order to accept or reject the prosecution's theory. This preserves the important point that in order to acquit, a juror need have no notion whatever about how to explain the evidence. It suffices that he reject the prosecution's explanation of it.

[6] "History is replete with examples of convictions based exclusively on circumstantial evidence" [Klotter, 1992, p. 69].

We ourselves are minded to challenge the first horn of the dilemma of Proposition 8.20. Great weight is placed against it by the doctrine of the reasonable person. In its most general sense, it requires that jurors perform as ordinary persons in the course of their reflections on the matters before them. They are then required to use this ordinary thinking to reach a verdict. Verdicts are not only open to be produced by ordinary thinking, but are *required* to be so produced, with one proviso: except when juridically constrained in some or other particular way. If this is right, a solution to the dilemma of Proposition 8.20 drops out. In the context of realistically constructed cases based on circumstantial evidence, ordinary thinking is frequently, if not typically, abductive. Since abductive thinking is inherently conjectural, not only is it left open that a verdict of guilty might be conjecturally based, but it is inevitable that this frequently, if not typically, be so. What remains is to show how conjecturally structured theories of a case manage to hit the required proof standard.

The core idea embedded in the standard makes a twofold claim on reasonability. First, the theory of the case for conviction must be such as to draw the favour of a randomly selected reasonable person. Secondly, that self-same reasonable person must also be disposed to the view that the facts of the case do not answer to a rival theory of them that could reasonably be accepted. Interpreted abductively, this requires that an abductively secured conjecture of guilt must be strongly secured, and that there is no rival conjecture that is strongly enough secured. However, as the Indiana Court of Appeals has made clear in a case from 1978, "Convictions should not be overturned simply because this court determined that the circumstances do not exclude every reasonable hypothesis of evidence" [Klotter, 1992, p. 69]. Accordingly,

Proposition 8.21 (Guilt and reasonable alternatives) *If a verdict of guilt is arrived at circumstantially it is not necessary that there not be other abductively reasonable theories of the evidence.*

For the present suggestion to pass muster, the idea of abductive strength requires clarification. To do so, it is important to emphasize that typically a conviction based on circumstantial evidence is a conviction *faute de mieux*, epistemically speaking. The qualification "typically" is necessitated by the fact that the law allows that, on occasion circumstantial evidence may be as strong or stronger than direct evidence. Also significant in an American case from 1969, "the trial court properly instructed the jury that 'the law makes no distinction between direct and circumstantial evidence but simply requires that the reasonable doubt be drawn from all of the evidence in the case,' including 'such reasonable inferences as seem justified, in the light of your own experiences'" [Klotter, 1992, p. 68]. The betterness that circumstantially based verdicts fail to achieve is the grade of epistemic attainment, whatever that is in fine,that attend conviction by direct evidence.

Thus we assume as a matter of epistemology, rather than of juridical pronouncement, that unrebutted direct evidence possesses an epistemic strength not usually possessed by circumstantial evidence in the face of competing and not unreasonable rival theories. In structural terms, let K be what the court knows of the matter before it by direct evidence. Since, by hypothesis, a conviction cannot be got from K, alone, it must be aimed for by some supplementation of K short of additional direct evidence. This constitutes an abduction problem for the prosecution. The prosecution must attempt to supplement K in ways that the contents of K itself make reasonable and without further direct evidence. The task of the juror is to determine whether the prosecutor's case is, in effect, a strong enough abduction without strong enough rivals. To achieve this standard, he must overcome the epistemic disadvantage implicit in the fact that sufficiently strong abductions won't hit the epistemic standard hit by K. Accordingly we shall say

Proposition 8.22 (Discounting epistemic disadvantage) *A successful abduction for conviction is one that is strong enough to minimize the epistemic disadvantage that inheres in abductive solutions. Correspondingly, a rival abduction is insufficiently strong when it does not minimize the inherent epistemic advantage to a sufficient degree.*

Corollary 8.22(a) *Implicit in the doctrine of the reasonable person is the principle that sometimes it would be unreasonable not to accept an abduction, or to accept it weakly, just because it failed to hit the epistemic standards reached by K.*

What we are here proposing is an epistemic commonplace. It is the idea that epistemic satisfaction is not only not typically achieved by epistemic optimization, but that, for large classes of cases, postponing epistemic satisfaction until greater strides toward optimization are achieved would be decidedly unreasonable. In the absence of contrary indications, you know that you are your parents' child if you arrived during the child-bearing years of their union. In the absence of contrary indications or some contextually required standard of proof, resort to DNA testing would be quite mad. The criminal law requires that those of its obligations that fall to jurors be discharged by persons who operate as ordinary thinkers. The criminal law requires that the epistemic endeavours of jurors rise to the standards of the epistemically ordinary person. The requirement of determinacy whether, in its turn, the prosecution's theory of the case achieves law's standard of proof is thus a requirement that a reasonable person can be expected to attain when operating as an ordinary thinker. What the criminal law clearly settles for is not optimization, but satisfization set against sufficiently high standards. In the case of circumstantially based conviction, what the criminal law clearly settles for as well is an abductive solution which an epistemic satisficer who knows the relevant standard of proof would confidently accept and whose acceptance would not be in any way troubled by the express recognition that this judgement did not rise to the epistemic standard of K. The juror has discharged his ultimate obligation if he finds himself in the role of the epistemic satisficer whose standards do not in this particular way rise to K's level.

8.6 Hypothesis-Discharge

It is widely assumed in the literature that discharging an hypothesis is intimately bound up with its subsequent experimental confirmation or other forms of validation. So understood, hypothesis-discharge is *post-abductive*. It lies in the confirmatory aftermath of a decision to send a proposition to trial. We have already had occasion to observe that a decision to send a proposition to trial is neither necessary or sufficient for hypothesis-engagement; hence is not intrinsically an abductive determination. Neither are favourable trial-outcomes necessary nor sufficient to that part of hypothesis-discharge that does remain fully within the ambit of abduction. This not to overlook the relative frequency with which in non-legal settings, an abductive conjecture is sent straight to trial; nor is it overlook that one way of shearing off a proposition's conjectural character is by demonstrating its truth experimentally or in some other way. Let us be clear in saying that, while confirmation of a proposition is sufficient for the cancellation of its conjectural mode of presentation, it is not part of the process of abduction. Accordingly, when a conjecture is sent straight to trial, abduction ends at that point. We remarked in chapter 4 that an abducer might reflect in his *selection* of a hypothesis his optimism that it would do well at trial, but, as we noted, thinking that one's hypothesis will do well at trial is not intrinsically tied to its selection. But even if it were, the trial itself would still be post-abductive.

This leaves the question of whether hypothesis-discharge is possible within abductive contexts and, if so, what its structure would be. As we have it so far, hypothesis-discharge is achieved by an inference to a H^c. H^c reflects a readiness to release H, on sufferance, for premissory work in future inferences. How does this hook up with what juries do?

The answer lies in what we have already discovered about the operation of the provisions of the beyond-reasonable-doubt standard for circumstantial criminal conviction. We summarize the main points of that finding.

1. A verdict in a criminal trial is not a conjecture. It is a *finding*; hence something that is forwarded assertively.

2. Even so, especially in cases built upon circumstantial evidence, verdicts are reached abductively. They are solutions of abduction problems.

3. The standard of proof beyond a reasonable doubt in effect requires a jury to discharge its theory of the case, that is, to forward it non-conjecturally. This

resembles what abducers in general achieve by forwarding H^c assertively.

4. Since, in such cases, there is no independent means of demonstrating directly the truth of a jury's finding, the jury's discharge of the hypothesis cannot be seen as post-abductive.

5. Accordingly, in reaching its finding in such cases, hypothesis-discharge is part of the jury's solution of its abduction problem.

This allows us to say that

Proposition 8.23 (Discharge) *Conditions on abductive hypothesis-discharge approximate to those governing circumstantial conviction in a criminal trial.*

Accordingly, it may be said that when a jury reaches its verdict, they have done something like draw an inference to $C(H)$ and a decision in the form H^c. "$C(H)$" expresses the jury's conviction that, although the evidence is only circumstantial, it may be taken with requisite confidence that the accused's guilt best explains it. In turn, H^c releases the verdict, "Jones is guilty", for work as a premiss in future inferences or decisions. For one thing "Jones is guilty" is a primary datum for subsequent decisions about sentencing. And thereafter, it states a legal fact. But here, too, it is a fact on sufferance, i.e. in the absence of an appeal that would eradicate it.

In our discussions so far, we have plotted the fortunes in their legal settings of three concepts of central importance to a logic of abduction — plausibility, presumption and relevance. When compared with how they fare in standard or non-legal contexts, an important methodological lesson presses for a hearing.

Proposition 8.24 (The distortions of law) *Given epistemic intrusions required by justice, it may be taken as a rule of thumb that cognitive concepts are not well-elucidated by the treatments they receive in legal contexts.*

Corollary 8.24(a) *Theorists who seek for satisfactory* general *explanations of cognitively oriented concepts, such as plausibility, presumption and relevance, should not expect to find them in theoretical jurisprudence.*

8.7 The Probativity Question

In chapter 4, we drew attention to a still unresolved contention among philosophers of science about the probativity of explanation. We pointed out that there is a considerable body of opinion an opinion shared by the present authors that the explanatory force of a proposition is not in the general case a satisfactory marker for its truth. At first sight, this is disastrous for the abductive theory of criminal conviction. For if the fact that the hypothesis of guilty as charged is indeed the best explanation of the evidence led at trial is a fact that is compatible with the *falsity* of that hypothesis, surely we are deluding ourselves in thinking that the common law offers to accused persons the safety of a fair trial, at least for the most part.

How shall we answer this objection? Perhaps this is the best place to drive home the point that the common law's criminal justice system does not offer accused persons epistemic guarantees. Another — and somewhat jolting — way of saying this is that the criminal justice system squarely faces accused persons with the prospect of outcomes that are not *known* to be true.[7] (If this doesn't drive a stake, once for all, through the heart of the common belief that guilty verdicts attain an unusually high standard of proof, nothing will.) Accordingly,

Proposition 8.25 (The fundamental epistemic fact) *The fundamental epistemic fact about criminal convictions is that they constitute verdicts that need not be known to be true (and in general are not known to be true) in order to qualify as both just and cognitively scrupulous.*

Proposition 8.25 bears on the structure of abduction itself. Suppose, contrary to what the present authors believe, that best explanations are probative. That is, suppose that best explanations are truth-conferring. Then it is easy to see that an inference to the best explanation cannot be a case of abduction. Abductive inference is ignorance-preserving; but (on the present assumption) best-explanation inferences are truth-conferring. So best- explanation abductions don't preserve the ignorance condition on abduction. Accordingly,

Proposition 8.26 (Non-probativity) *If theories of the evidence are best-explanation abductions, explanations are not truth-conferring.*

Corollary 8.26(a) *Corollary 31(a) By the fundamental epistemic fact (Proposition 8.25), best- explanation inferences are not truth-conferring in judicial settings.*

8.8 Revision Structures

A decision to discharge an abductively successful hypothesis H bears on the parameter $K(H)$ in two ways, one of which is well-recognized in all the standard treatments, and the other of which often goes unnoticed. In both the AKM- and GW-schemata, the inference to H^c presupposes that $K(H)$ stands in the right kind of relation to a target T. $K(H)$ is a revision of a knowledge-base K. $K(H)$ arises from K by addition of the assumption of H's truth, together with whatever adjustments are necessitated by the presence of H in $K(H)$. There are no uniform fixed requirements on K-revision, apart from its bearing on $K(H)$ in such a way

[7] This proceeds not only from the abductive character of verdicts but also from the admissibility of testimony.

that it may be inferred from this that the abducer's target T has been presumptively attained. Consider again the issue of consistency. If the abduction in question is consequentialist and if $\rightsquigarrow$ is classical consequence, it is clear that $K(H)$ must be consistent; but, as we have seen, it is inadvisable to make consistency a strictly necessary condition for abduction as such.

A further condition on the selection of H is that $K(H)$ not generate consequences that the abducer is not prepared to commit to. That this is not represented in either the AKM-schema or the GW-schema is a significant mission. Accordingly,

Proposition 8.27 (Unwanted consequences) *It is necessary to revise the abduction schema so as to reflect the requirement that even though $K(H)$ may presumptively attain T, it is not justified to infer $C(H)$ or H^c if $K(H)$ also carries consequences that the abducer is not prepared to accept.*

The second way in which a winning H involves $K(H)$ is at the point at which H is discharged. As we have emphasized, a decision to discharge H is a decision to release it for further premissory work, subject to the requirement that its conjectural origins be duly noted. It is fair question as to how roving a premissory role such an H might be assumed to have. It would appear that there is no wholly general answer to this question. Even so, it can be said with confidence that *at a minimum*, the domain in which H^c now functions as a premiss (or a datum) is $K(H^c)$ itself. It is well to note that $K(H^c)$ is just like $K(H)$ except that $K(H^c)$ is also discharged. Consider again the quantum hypothesis Q. In invoking it, Planck didn't merely add a new hypothesis to the physics of 1900; rather in adding it, he radically changed the character of physics.[8] As long as Q lacked for empirical confirmation, the new physics $K(Q^c)^c$ was itself conjectural. There are two points to be clear about. Let loose a conjecture might radically and massively recognize a given knowledge-base; and it will do this conjecturally for so long as the added hypothesis is itself conjectural. Accordingly,

Proposition 8.28 (Closure) *Conjecturality is closed under the revision of K to $K(H^c)$. (We reflect this fact by denoting K-versions by the expression $K(H^c)^c$).*

We may take it that, whenever H is an abductively successful hypothesis with regard to $\{K, T\}$ ignorance problem, there exists a revision structure $K(H^c)^c$ for K. One of the conditions that $K(H)$ had to meet was that it bore $\rightsquigarrow$ to a payoff V. A further requirement is that $K(H)$ have no unacceptable consequences. $K(H^c)^c$

[8] Another example of the *wide effect* of even a *narrow adjustment* is the Anderson-Belnap approach to relevant logic. Originating in the downgrading of Disjunctive Syllogism from the status of a valid rule to that of an admissible rule, the wide effect of this was the loss of extensionality. Nothing in the historical record indicates that the repudiators of Disjunctive Syllogism had any conception of the radicality of their proposal.

invokes these same requirements. So a minimal constraint on the composition of $K(H^c)^c$ is that its members severally or collectively comply with these two requirements. It is well known that knowledge-sets of even fairly low finitude cannot, by beings like us, be checked even for truth-functional consistency. The same is true of their closures under consequence. This in turn requires us to say

Proposition 8.29 (Constructing revision structures) *Notwithstanding that for any winning H a revision structure $K(H^c)^c$ exists, it is not in general possible for abductive agents to construct such structures.*

Proposition 8.30 (Adverse consequences) *It is not in general possible for the individual abducer to verify that his $K(H^c)$ is free of adverse consequences.*

Corollary 8.30(a) *The parallel between revision structures and filtration structures is also evident.*

8.8.1 Proof Standards

An exception to Proposition 8.24 and its Corollary would appear to be the matter of intra-abductive hypothesis-discharge. It is a welcome exception. It helps correct a considerable misconception about the standard of proof in criminal cases. This is the idea that the standard is artificially high. In fact, it is not artificial, and it is not especially high — certainly it is no kin of mathematical proof or experimental confirmation of the sort required in drug trials. It is perfectly true that, in the name of justice, the law artificially constrains what evidence a jury can hear and, at times, the weight that a jury can give it; but this same artificiality is not intruded into the standard of proof itself. What shows this to be so is the commonplaceness of the constraints under which the standard is honoured in actual judicial practice. Key to a proper understanding of them is the idea of *satisfaction*. (See here, [Woods, 2005a]). What the law requires is that jurors attain a certain level of doxastic satisfaction. They must be satisfied that the picture that the evidence suggests to them is undisturbed by the fact that it is not an epistemically optimal theory of the case. The other is that the failure of a rival theory of the case to satisfy them is not something that counts against it in an epistemically optimal way. But this is the condition in which the epistemic satisficer finds himself quite routinely. It is the hallmark of the reasoning of an ordinary reasoner when reasoning in the way of ordinary reasoners about just about anything. What counts, both in the general case and in the case of proof beyond a reasonable doubt, is that these occasions of possible error do not disturb the reasoner's doxastic repose. (The language of the law is replete with the idioms satisfaction and repose. Judges tell juries that, to convict, they must be *satisfied* that such-and-such and so-and-so. When counsel have presented their case, they *rest*.) In this model of juridical determination, it is difficult to over-estimate the pivotal importance of satisfaction. Satisfaction is the dual of cognitive irritation, which is what occasions the need for abductive reasoning in the first place. Accordingly,

Proposition 8.31 (Doubt and satisfaction) *A jury's verdict meets the standard of proof beyond a reasonable doubt when its members are in a state of doxastic satisfaction achieved by the procedures of ordinary reasoning in response to the evidence led at trial.*

Proposition 8.32 (Competence) *The present model of cognitive satisfaction presuppose the competence of individual jurors; in particular that the satisfaction required by the standard would not be achieved by a competent reasoner unless he were untroubled by the fact that his theory of the case did not attain standards of epistemic optimality and by the fact that his exclusion of rival theories did not attain it either.*

The key to hypothesis-discharge lies in the structure of the abducer's doxastic satisfaction. When a proposition is held conjecturally, what the reasoning agent is satisfied about is that it is a proposition that merits conjecture. When a proposition is abductively discharged, what the reasoning agent is satisfied with is *it*. He is satisfied with its propositional content. A reasoner moves from $C(H)$ to H^c when he moves from the first kind of satisfaction to the second.

We have seen that jurisprudential contexts occasion significant distortions of most concepts of relevance and all standard conceptions of presumptiveness. This is a reflection of the epistemic compromise that justice negotiates with truth. It arises from the law's fundamental operating principle that epistemically wrongful convictions should be minimized even at the cost of epistemically wrongful acquittals. These, we say, are significant distortions, but they are significantly redressed by the circumstance that in achieving even the high standard of proof required for a criminal conviction, the juror's reasoning, step by step, need not — and should not — aim at or attain a standard higher than the standard achieved by a reasonable person when reasoning as an ordinary being; i.e., including the drawing of "such inferences as seem justified, in the light of [his] own experience" [Klotter, 1992, p. 68]. This places the phenomena of circumstantial conviction in the spotlight, and gives us a point worth repeating. It gives us occasion to provide an interpretation of proof beyond a reasonable doubt according to which the juror is an abductive satisficer concerning the verdict he proposes, whose confidence in it is not shaken by his recognition that his own solution does not optimize to the level of K or higher, and for whom there is no rival abduction that could appeal to his obligations as a satisficer. We have it, then, that

Proposition 8.33 (Beyond reasonable doubt) *The judicial question of proof beyond a reasonable doubt has a solution in the logic of abduction.*

8.8.2 Analogy

In one of the epigraphs to the present chapter, Bas van Fraassen proposes that

Proposition 8.34 (The principle of symmetry) *Structurally similar problems must receive correspondingly similar solutions [van Fraassen, 1980].*

The Principle of Symmetry also embeds an important insight about analogical reasoning. We have already made much of the point that hypothesis-engagement presents the would-be abducer with what is arguably his most difficult task, notwithstanding the frequency, speed and apparent ease with which such tasks are performed. The phenomenon of cut-to-the-chase abduction attests to this point in an especially vivid way. For an abduction problem, the space of possible hypotheses is up to arbitrarily large. If the abduction problem has a solution, it will be true in the general case that the winning hypothesis is an element in this space. Yet for what can easily be seen as the large majority of satisfactory abductions, the abducer makes his selection without searching the space in which it resides. As we have said, there are two different explanations of this appearance. One is that the appearance is also the reality; i.e., that hypothesis-engagement is routinely achieved without such searches. The other is that these searches are actually made, but in ways that leave no behavioral or introspective trace; i.e., they are made tacitly, using the connectionist wherewithal of cognition down below. In preceding sections of this chapter, we have considered the highly plausible suggestion that what facilitates the process of engagement without search is the human reasoner's striking capacity for evading irrelevancies, a trait that Harman re-creates as a supposed rule:

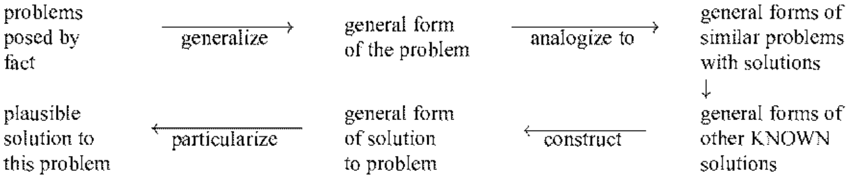
Proposition 8.35 (The clutter avoidance principle) *Do not clutter up your mind with trivialities [Harman, 1986].*

Much the same can be said of our facility with analogies. The ease with which we recognize analogies facilitates the process of engagement without search. The link between the two traits is reflected in turn in what might be said to be the

Proposition 8.36 (The fundamental principle on analogies) *The greater the relevant similarity, the stronger the analogy. The less the relevant similarity, the weaker the analogy.*

Both Proposition 8.32 and Proposition 8.33 carry essential qualifications. Without the qualification "structural", the Principle of Symmetry becomes highly dubious. Without the qualification "relevant", the Fundamental Principle on Analogies likewise fails to convince. It is worth pointing out that although both principles can claim legions of supporters, there is virtually nothing in the literature to guide us in the interpretation of these qualifications. This is not something to make light of, but neither should it alarm us unduly. If we take the relevance constraint in the manner of agenda relevance, it produces the truism that the similarities that make for strong analogies are those that help us analogize strongly. The circularity is benign, but otherwise unhelpful. If, on the other hand, we understand the constraint in any of the other senses of relevance canvassed in this chapter — topical relevance, full-use relevance, irredundancy-relevance and probabilistic relevance — we are left with a suggestion that is far from useless. The suggestion is that the similarities in which analogies are grounded are those that stand to one another in nontrivial semantic relations.

Our task in what remains of this chapter is to chart the role of analogy in abduction. We leave it as an exercise to describe with greater precision the interconnections between analogy and relevance. We begin by revisiting two of the schemas for abduction briefly met with in chapter 3. Darden proposes 'the following general schema for a pattern of reasoning in hypothesis-construction: *Darden's Schema* [Darden, 1976, p. 142]


problems posed by fact
generalize
general form of the problem
analogize to
general forms of similar problems with solutions
general forms of other KNOWN solutions
construct
general form of solution to problem
particularize
plausible solution to this problem

Darden's Schema reflects Hanson's influence, as evidenced by the following summary

Hanson's Schema [Hanson, 1961, p. 33].

1. Some surprising, astonishing $p_1, P_2, p_3, \ldots$ are encountered.

2. But $p_1, p_2, p_3, \ldots$ would *not* be surprising where a hypothesis of H's type be obtained. They would follow as a matter of course from something like H and would be *explained* by it.

3. Therefore there is good reason for elaborating a hypothesis of the type of H; for proposing it as a possible hypothesis from whose assumption $p_1, p_2, p_3, \ldots$ might be explained.

Darden bids of the abducer to analogize. Hanson bids him to reason from types. There is a ready-built theory of analogical argument that gives instruction on both these matters. It is the Meta-Argument Theory of Analogical Argument (MATAA), to which we now turn [Woods and Hudak, 1989].

8.8.3 The Meta Approach

Consider now what may have been one of the most discussed analogical arguments of the century just past [Thomson, 1971, p. 49].

> You wake up in the morning and find yourself back to back in bed with an unconscious violinist. A famous unconscious violinist. He has been found to have a fatal kidney ailment, and the Society of Music Lovers has canvassed all the available records and found that you alone have the right blood type to help. They have therefore kidnapped you, and last night the violinist's circulatory system was plugged into yours, so that your kidney can be used to extract poisons from his blood as well as your own. The director of the hospital now tells you, "Look, we're sorry the Society of Music Lovers did this to you — we would never have permitted it if we had known. But still, they did and the violinist now is plugged into you. To unplug you would be to kill him. But never mind, it's only for nine months. By then he will have recovered from his ailment and safely be unplugged from you." Is it morally incumbent on you to accede to violation? No doubt it would be very nice if you did, a great kindness. But do you *have* to accede to it? What if it were not nine months, but nine years? Or longer still? What if the director of the hospital says, "Tough luck, I agree, but you've now got to stay in bed with the violinist plugged into you, for the rest of your life. Because remember this: All persons have a right to life, and violinists are persons. Granted you have a right to decide what happens in and to your body, but a person's right to life outweighs your right to decide what happens in and to your body. So you cannot ever be unplugged from him"

Thomson develops her analogical argument as follows. If you judge *The Violinist* to be a good argument, then there is another argument similar to The Violinist that, by parity of reasoning, you must also judge good. This other argument (call it *The Pregnancy*) is one that concludes that maintaining a pregnancy from rape is not morally obligatory. Thomson is here invoking, without naming it, the Fundamental Principle of Analogy, which requires that relevantly similar cases be treated similarly.

What the Fundamental Rule does not make clear is what the relevant similarities between *The Violinist* and *The Pregnancy* would consist in. This is answered by the Meta-Argument Theory of Analogical Argument. It would be well to bear in mind at this juncture that Darden's schema has the reasoner *generalizing* and Hanson's schema has the reasoner arguing from *type*. So consider the following schema, which we'll call *The Generalization*.

The Generalization

Human beings H_1 and H_2 are so related that without H_2's consent, H_1 has placed H_2 in a state of vital dependency;

1. the period of dependency is indeterminate (perhaps nine months, perhaps nine years, perhaps forever).

2. the dependency is a grievous impediment both to locomotion and to (stationary) mobility;

3. the dependency constitutes a grievous invasion of privacy;

4. it is an invitation to social disaster, for H_2 (and H_1 as well) is a laughing stock,

5. it threatens H_2's economic self-sufficiency;

6. therefore, it would be morally permissible for H_2 to terminate the vital dependency.

Suppose, for present purposes, that Thomson is right in claiming parity between *The Violinist* and *The Pregnancy*; then *MATAA* furnishes the answer to what it is that these two quite different arguments nevertheless have in common. The *MATAA* proposal takes seriously the mention of *cases* in the Fundamental Rule to treat similar cases similarly. Accordingly, it is proposed that what *The Violinist* and *The Pregnancy* are cases *of* is a common deep structure represented here by *The Generalization.* Thus an analogical argument is a meta-argument; it is an argument about arguments. It is an argument in the form

Meta-Argument

1. Argument A possesses a deep structure that provides that the premisses of A bear relation R to its conclusion.

2. Argument B shares with A the same deep structure.

3. Therefore, B possesses a deep structure that provides that its premisses likewise bear R to its conclusion.

4. Hence, B is an analogue of A, A and B alike are good or bad arguments, by parity of reasoning, so-called.

In our example, argument A is *The Violinist. R* is a strong consequence relation. B is *The Pregnancy.* The deep structure is *The Generalization.*

The *MATAA* analysis requires that the analogizer generalize on an argument whose assessment is already settled, and then instantiate to a different argument. To achieve the desired parity, it is essential that the property that the original argument is assessed as having (e.g., validity) is preserved by the generalization and preserved by the subsequent instantiation (Darden calls this "particularization"). It is well to note the significance of target property-preservation. For ease of exposition let us agree to denote by P that property of *The Violinist* in virtue of which we regard it as a successful argument (assuming that we do). Then if the move from *The Violinist* to *The Generalization* is P-preserving and the instantiation from it, in turn, to *The Pregnancy* is also P-preserving, then one cannot in strict consistency ascribe P to *The Violinist* but not to *The Pregnancy*. This is important. It shows that good *MATAA* reasoning makes it a requirement of consistency that *The Pregnancy* be given the identical evaluation as *The Violinist*. But let us also note in passing that this is not at all the same as holding that one cannot in strict consistency accept the *premisses* of *The Pregnancy* and withhold or reject its conclusion. For nothing requires that the P that gets preserved in successful *MATAA* reasoning is always the property of validity. It is easy to see that the *MATAA* model incorporates what Paul Bartha calls the *common core* of good analogical arguments.

> That common core ... is captured by two simple and fundamental principles: *prior association* and *potential for generalization* [Bartha, forthcoming, sec. 7.1].

How might this bear on abduction? In a typical consequentialist case the abducer has a target T for which he seeks an H and V such that $K(H) \rightsquigarrow V$. His principal task is to find the requisite H. By Darden's and Hanson's lights, the abducer reasons his way to H by analogy. Doing so involves generalization and instantiation, and reasoning from type. If he proceeds in the manner prescribed by *MATAA*, the consequentialist abducer looks for a distinct consequence structure

1. $K^*(H^*) \rightsquigarrow V^*$
   that generalizes to

2. $K'(H') \rightsquigarrow V'$

in ways that preserve the truth of (1). A condition on this generalization is that the payoff V of the abducer's present abduction problem instantiate the V' of the generalization of the original consequence structure. He then instantiates from K' and H' to the required K and H. What this represents is a rather commonplace situation in which there is a conclusion the abducer wants to find some justification for. This is a premiss-selection task. He looks for a conclusion whose sole similarity to his intended conclusion is that it is justified by premisses of a type (Hanson) that would likewise justify the intended conclusion. So he looks for the requisite similarities among putative premisses.

Except for tightly circumscribed situations, there are no known algorithms for this procedure that abducers actually execute. On the other hand, the *MATAA* has some attractive advantages. One is that it reduces analogizing to target property-preserving generalization and instantiation, which, given their commonness is a significantly simplifying reduction. The other that it gives some content to the schematic insights of Darden and Hanson. Even so, *MATAA* is not problem-free. One of its difficulties is its narrowness. *MATAA* reasoning always has as its target some or other proposition which the abducer wants to *establish*. Thus, the *MATAA* approach will fail for any abductive target for whose attainment an *argument* is not required. It may well be, however that *MATAA* admits of adaptations which would make it an appropriate kind of approach for other targets, such as explaining a proposition, or simplifying a body of knowledge. If such adaptations were possible, they would all pivot on P-preserving generalization and instantiation, where P is whatever virtue the abducer wants to hold fast to in the attainment of his target. (Intuitively, we can think of P as varying over such standards as *strong explanation*, or *accurate prediction*, or *effective simplification*.) It is in this factor of P-preservation that a second problem arises. It is that when a target is hit on the basis of *MATAA* reasoning or by reasoning adapted from it in the ways we are presently considering, the reasoner's problem may be solved, but it is not solved in ways that honour the nescience-condition.

Proposition 8.37 (When analogizing is not abductive) *Analogical reasoning cannot be abductive on any account of analogy that turns on P-preserving generalization and instantiation.*

The truth of Proposition 8.34 is perhaps best exemplified by returning to our *MATAA*-reconstruction of Thomson's argument. Here the problem is to establish that there is no moral obligation to persist with pregnancies arising from rape. We may take it that the reasoner's knowledge-set K contains a proof that there is no moral obligation not to unplug the violinist. The analogy has two further components, each of which may reasonably be supposed not to be in K (or not expressly in K) at this state of the resolution process. But the process cannot terminate satisfactorily unless the reasoner, in effect, updates his K by inclusion of the component we've been calling the *Generalization*. Given what he then knows, the problem is solved if he infers from it The Pregnancy and does so with the same degree of cognitive virtue possessed by the original argument, The Violinist. In sum, the analogizer hits his target by *expanding his knowledge-set*. But this is not abduction. Not only is the ignorance requirement not met, neither is there occasion for conjecture. One is not here conjecturing that *The Generalization* is a generalization of the *Violinist* and that the *The Pregnancy* is a case of it. Rather, the analogizer is as satisfied about the truth of these claims as he was in the beginning about the soundness (i.e., the P-adequacy of the *The Violinist*.

Guarantees of P-preservation violate the ignorance condition on abduction. One of the principal parts of the *MATAA* model of analogical argument is that such arguments are good because they are P-preserving. It would appear as we have said, that analogical argument has no legitimate role in abduction. But perhaps this is too abrupt a conclusion to draw. Central to the *MATAA* model is the case-of relation, expressed in extensional languages by the syntactic relation of instantiation. But the case-of relation is also a semantic relation which depends heavily on meaning relations between natural language expressions. The *MATAA* model provides no general account of this relation. Theories of meaning relations between predicates and other terms are thin on the ground and not very persuasive even when they exist. So *MATAA*'s failure to provide one is not something to be over-severe about. Even so, it is a consequential omission. One of the more important tasks of such a theory is to achieve satisfactory control of the distinction between P-preservation and approximate P-preservation, in particular, such theories can be expected to capture the difference between *being a case of* and *being similar to*. The very fact that the distinction exists, and is important, justifies our interest in having an account of it. But it also suggests a relaxation of the ban on analogical arguments in solutions to abduction problems. Here is how. Consider an abductive context in which T is the target. The would-be abducer seeks an hypotheses H which, in concert with what he now knows, will bear the $\looparrowright$-relation to some payoff V on an interpretation in which this fact suffices for the attainment of T. Suppose that included in what the consequentialist reasoner knows are the following.

1. There is a T' and a V' resembling T and V, and a proposition H such that $K(H) \looparrowright V'$ obtains in a way that suffices for the attainment of T'
2. Proposition H' resembles H and is such that $K(H') \looparrowright V$ in ways that attain T.

But the abducer neither knows nor has adequate independent evidence to support H'. So he handles H' abductively. He concludes that H' is safe but for conjecture on grounds that if it did obtain in fact, it would play an integral role in a P-approximating analogical argument.

This is an instructive turn. It demonstrates that analogical considerations can guide a successful abduction even though the abductive inference does not itself qualify as an analogical argument, even in our relaxed P-approximating sense.

As may now be apparent, our adeptness at analogizing is a particular case of the more general knack for knowing what to make of similarities. Knowing what to make of a similarity is tantamount to recognizing similarities that are relevant. In a *MATAA* — setting, the the relevance of a similarity is that thanks to which it approximates to without being the case-of relation. It is a virtue of that interpretation that it avoids the circularity that attends defining the similarities that are relevant to abductive solutions as those that advance or complete such solutions. This evacuates important claims of information-content, claims such as "Relevant similarities advance or close abductive agendas". Still, outside of such definitional contexts, relevant similarities are in fact those similarities that prove helpful in the discharge of cognitive tasks. Our facility with relevant similarity-recognition is part and parcel of our general facility for evading irrelevancies and staying on point. This, in turn, is part and parcel of our general aptitude for problem-solving. Relevant similarity is, therefore, a notion that falls rather easily into the ambit of agenda relevance.

8.8.4 Similarity

The notion of similarity is fundamental to all cognitive practice. It parallels the fundamentality of relevance recognition. Concerning the latter, we are, as we said earlier, awash in information at each moment of the exercise of our cognitive devices. Essential to their satisfactory exercise is, occasion by occasion that most of this information that befalls an agent function, on that occasion, as noise and that what is not thus filtered out plays a positive role in completing the task at hand. In like fashion, the human cognitive agent is awash in sheer difference. Just as the detection of islands of relevance in oceans of irrelevance is required for cognitive survival, so too is the timely and accurate detection of similarity in difference, one of the most basic exemplifications of which is the capacity to classify. Discerning that Jasper is a crow requires that the classifier not attend to the substantial differences between Jasper and the other crows of his acquaintance (or the prototype of crows stored in memory, if that is what it takes). Linked to this capacity, and dependent upon it, is our adeptness in recognizing natural kinds, which underlies our (approximately) good record as hasty generalizers. (See again the discussion in chapter 2). Neither is plausible to suppose that type-instantiation is in general inferential. Recognizing that Jasper is a crow or that Sarah is a woman involves attending to similarities and discounting differences. But this is not abduction unless the attributions are, in relation to the cues that prompt the classification, subpar. This is not for the most part the case.

Something cognitively as basic as similarity recognition is bound to have a role in something cognitively as basic as abduction. To the extent that analogies are similarities we take cognitive note of, analogy, too, is bound to have a role in abduction. But again we should note that although analogical considerations may well influence the outcome of an abductive exercise, analogical and abductive reasoning are separate types of cognitive practice, mutually independent both procedurally and conceptually. This lends a degree of corroboration to Darden's scheme for abduction, in which analogizing has a role, and Hanson's schema for abduction, in which reasoning from type is emphasized. But as we now see, these links to analogy and to types are contingent. They do not play at the heart of abduction as such.

8.9 Analogy in Law

In the Anglo-American common law tradition, analogical reasoning is fundamentally a matter of taking due notice of legal precedents[9] The doctrine of precedent is called *stare decisis*. In English law, which embodies the strictest version of the doctrine, precedent is governed by three main conditions [Cross and Harris, 1991, p. 5].

1. Decisions of any superior court must be respected.
2. Any decision of such a court constituents a binding precedent for that court and for any lower court.
3. Any decision of such a court constitutes a persuasive precedent for higher courts.

A precedent is binding when it requires a judge in a given case to decide it in the same way as the previous case irrespective of the merits of the case presently before him. A precedent is pervasive when a judge must honour it in the present case, unless he has sufficient reason not to.

Deciding a case in the same way as an earlier case involves the application of the *ratio decidendi* of the prior case to the present one. A *ratio decidendi* is a general principle of law, or set thereof, on the basis of which the judge reaches his decision. In actual juridical practice, discerning such principles is often far from easy and, in any event, always contextually influenced by the particular facts of the case in question [Cross and Harris, 1991; Levi, 1949]. There is a famous case in which the House of Lords found for the plaintive in an action brought against a producer of ginger beer. The charge was that the manufacturer was liable for the plaintiff's illness upon consuming the beverage from a bottle containing a dead beetle. The *ratio decidendi* of the finding included the determination that

> ... a manufacturer of products, which he sells in such a form as to show that intends them to reach the ultimate consumer in the form in which they left him with no reasonable possibility of intermediate examination, and with the knowledge that the absence of reasonable care in the preparation or putting up of the products will result in an injury to the consumer's life or property, owes a duty to the consumer to take that reasonable care. (*Donoghue v. Stevenson* 1932 AC 599.)

[9] We here follow [Bartha, forthcoming].

In a case from a court of equal or higher jurisdiction, the judge in a present case must apply the reasoning of the previous case unless there exist relevant differences between the two. Should the present judge determine that relevant differences exist, he must weigh the possibility that the previous *ratio* generalizes in such a way as to override these differences. We note in passing the adaptability of the doctrine of *stare decisis* to the *MATAA* model of analogical argument. In the preceding case we find two elements of this model. The judge's *ratio decidendi*, which corresponds to *MATAA*'s *Generalization Argument*, and the judge's finding, which corresponds to *MATAA*'s instantiation of the *Generalization Argument* (which corresponds to The Violinist in our reconstruction of Thompson's example). The third component of the *MATAA* model is embedded in the subsequent trial. It is the court's finding in that case (corresponding, again, to *The Pregnancy*), drawn by instantiation from the previous *ratio* (corresponding to *The Generalization*). The law's unanalyzed notion of relevant difference is easily handled in the *MATAA* model. There is no relevant difference between the present and prior cases just when they each instantiate a common *ratio* in a P-preserving way. P-preservation here answers to the legal notion of reasoning adequate for the determination of a legal fact.

Where a judge finds relevant difference to exist between his present and some earlier cases, the duty to ascertain whether the *ratio* of the preceding case can be generalized to the present case may strike two different forms, both of which can be accommodated in the *MATAA* model. In the first instance the judge attempts to determine whether the prior *ratio* generalizes in a P-preserving way to a more general form of reasoning from which a like finding in the present case could be seen as a P-preserving instantiation. In the second instance, the present judge leaves the generality of the preceding *ratio* untampered with, and tries instead to determine whether the facts of the present case constitute an approximately P-preserving instantiation of the previous ratio. As we saw earlier in this chapter if the judge proceeds in the first way, his finding in the present case satisfies precedent but is not abductive, since P-preservation conflicts with the ignorance condition. If the judge proceeds in the second way, he allows himself a presumptive finding for the facts of his present case from the prior *ratio*. But this too is not abduction. In finding that the facts of the present case approximate to an instantiation of a prior *ratio*, it is not in general a requirement that his estimate of the similarities be conjectural.

The authority of legal precedent is subject to loose and strict interpretations [Llewellyn, 1930]. Precedents are usually interpreted strictly when they are considered as defective in some way. Strict interpretations limit the harm done by bad juridical determinations. In contrast, precedents are interpreted loosely when they are thought good enough instances of legal reasoning to justify their widest possible application. A strict interpretation of a precedent often involves the judge in finding a relevant difference in the present case which almost certainly would be regarded as nonexistent under a loose interpretation. This is natural occasion for a foolish or inexperienced judge to make mistakes of relevance. Thus in English legal practice strict interpretations of precedent tend to be reserved for highly experienced judges.

The contrast between relevant and irrelevant similarities works in tandem with the contrast between strict and loose interpretations of precedent. We have indicated how the first contrast is captured by the *MATAA* distinction between P-preserving instantiation and P-approximating instantiation. Likewise, a judiciously made strict interpretation will require P-preservation for strict values of P. Loose interpretations, in turn would apply to those more open to P-approximation for less strict values of P. Accordingly

Proposition 8.38 (*Stare decisis in* MATAA) *Virtually all the essentials of the legal doctrine of* stare decisis *are preserved in the* MATAA *model of analogical reasoning.*

Proposition 8.39 (Bad precedents in *MATAA*) *A significant exception to Proposition 8.36 is that* MATAA *(rightly) does not presume the "badness" of Generalization arguments involved in P-preserving analogies with high values forP.*

Corollary 8.39(a) *The presumption of bad reasoning in precedents subject to strict interpretation is a feature of analogical reasoning that is peculiar to legal contexts.*

8.9.1 Precedent

The duty to honour legal precedent cuts across the distinction between strict and loose interpretation of them. As we have seen, the duty is a defeasible one subject to the possibility of exception. Since 1966 the House of Lords, which is England's highest court has had the right to deviate from its own precedents when it appears right to do so, or when not doing so would create an injustice, or when not doing so would unreasonably impede the proper development of the law [Cross and Harris, 1991, p. 104]. Even so, normally its own precedents are binding upon the Lords.

The legal duty to honour precedents resembles the epistemic "duty" to treat similar cases similarly. In other words the doctrine of *stare decisis* has a natural counterpart in the Principle of Symmetry. *Stare decisis* has an avowedly pragmatic justification. For all its faults, precedential reasoning fosters some important social benefits — consistency, certainty and finality in the law. Does the counterpart Principle of Symmetry enjoy a similar backing? More particularly, can it be established that the *scientific* value of the symmetry principle rests on the same kinds of pragmatic consideration? One of the attractions of [Bartha, forthcoming] is the case it makes for an affirmative answer to this question. If Bartha is right, he has made an important contribution to the epistemology of science, by emphasing the closeness of both the value and the structure of analogical reasoning in science to how these matters fare in English common law.

Given that our purpose here is to elucidate abduction and that we have found nothing that is intrinsically abductive about legal analogizing, further discussion of Bartha's thesis is beyond our reach. But we note in passing that if Bartha's thesis is correct, neither is there anything intrinsically abductive about scientific analogizing. This would be a highly significant result, if true. Even under its loose interpretation, *stare decisis* reasoning lacks the general structure of abductive resolution. For let the task of the judge in a present case be to determine whether something is a legal fact. In reaching this determination, the judge has possession of certain facts led in evidence. His task is to decide whether such evidence E is a sufficient basis for the assertion of a legal fact F. Guiding the judge is a prior ruling in a case in which a legal fact F' was asserted on the basis of evidence E'. This being so, we have it that for a suitable interpretation of $\looparrowright$, the judge in the precedent-setting case, found that $E' \looparrowright F'$ obtained, and that it obtained in such a way that an inference of F' from E' would achieve some required cognitive standard P. If in the present case, the judge finds E and F to be such that an assertion of F on E is indicated by a loose interpretation of the precedent, one way in which this finding could be constructed is as follows.

1. The judge knows that $\langle E, F \rangle$ is a P-preserving inference.
2. The judge notes that E' resembles E' and F resembles F'.
3. The judge notes that these similarities are such that the ingerence $\langle E', F' \rangle$ is approximately P-preserving.
4. The judge knows that E'. Accordingly, he infers that F'.

Except where *stare decisis* is persuasive the judge has a duty to affirm F' as a fact if he is persuaded that $\langle E', F' \rangle$ is P-approximating to a sufficient degree. Even when the authority of precedent is only persuasive, the judge may not involve the precedent unless he finds that $\langle E', F' \rangle$ is P-approximating to a sufficient degree. But what is missing in the structure of this reasoning is occasion to conjecture a hypothesis that facilitates the hitting of the judge's target. What the judge must determine is whether E' gives him adequate basis to declare F'. In those cases in which the legal finding is determined by precedent he is required to find that E' and F' bear appropriate similarities to counterpart parameters in the precedent under consideration. In discharging none of these tasks is he required to conjecture that E' or that F'. In fact he cannot conjecture that E' (since it is already known to him) and he must not conjecture that F' (since his duty is to *assert* F' if it passes the relevant tests). So we say again that whereas the structure of *stare decisis* is analogical as such, the same cannot be said for its abductivity.

8.10 Analogue Modelling

Similarity recognition is a *sine qua non* of cognitive practice. Darden rightly observes that "the term 'analogy' has been used to designate any similarity relation" [Darden, 1991, p. 246]. As against this, she suggests that similarities constitute

> a continuum from identical (a thing is only identical to itself) to inductively similar (one can class the two things together and form a scientific law about them) to merely analogous (the two things are from different fields but have some similarities) to completely different (the two things have no similarities) [Darden, 1991, p. 245].

She observes that it

> may not be clear exactly where inductive similarity stops (e.g., both these two bodies have mass, so a general law can be formed encompassing them) and analogy begins (e.g., both sound and light are waves and thus analogous) [Darden, 1991, p. 245].

Inexact as these boundaries are, we join with Darden and others in thinking that significant differences mark the contrasts among inductive similarity, cross-field similarity, and analogies. [Hesse, 1966] effectively pleads the case for a real difference between inductive similarity and analogy, [Gentner, 1983, pp. 72–77] presses a similar case for distinguishing analogy from what she calls "literal similarity".

It is interesting to note that on Darden's continuum, the *MATAA* model of analogy is more a matter of inductive similarity than analogy. In Thomson's example, there are two different sets of data which the arguer attempts to place within the ambit of a single *Generalization* argument. Apart from the qualification "scientific", this is just about a perfect textbook example of "class[ing] two things together and form[ing] a scientific law about them." This being so, we should flag our claim of late in the preceding section that analogy is not intrinsic to abduction. What requires this reconsideration is that, whatever the details of Darden's account of analogy (something she doesn't furnish in [Darden, 1991]), hers is not a *MATAA* conception of it.

The key to Darden's approach is that analogical reasoning must involve the pairing of items not just of different kinds, but of so marked and deep a difference as to require that judgements in question be construed *metaphorically*. When we say that the interior of an atom is a kind of solar system, there is (as Gentner reminds us) nothing in the claim that is literally true. In this respect, analogue models in science resemble what some investigators call "analogical characterization", or, more traditionally, "analogical predication". (See here [Woods and Hudak, 1992] and references therein.) Analogical predication is typified by claims such as "Philip Mountbatten is the First Lady of the United Kingdom" or Stravinsky's pungent observation "Puccini is the Verdi of music". Analogue modelling preserves this factor of non-literalness. Let us suppose that the modeller's task is to account for certain phenomena F. Suppose that he observes that F resembles F', that E accounts for F and that E' resembles E. Even in this elementally schematic form, it is easily seen that the modeller may be attracted to either of two different conjectures:

1. E' might account for F'
2. Since E' might account for F', E' might be the case.

If we are intent on preserving what is distinctive of the Darden-Gentner conception of analogy, it is necessary to construe the embedded resemblance claims non-literally. Such claims constitute a special class of similarity-in-difference propositions. It is extremely difficult to capture assertability conditions for such propositions ([Woods and Hudak, 1992], and, for criticism, [Lichter, 1995, pp. 104, 285–297]). But some hings are clear. The analogy between the consort of the British Monarch and the wife of the President of the United States will fail unless at least two conditions are met. One is that there must be certain highly pertinent similarities between the two. The other is that their differences run so deeply as to make the analogical claim literally false. These requirements jointly provide that the similarities of the case be so significant as to endow a falsehood with the dignity of an analogical truth. We say again that the details of such structures are still not well known. But there is no difficulty in seeing why such similarities should prove so helpful for hypothesis-generation and hypothesis-engagement. Such similarities we might call *telling*. This enables us to say that what is characteristic of analogue modelling, [Darden and Cain, 1989], certain forms of type-theoretic reasoning [Genesereth and Nilsson, 1987; Gentner, 1981], and of cross-field similarity reasoning [Boyd, 1979] and visual modelling [Magnani, 2001a], is almost certainly what makes for the utility of analogies in hypothesis-search and selection tasks. In his attempt to pick a winning answer from up to arbitrarily large sets of alternatives, the sooner the would-be abducer can discard irrelevant possibilities the better. One might think that this task is essentially completed when the reasoner discards all possibilities that are inductively dissimilar to the data of his problem. In fact, however, this could be ruinous for the abduction in progress, since lots of abduction problems have solutions that exceed the reach of any inductively relevant similarities presently available to the abducer. If the abducer is to succeed he must operate without the comfort of this kind of inductive support. His capacity for discerning telling similarities is a case in point. Reasoning from telling similarities is serious example of serious thinking in the absence of inductive determinants. Whether our abducer has used analogical means in his choice of hypothesis, that he is able to do so at all shows that he is able to operate outside the inductive box. This is something shared by analogical characterization and hypothesis-selection even when it is not analogically based.

In both cases, the reasoner is able to function non-inductively in Darden's sense. It is far from surprising, therefore, that these two ways of non-inductive proceeding should so naturally converge in hypothesis searches in abductive contexts.

As a last word, what now do we make of the Darden schema for abduction? In the section preceding this one, we noted that if analogies are construed in the *MATAA* model, abduction is not intrinsically analogical. As we now see, Darden's analogues are not *MATAA* analogies. They bear a considerable similarity to predicational analogies. It is doubtless true that analogies of this kind often facilitate the task of finding hypotheses, but is equally clear that they are not a necessary condition of doing so. So again we conclude that we have not yet found a conception of analogy for which the Darden schema is true for abduction as such. We note also that analogical predication fails to preserve Bartha's common core of good analogical arguments (viz., prior association and potential for generalization). Where is the potential for generalization in "Philip Mountbatten is First Lady of the United Kingdom"? Bartha's core is the core of analogical *arguments*; and analogical predications are not arguments.

Chapter 9

Interpretation Abduction

"John is an Englishman. Therefore, John is brave"

Paul Grice

9.1 Hermeneutics

Hermeneutics is the science of interpretation. The cases we examine in this chapter fall — or appear to fall — into a general category called *interpretation problems*. We shall be dealing mainly with the interpretation of linguistic data, and will touch only briefly on how this bears upon visual data. Linguistic interpretation problems arise in two principal ways — in the interpretation of texts, and in the interpretation of interactive discourse. The general form of the problem is one in which a portion of text or a fragment of discourse carries a meaning that is not directly expressed. The interpreter's task is usually taken to be one that cuts across a distinction between the hidden meanings of the data themselves and the hidden meanings of the producers of those data. Concerning the first, it might be argued that the sentence "All John's children have the measles " carries the meaning "John has children", attested to by various theories of presupposition [Levinson, 2001; Kempson, 1975][Stalnaker, 1973, pp. 447–457]; and concerning the second, it might be part of what the utterer means, or what the utterer intends his interlocuter to understand, that John's children don't now need to be innoculated against measles. If we take it that the interpreter's K is what he knows the words and sentences of the text or the utterance to mean directly, the trigger is that K appears not to tell him what their (and their utterer's) hidden meaning is [Grice, 1989; Heim, 1991].

We should not make too much of the fact that, in one of its standard uses, "hidden" means "secret" or even "furtive". Here it has a more relaxed sense. Hidden meaning resembles circumstantial evidence. In law, a fact is secured by circumstantial evidence when it arises from it by inference. This contrasts with direct evidence. A fact is secured by direct evidence when the fact is concluded without inference. Hidden meanings likewise are meanings that are inferred from the words of a text or utterance, rather than understandings that flow from them directly. Some linguists (and Peirce, too) are attracted to the view that *all* cases of meaning-comprehension are inferential in character. Right or wrong, our distinction is still left standing. Even omni-inferentialists allow that there is an intuitively workable contrast between the sentence-meaning of "There's beer in the fridge if you'd like some" and the utterer's meaning to the effect that his interlocuter should feel free to help himself. If both understandings did arise inferentially, it is possible regardless to preserve the contrast by characterizing the second of the two inferences as *presumptive* (which is precisely what Stephen Levinson does in the title of his book, *Presumptive Meanings* [Levinson, 2001]). Similarly, if grasping the utterance-meanings of "All John's children have the measles" and "John has children" is in each case a matter of inference, only the latter makes a plausible claim on presumptive meaning determination.

The interpreter's target T is to construct (or otherwise attain) a suitable interpretation of full meaning, both direct and presumptive. When the interpreter's situation is such that the target cannot be hit simply by extending K (say, by reading some adjacent text or asking the utterer what he means), his problem might be thought to be consequentialist abductive. If so, a solution would be constructed along the standard lines, in which the interpreter conjectures a semantic hypothesis H which, in concert with K, makes it true that, for a payoff V, $K(H) \looparrowright V$ holds in such a way that T is reached. As before, this requires that $\looparrowright$ itself bear an interpretation appropriate to the role performed by $K(H) \looparrowright V$. So here the $\looparrowright$-relation might seem to operate as a kind of nonformal semantic consequence relation. It is a matter of some contention as to whether meanings carried by deductive relations such as entailment, or pragmatic relations such as presupposition, honour the nescience conditions that are integral to abduction. Given, for example, that "Harry fell down" entails that someone fell down, and that "All John's children have the measles" presupposes that John has children, whether we allow that any K containing "Harry fell down" and "John's children have the measles" will also contain "Someone fell down" and "John has children" will pivot on the K-closure conditions for entailment and presupposition. Important as it is, we will not press the issue here. It suffices for our purposes that there are interpretation problems in which the fuller meaning of a source does not lie in the K-preserving closure of what is already known of it. As long as we can be assured that the requisite distinctions exist, it will be possible to explore whether the determination of implicit meanings is intrinsically abductive, without having the complete story of how the distinctions operate. Among logicians, the best-known interpretation task is the *enthymeme* resolution problem, also called (with less than perfect accuracy) the problem of missing-premisses [Hitchcock, 2002].

9.1.1 Enthymemes

It is well to note how narrowly circumscribed enthymeme problems are. They are restricted to utterances of a given type (arguments) and they seek the 'repair' of the utterance in respect of a single designated property (validity). In real-life situations, most articulate utterance is incomplete in one way or another, and yet argumentative utterance is comparatively rare. Even when a transmission does take on the features of an argument, the problem for the resolver of the enthymeme is rarely to find a way to make it valid. The first point reflects the fact that most human utterance is non-argumental; and the second gives us occasion to notice that, even when making arguments, participants more often than not strive for standards other than validity. This, in turn, may reflect the fact that, except when trivially realized, validity is an expensive commodity, and that guarantees of truth-preservation do not in general come cheap.

In a way, it is unfortunate that enthymeme resolution is treated as a kind of presumptive meaning determination. If it is a normalic truth that human argumentation is not bound by the validity-standard, then on any occasion of argument-interpretation, the interpreter's default position will be that the arguer is not intending to make a valid argument. There are important exceptions, of course, typically indicated by context and sometimes by subject matter (e.g. a geometric proof). All agree that it is the task of enthymeme resolution to interpret a text or an utterance in such a way as to transform an invalid argument into a valid argument. Given the point at hand, the enthymeme interpreter's first chore is to discern whether the text or utterance is in fact bound by the validity standard. It is surprising that so critical a factor is so frequently overlooked. Nor are these omissions reserved for cases in which the interpreter is trying to discern what the arguer's standard actually is. A similar omission is also a commonplace in philosophical disputation, in which a charge in the form " But that doesn't follow" is routinely allowed to be decisive if true. In fact, however, although philosophical arguments sometimes aim at validity, for the most part they do not. For every proof of God's existence, there are, for example, dozens of abductive arguments against most of the garden variety scepticisms. The frequency with which philosophers fold under the accusation, "It doesn't follow", is sharply orthogonal to the infrequency with which philosophical case-making is validity-bound.

It would be ill-advised, however, to make too much of the classical tie between enthymeme's and the standard of validity. It is possible to soften the targeted standard in any degree that preserves the reneral idea of a good argument, without having to alter the basic structure of enthymeme resolution problems. Let S be any standard which, if attained, would render an argument in some sense good. Intuitively, in addition to validity, S ranges over properties such as inductive strength, statistical support, plausibility, and so on. Now unshackled from its exclusive tie to validity, enthymeme's are arguments which, for some S, fail to be S as expressly presented; and an enthymeme resolution is one which under appropriate conditions transforms the originally not-S argument into one that is S. In what follows, for ease of exposition we will take the classical cases (i.e., the validity-seeking cases) as stand-ins for them all.

There is, then, a significant tension between an attempt to understand more fully what an utterance or utterer says and an attempt to find an interpretation which transforms it from an invalid to a valid argument (or, more generally, from a non-S to an S argument). As long as the two tasks are conceived of as having to be jointly performed, the interpreter is met with a number of discouragements. To take just one example, every invalid argument has a valid transformation. If it is contextually or in some other way indicated that the argument in question is bound by the validity standard, then the joint performance of this pair of tasks obliges the interpreter to attribute to the utterance or the utterer a meaning that meets the validity standard. Since that standard can always be achieved, the interpreter may find himself in the implausible situation of having to suppose that *understatement* is the only way for anyone actually to make an invalid argument.

Perhaps this problem can be averted if we conceive of enthymemes as a proper subset of the set of invalid arguments, whose members could be described as *enthymematically valid.* Let us say that

Definition 9.1 (Enthymematic validity) *An argument is enthymematically valid iff (1) it is invalid, and (2) there is a valid extension of it under requisite constraints* $C_1, \ldots, C_n$.

This is intended to capture the intuition that some arguments are invalid never mind what is subsequently done to tat them up, and others are invalid only through reparable omission. With our schematic definition at hand, we can try to be more specific in detailing the tasks that enthymeme resolvers take on. The enthymeme resolver is required to specify the C_i in ways that preserve the fact that enthymematically valid arguments are a proper subclass of invalid arguments. Another way of saying this is that the agent's central task is that of determining whether the other party's original argument is, as stated, enthymematically valid. Keeping in mind that not every invalid argument is enthymematically valid, the interpreter is precluded from selecting a resolution strategy that obliterates this distinction. So we may take it that although in the general case arguers are committed to the truth of their own premisses and conclusions, hence to the truth of propositions that jointly entail the argument's own material conditionalization, attributing it as part of the argument's or arguer's presumptive meaning will indeed collapse the distinction between enthymematically valid and merely invalid arguments whenever it can reasonably be supposed that the original argument was forwarded sincerely (which is nearly always.)

We find ourselves in the jaws of two conflicting intuitions. When it comes to determining what someone means in uttering something (or what that utterance alone means) it is best to restrict our postulations to those already suggested by the original utterance. On the other hand, we want also to proceed in ways that preserve the distinction between enthymematic validity and unqualified invalidity. If, accordance with the first intuition, we restricted our attributions to propositions for which the original itself offers some support, it would be difficult to see how the interpreter could avoid postulating the material conditionalization of the subject-argument, the premisses and conclusions of which having already been forwarded as true. But again, if so, the distinction between enthymematically valid invalid arguments and unqualifiedly invalid arguments is all but lost. On the other hand, the best way of preserving that distinction is to restrict one's posits to those propositions which, once attributed, validate the argument, but which, apart from that, are unimplicated in features of that argument in its original form. An example may help with the present point. Consider, imprecise as it assuredly is the following pair of arguments.

	I		II
1.	Socrates is a man	1.	All men are mortal
2.	∴ Socrates is mortal	2.	∴ Socrates is mortal

With argument I, if we restricted our supplementation to propositions to which the argument or arguer are already pledged, or to immediate consequences thereof, we would add to I the proposition "Socrates is a man $\supset$ Socrates is mortal". This would validate the argument in ways that flow from entirely reasonable attributions; but, since the process applies with full generality, all arguments of type I would be enthymematically valid. Accordingly, we should select our attributions from propositions we have reason to suppose the utterer thinks true but which are not implicated in the original argument in the way that "Socrates a man $\supset$ Socrates is mortal" is, and which, if added, would make for the desired transformation. "All men are mortal" fits this bill.

Argument II is both similar and different. Although "All men are mortal" and "Socrates is mortal" yield the true and validating material conditional "All men are mortal $\supset$ Socrates is mortal", this too holds for every argument of type II. Attributing it as the missing or hidden premiss would again cost us the distinction between enthymematically valid and unqualifiedly in- valid arguments of this type. Accordingly, the leading intuition among enthymeme investigators is to posit "Socrates is a man" as the missing or hidden premiss. "Socrates is a man" fits the bill. It is not implicated by the original argument in the way that its own premiss and conclusion are, and their own joint immediate consequences are. It is nevertheless not implausible to attribute it; and once attributed, validity is attained without collapsing the distinction between enthymematic validity and invalidity.

But consider a third argument,

III

1. All men are mortal
2. Socrates is a man
3. $\therefore \varnothing$

in which the conclusion is left unexpressed. No one doubts the intuition that tells us that "Socrates is mortal" is the obvious candidate. Since "Socrates is mortal" is an immediate consequence of the argument's own premisses, the present resolution strategy precludes our positing it. So, quite apart from its imprecision, there is something wrong with it considered as a general strategy.

9.1.2 Fermat's Last Theorem

Possibly the most celebrated enthymeme resolution problem of the past several centuries is Andrew Wiles' quite extraordinary proof of Fermat's Last Theorem. Fermat's Last Theorem asserts that for any $n > z$ there exists no solution in the whole numbers of the equation

$$x^n + y^n = z^n$$

In 1630, ... Fermat inscribed this theorem in the margin of a book he was reading at the time, Diophantus' *Arithmetica*. He added that he had a "truly marvellous demonstration" of his claim, which, given the narrowness of the margin in which he was writing, he omitted to record. There is now ample reason to think that Fermat was mistaken in his claim to have found a proof, but this was not apparent to Wiles at the beginning of his long journey towards the demonstration he would achieve in 1995[1] As is made clear by a 300 year history of failed attempts and by the length and subtlety of Wiles' own proof, the enthymeme resolution problem which Fermat bequeathed to mathematics is one of daunting scope. It is the problem of finding mathematically true propositions (some of which would require subproofs) which jointly considered would serve as premisses in a proof whose conclusion is that for any $n > z$, the equation "$x^n + y^n = z^n$" has no

[1] In a lecture delivered at the Isaac Newton Institute in Cambridge on 23 June 1993, Wiles announced that he had a proof of the theorem. When, in the months following, the proof was being readied for publication, an error was discovered. With the assistance of Richard Taylor, Wiles was able to correct the error, and the proof appeared in print [Wiles, 1995].

solution in the whole numbers. The "missing premisses" of Wiles' proof run to 108 printed pages. When asked, "... is your proof the same as Fermat's"? Wiles replied, "There's no chance of that. Fermat couldn't possibly have had [the 1995] proof ... It's a 20th-century proof. It couldn't have been done in the 19th century, let alone the 17th century. The techniques used in this proof just weren't around in Fermat's time" [Wiles, 2000, pp. 5-7]. When asked further whether "Fermat's original proof is out there somewhere", Wiles replied, "I don't believe that Fermat had a proof" [Wiles, 2000, p. 6].

These are instructive remarks. Not only do they call to our attention the strong likelihood that Fermat mistakenly thought that his theorem was the conclusion of a proof that he himself possessed or could produce at will, but, for want of space, had left unexpressed. But also, given that his marginalia contain not a hint of discouragement about finding such a proof, it may also be supposed that Fermat thought that the missing premisses of his proof were within the competency of mathematicians of the day to reconstruct. This being so, it appears that what Fermat himself believed was, in effect, that his marginalia constituted an enthymeme resolution problem for mathematically competent reasoners. If, however, Fermat had no proof, must we say that he was again mistaken in thinking that his scribblings amounted to an enthymeme? And how, in turn, are we to characterize Wiles' undertaking? If, as Wiles believes, Fermat did not have a proof of the theorem, how can it be correct to say that Wiles' own task was that of solving Fermat's enthymeme?

We find ourselves facing two problems. One is the problem of determining whether Wiles' proof is in fact the solution of an enthymeme problem. If so, the second is that of sorting out whether, in constructing his enthymeme resolving proof, Wiles was solving an interpretation problem. Concerning the first, it is not implausible to suggest that even though Wiles' task cannot have been to reconstruct *Fermat's* own proof, it was his objective to produce *a* proof. Notwithstanding that no proof (enthymematic or otherwise) existed until Wiles, his task remained throughout one of deploying the mathematical wherewithal that would secure the theorem in the appropriate way. Since Wiles' project was, irrespective of whether Fermat himself had a proof, to find premisses (and subproofs) from which the theorem would soundly flow, his project may reasonably be characterized as a search for premisses not theretofore present. To that extent, Wiles' proof has the same structure as a search for missing premisses, where the sole difference (that the premisses sought for had no prior occasion to go missing) is the merest contingency. So our answer to question one is that Wiles' proof qualifies as (a particularly dramatic) resolution of an enthymeme problem.

Our second question was whether in producing the proof that would solve this enthymeme problem, Wiles was doing anything that would justify us in thinking that he was also solving an interpretation problem. In the NOVA Online interview, Wiles allows that in his "early teens, I tried to tackle the problem as I thought Fermat might have tried it" [Wiles, 2000, p. 3]. In time he realized that this was not the way to go. Even if he were mistaken in so thinking, even if he were justified in trying to imagine how Fermat himself would proceed, there is nothing in this that remotely qualifies as either an interpretation of the hidden meaning of the utterance "For all $n > z$, there is no solution in the positive integers of the equation "$x^2 + y^2 = z^2$" or an interpretation of what Fermat himself meant in uttering this sentence. From this we may conclude that

Proposition 9.2 (Prior utterance) *In an enthymeme resolution exercise the premisses (or conclusion) sought for need not be implicit in some prior utterance.*

Proposition 9.3 (The tie to interpretation) *There is no intrinsic general link between enthymeme resolution and the interpretation of either utterance — or utterer's meaning.*

9.2 Enthymeme Resolution as Abductive

Our task here is not to furnish a detailed, still less the definitive, account of enthymematic resolution strategies. The objective rather has been to achieve some headway with a further pair of questions. One is whether, or to what extent, enthymeme resolution is a kind of presumptive meaning determination; and the other is to what extent enthymeme resolution has an abductive character. Concerning the first of these, we have been detailing the respects in which understanding an utterance (or utterer) is different from, and often in some sort of tension with, the validation of an invalid argument. And, using the example of Wiles' proof of Fermat's Last Theorem, we have taken the point further, claiming that enthymeme resolution is not intrinsically linked to the problem of determining presumptive meanings. To the second question we now turn.

We have a basis for saying that enthymeme resolution problems are not by and large interpretation problems. Since the burden of this chapter is to bring some light to interpretative abduction, we have here, save for just one consideration, no occasion for persisting with our discussion of enthymemes. The exception is that, interpretative or not, enthymeme resolution seems intuitively to have an abductive character. So we should pause briefly to investigate this intuition. We begin this task with a clarification. When Andrew Wiles started his quest, he thought that the proof he sought existed but had "gone missing". Thinking that Fermat had the proof, Wiles attempted to think like Fermat. In this particular case, Fermat's marginalia offer no clue as to the structure of the proof, beyond what its conclusion would have to be. So, in trying to get into Fermat's frame of mind, interpreting the scribblings would have availed Wiles nothing. Even so, there are other cases in which a person's writings or utterances do give some clue as to how that person might have thought of the matter that the enthymeme resolver wants to get clear about. To that end, the enthymeme resolver might be well-served by plumbing that text or that utterance for its presumptive meanings. But it will not be the case in general that, in determining these presumptive meanings, the problem-solver will have unearthed any direct link to the full argument that finishes off his enthymeme problem. But we should also bear in mind that Wiles' objective was not to expose Fermat's proof, except in so far as *Fermat's* proof turned out to be *the* proof. At this early stage of his search, we can discern a threefold distinction in what Wiles was doing. First, he wanted a proof of the theorem. Second, he wanted to expose Fermat's own (presumed) attempt at one by putting himself in Fermat's way of thinking. Third, although given the scantiness of Fermat's marginalia, Wiles had no actual occasion to put himself in Fermat's way of thinking by determining the presumptive meanings of those scribbles, this is the sort of thing that could be tried in cases in which the text or the utterance is more forthcoming. With all this said, it bears repeating that enthymeme's are not typically solvable by the determination of presumptive meanings.

What, then, is to be said of the supposedly abductive character of enthymeme resolution? With Wiles' proof again in mind, there is a point of some importance to emphasize at the outset. It is that

Proposition 9.4 (Enthymemes and regressive abduction) *Enthymeme resolution does not have the character of regressive abduction.*

In chapter 5 we saw that, as understood by Russell and Gödel, regressive abduction can be schematized as follows. Let T be the task of justifying a non-obvious ("recondite", is Russell's word for it) principle of logic. Call this proposition H. Suppose that there is an obvious truth of mathematics, V, such that for some set of propositions K, $K(H)$ constitutes a proof of V. Since nothing else (so far as one can see) counts in favour of the truth of H, H is forwarded conjecturally solely on the basis of the role it plays in the proof of Q.

This contrasts with Wiles' solution of Fermat's enthymeme. Then Wiles' T' was to construct a proof of V'. His agenda was both similar to and different from that undertaken by regressive abducers. What the regressive abducer also wants is a proof of V. But that is not all that he wants, or even the most important part of it. What he wants is a proof of V that contains H as a prior line. The regressive abducer's main task is to justify H, not prove V. Proving V is a subagenda whose whole motivation is the support it lends not to V but to H. Wiles, on the other hand, was wholly absorbed with V'. To that end, there are thousands of prior lines, towards none of which did Wiles' show the kind of interest that defines regressive abduction. Wiles took great pains to construct his derivation of V' from mathematically sound premisses (and subproofs), but nowhere in that vast undertaking is there the slightest indication that Wiles' wanted his proof of Fermat's theorem to constitute a justification of any of the premisses he actually deployed. Regressive abducers target Hs. Enthymeme resolvers target V'. This suffices to show that even if enthymeme resolution did have an abductive character, it does not have the character of a regressive abduction problem in the manner of Russell and Gödel.

Is enthymeme resolution abductive? Suppose that it is. Then for any such problem there will be a trigger and a target. The trigger is the recognition that an argument A is not valid. The trigger is to supplement A so that the resulting argument is valid. Let us suppose that P is a proposition that turns this trick, that A^P (as we will write it) is a valid extension of A. These are only the barest details, needless to say. We are simply taking it as given that in his selection of P, the enthymeme resolver does not have *carte blanche*. (Thus he may not add as premiss the negation of an original premiss, or replace the original conclusion with the disjunction of it and its negation, or add as premiss the corresponding conditional of the original argument). Suppose that the enthymeme resolver succeeds in finding a P that does the trick, without having to cheat in any of these ways. Does his success offer one jot of justification to treat P as a *hypothesis*. to conjecture that P is *true*? Does it constitute the adequate grounds for submitting P to trial? These questions answer themselves. Enthymeme resolution is not abductive as such. Having made and repeated the point about the independence of enthymeme resolution and determination of presumptive meaning, it is well to enter an important caveat.

Proposition 9.5 (When enthymeme resolution is abductive) *Let A be an enthymeme. Then A is an argument of a type that fails to achieve some contextually indicated standard. Let A' be the argument that constitutes the resolution of the enthymeme problem created by A. Then if A itself has the structure of an (incomplete) abduction argument, A' too is an abductive argument.*

9.2.1 The Attack on Analyticity

In 1950, W.V. Quine delivered to the Eastern Division of the American Philosophical Association, meeting in Toronto, a bombshell. In a paper entitled "Two Dogmas of Empiricism" [Quine, 1951], Quine mounted a concerted attack on the concept of an analytic truth, i.e., a proposition made true solely by the meanings of its constituent expressions. The notion of analyticity had been a methodological staple of philosophers since Leibniz at least, and was given a central place in Kant's philosophy. It had been appropriated by the logical empiricists as fundamental to their theory of knowledge. In as much as the idea of analyticity had become the object of a deeply embedded presumptive legitimacy among philosophers, Quine's attack was shocking and his thesis was found to be seriously counterintuitive.

Had Quine's reservations about analyticity proved tenable, he would have succeeded in causing a significant derangement of modern philosophy, and would have administered an all but fatal blow to the methodology of contemporary analytic philosophy. "Two Dogmas of Empiricism" is the single most re-printed and anthologized philosophical paper in English of the 20th century. It has, in turn, generated an immense interpretative and critical literature. It is not hard to see why. The thesis of Quine's paper is as radical as it is far-reaching. It is a thesis whose importance requires the clearest possible formulation and the clearest possible supporting argument. Quine's paper fails this pair of expectations dramatically. It is not clear what fault Quine finds with the concept of analyticity and it is not clear what argument he uses to demonstrate the existence of that fault. Concerning the first matter, there are at least six different versions of what Quine's thesis actually is, as advanced by the best of a vast army of Quine-commentators.

Given the sheer size of the reaction to "Two Dogmas" and the intractability of the interpretational dissensus to which it has given rise, it may well be the case that making sense of Quine on analyticity has turned out to be analytic philosophy's largest interpretation problem in a hundred years. Despite the lack of interpretational consensus, it is easy to see that what the best of Quine's commentators seek for is a convergence between an interpretation of Quine's charge that is favoured by an independently attractive supporting case, and an interpretation of Quine's case that is made attractive by the nature of the thesis to which it lends support. All of this is interpretational. Some say that it involves trying to put oneself in Quine's shoes (but see below). It also involves putting the best case that Quine's writings will allow which supports, subject to the same proviso, the best understanding of his thesis. "Two Dogmas" lays fair claim to constituting for its readers a radical enthymeme problem, made so by the fact that there are missing elements everywhere, both in the premisses of Quine's defence of his thesis and in the thesis itself, hence in the argument's conclusion.

This gives rise to a pair of distinct interpretation problems, each bearing a link to enthymeme resolution. The interpreter might wish to determine how Quine himself would have explicitized his own thesis and his own argument for it; or the interpreter might wish to reach to find the best claim against analyticity that he finds to be compatible with Quine's express writings on the matter, and with the same proviso, the best full case for the thesis thus interpreted. In the first instance, the interpreter fails in his quest if he fails to identify Quine's own understanding of the thesis and the case he (Quine) thinks supports it. In the second instance, the interpreter may (and perhaps should) fail to achieve these goals without defeating his own purpose. What he desires to achieve is not Quine's own understanding, but rather his own rendering of the best that passes muster with Quine's words. The first resembles an utterer's meaning problem. The second may be likened to an utterance-meaning problem. The first seeks Quine's understanding. The second seeks a Quinean understanding, but even so an understanding that is the interpreter's own.[2] Once again, it is possible to discern in the ways in which these problems are solved aspects of abductive reasoning. A critic might conjecture that the best determination of what Quine himself takes his thesis to be and that what justifies the conjecture, apart from comportability with Quine"s own words, is that H is also compatible with other doctrines Quine is known to espouse and that H is, or strikes the interpreter as, defensible in ways that Quine would likely approve of. Equally, apart from its conformity with Quine's words, a critic might conjecture that A is the best argument for H to attribute to Quine on grounds that it marshalls considerations that there is reason to believe Quine would be well-disposed toward, and in ways that Quine is known to favour.

All of this qualifies for the status of abduction and all of this is linked to enthymeme resolution. But it is not linked in such a way that the interpreter's success with his interpretation problem is constituted by his success at enthymeme resolution. This is guaranteed by the conjectural character of abduction. In providing what he takes to be Quine's understading of his own thesis and of the full case that he actually makes for it, our interpreter conjectured an enthymeme resolution is not his own but rather is one attributed to Quine.

Much the same can be said for the second kind of case, in which the interpreter undertakes to assemble his own thesis and his own case for it subject only to the condition that neither offend against Quine's express views. Here, too, there is plenty of room for conjecture. The interpreter might fix on H' as the thesis in question because of its plausibility or because he thinks he sees a way of defending it, and he may select A as the best defence of H' partly because he thinks that A makes a plausible case for H'. If it were possible to make these determinations fully explicit, we would see a reconstructed argument $A' = \langle P_1, \ldots, P_n, H' \rangle$. This argument would be a solution of an enthymeme problem if it hit a standard S (plausible inference from plausible premisses, say) by which the argument would be suitably good. Here the interpreter has selected the components of A by abduction, notwithstanding that A itself is not an abductive argument unless the enthymeme it constructs was itself an abductive enthymeme.

9.2.2 Inarticulacy as Economics

Embedded in our discussion so far are two questionable assumptions. One is that whenever an utterer knows what his utterance means he also knows what its full

[2] Quine returned to his analyticity thesis in his copious writings after 1952, in many of which he made replies to his critics and offered clarifications. (See here[Hahn and Schilpp, 1998].) Regardless, the interpretational dissensus persists, and amazingly, the received opinion among analytic philosophers of the present day is that Quine's thesis about analyticity is correct. In an important retrospective of Quine's views in 1969 [Davidson and Hintikka, 1969] there is no paper on analyticity. Presumably, the editors considered the analyticity issue was a closed book in 1969.

meaning is; and that whenever he knows what his utterance means, he also knows what he himself means in uttering it. The other is that in interpreting the meaning of another's utterance or of what the other means in uttering it, it is best for the interpreter not to permit subjective factors enter into his reflections. Given the first assumption, whenever we seek to determine what Quine's utterances mean and what he means in uttering them, it is always best and always decisive to ask Quine himself. Where direct confirmation is not available, our second assumption provides that in seeking interpretations of what Quine's utterances mean and of what he himself meant in uttering them, it is not desirable that the interpretations attributed to Quine flow from the interpreter's own favourable impression of them. In the present subsection, we briefly reflect on reasons for doubting both assumptions. ([Polanyi, 1966; Howells, 1996; Fleck, 1996]). We approach this task with an examination of the phenomenon of *inarticulacy*.

The inarticulacies that we intend to track fall into two groups, circumstantial and constitutional. In a rough and ready way, a circumstantial inarticulacy turns on something omitted by a speaker in the face of conventions or under press of contextual cues which the speaker is at liberty in principle to violate or ignore. Constitutional inarticulacies arise from omissions made by a speaker owing to the way he is structured as a cognitive agent. In the general case, these are omissions which it is impossible, or anyhow very difficult, for a speaker to repair. Note that if constitutional inarticulacies exist, then assumption one, above, cannot be true as it stands.

Our remarks remarks here draw on the model of cognitive agency developed in chapter 2. We shall attempt to show that the dominant motivation for a speaker's inarticulacies is his need as an individual cognitive agent to proceed economically in thought and speech. As we have said, the individual is a cognitive agent who frequently must prosecute his agendas on the cheap. Things left implicit are an important feature of his cognitive economy. Speaking in enthymemes is only one way of achieving such economies. But in as much as enthymemes have been on our minds of late, they may serve as a representative sample.

In interactive settings, the enthymeme poser addresses an argument to an interlocuter, who is the enthymeme solver. Resolution need not be explicit. Rather it is supposed that in interpreting the utterance of the enthymeme maker the enthymeme solver himself *implicitly* supplies what is missing. There is a reason for that. Doing so preserves the economies achieved by the original utterer.

While the enthymeme solver normally lacks occasion to voice details of his solutions, this may be something he is able to do at will (and, on the traditional account, can do). For his part, although the enthymeme poser leaves certain things unsaid, he could in principle, on the traditional account, have completed his argument. Whether or not to proceed enthymematically, is left to the individual's discretion. Enthymeme posing and enthymeme resolution would therefore appear to fall into the category of circumstantial inarticulacy. We repeat in passing that enthymemes are just a special case of a much large richer class of circumstantial inarticulacies, which we might think of generally as *understatements*.

Speaking and understanding enthymemes is utterly common in human discourse and is something we seem to manage with ease by and large. The commonness and ease suggest the operation of comparatively primitive skills. The suggestion is born out by empirical studies of language acquisition, according to which understatement is a feature of speech from the earliest manifestations of a child's literacy. Understatement would seem to be an intrinsic feature of linguistic competence. *To state is to understate.*

9.2.3 Some Virtual Guidelines

As we have seen, it is surprisingly difficult to specify the right virtual rules even for enthymeme problems. Once the constraints imposed by the validity-only target are lifted, rules become even more complicated to get right. Notwithstanding, we now suggest some general guidelines, which for the most part are *virtual* rules (since most resolution takes place 'down below'). To make our exposition manageable we shall, however, continue to focus on the interpretation of arguments.

1. Arguments are often incompletely formulated. In their express form they fail to possess certain properties of arguments in which the arguer can be supposed to have an interest.

2. The job of the interpreter is to attempt to supplement the argument as stated in ways that secure the desired property or properties.

In so doing, the interpreter is under two general obligations in whose joint fulfillment there can be, as we have said, no antecedent guarantee of consistent compliance. The interpreter must, on the hand, construe what his interlocuter meant to say, and on the other, must try to find in such an interpretation factors that will, once assumed, achieve for the argument its desired objective.

What *properties* an argument is aiming for are usually contextually available to the interpreter. What *interpretation* to consider is in turn a matter of common knowledge and context.

Common knowledge (also discussed in chapter 7) is knowledge had by a community, K. The community may be as big as the entire population of the world or as small as the spousal unit created by marriage. Knowledge is common when and to the extent that it can be presumed that members of K have it and that being members of K suffices for their having it, other things being equal.

The commonality of what is commonly known makes for considerable economies, as we had occasion to emphasize in our discussion of presumptive reasoning in chapter six. If P is a proposition that Harry claims to know by common knowledge, then by its commonness it is also what Harry would expect that the others would claim to know; and this he can know without having to ask, so to speak.

A further economic feature of common knowledge is its *tacitness*. It is entirely explicable that this should be so. If P is a proposition of common knowledge, then it may be presumed that Harry knows it, and that if he knows it as a matter of common knowledge that he knows that you know it, too. Then, since we are spared the nuisance and cost of publicizing P, there is less occasion to give it expression than other things we know.

A context furnishes guidance to those whose actions and interactions are influenced by a shared knowledge that arises from particularities of their interaction. Context makes common knowledge possible for pairs of interacting participants. So contextual knowledge is a species of common knowledge. The two are also connected in a further way. What is common knowledge for Harry is a matter of the epistemic community to which he belongs. Knowing that these, but not those, are the epistemic communities to which he belongs is itself a matter of context. It is a matter of the particularities of his interaction with his fellow beings that answers that question.

9.2.4 Background Knowledge

Like common knowledge of all stripes, contextual knowledge is also often tacit. The same is true of the *background knowledge* of a scientific theory. In some respects, a theory's background is the knowledge common to fellow scientists, and in other respects, a theory's background is a matter of, say, details of the functioning of the field team's apparatus. What is common to background knowledge in scientific theories is the following.

1. If P is a proposition of background knowledge for a theory Θ, P is not likely to be used expressly as a premiss in formal derivations of the laws and theorems of Θ.

2. Notwithstanding, P's truth affects positively the epistemic context in which it was possible for theorists to derive Θ's laws and theorems.

3. In the event that Θ is met with counterevidence, the likelihood increases that it is P that will be forwarded expressly, and considered for express rejection, rather than some expressly derived law or theorem of Θ.

Feature (1) suggests that scientific theories, like enthymemes in a certain way, are expressively incomplete, and that their correct understanding requires an interpreter to marshall the appropriate Ps of background knowledge, and further that this be done for the most part tacitly.

Characteristic (2) suggests the need to understand with care the distinction between a theory's heuristics and its laws and theorems. Once way for a proposition P to contribute positively to the epistemic context in which laws and theorems are derived is to influence the theorist in figuring out how to make his proof. Another way is for P to function as a tacit premiss, or as a tacit rule, in these formal derivations.

Property (3) bears on refutation. A refutation problem with respect to a theory Θ is the problem of determining with optimal specificity that aspect of Θ which an apparent counterexample to one of its laws or theorems makes necessary. It is a matter of figuring out where to 'pin the blame'. In those cases in which it is judged best to pin the blame on P, a proposition of the theory's background, there are two consequences of note. One is that refutation management is a standard way of moving P out of its former anonymity, or bringing P out of the closet, so to speak. The other is that it lies in the very nature of pinning the blame on P that we settle the question of whether P was a tacit premiss in Θ's formal derivations, or whether the tacit recognition of P facilitated — causally assisted — the theorist in thinking up the ways and means of deriving a law or theorem of Θ.

We said that the task of the interpreter of an incompletely expressed argument is to furnish an interpretation that can plausibly be attributed to the other party and which strengthens the argument's claim on some contextually specified target property. Doing this, we said, was a matter of common knowledge and of context. How?

We postulate the following virtual rules.

Proposition 9.6 (The common knowledge rule) *For reasons of economy, and of plausibility as well, the interpreter should try to draw his supplementations of the original argument from what is common knowledge for both interpreter and arguer.*

If this is right, our second assumption from the subsection just above is also called into question. The present rule bids the interpreter to allow his attributions to reflect the fact that, independently, they find favour with the interpreter himself. The rule pivots on the commonness of common knowledge. If P is common knowledge for the interpreter and there are contextual indications that both parties inhabit an epistemic community of a sort appropriate to discussion of the original argument, then P is knowledge for the other part, the interpreter knows this; and thus can attribute it without incurring the costs of further enquiry, and can attribute it plausibly.

Proposition 9.7 (The contextual knowledge rule) *If a candidate, P, for interpretation of the original argument is not upheld by the prior rule, the interpreter should seek to make his selection from the contextual knowledge furnished by particularities of his interaction with the original arguer.*

The rationale is the same as before: economy and plausibility.

Our cases admit of some straightforward formalization. If the original argument is

$$\begin{array}{ll} 1. & A \\ \hline 2. & B \end{array}$$

then the judgement that it is an enthymematically valid argument is conveyed by acknowledging that

$$K{:}\quad \frac{A}{B}$$

where K is background knowledge accessible in principle to utterer and interpreter alike, and is read:

> K gives that since A, B.

Then the interpreter's task is to specify a subset k of K such that

$$\begin{array}{c} k \\ A \\ \hline B \end{array}$$

is valid (or otherwise S).

Seen this way, enthymeme resolution is a specification task, itself a kind of *revision task*. Refutation management has a similar structure. A theory Θ always has a background K. If something contradicts Θ (as is said loosely and conversationally), the theorist often has the option of 'blaming' Θ or 'blaming' K. But either way his further task is to specify the source of the error, whether in Θ proper or in K.

As has been said, the two fundamental tasks of the linguistic interpreter may prove not to be consistently co-performable. It is one thing to figure what the original arguer was meaning to say; it is another thing — and sometimes a thing incompatible with the first — to spin the original in ways that make it a better argument in some contextually indicated way. How are such tensions to be avoided or minimized? Can they *be* avoided or minimized? Enter the Principle of Charity.

9.3 Charity

The Principle of Charity found its place in the common currency of argumentation theory by way of two sentences from a First Year logic textbook.

> When you encounter a discourse containing no inference indicators which you think may nevertheless contain an argument, stop and consider very carefully whether such interpretation is really justifiable.... A good rule is 'the Principle of Charity': If a passage contains to inference indicators or other explicit signs of reasoning and the only possible argument(s) you can locate in it would involve obviously bad reasoning, then characterize the discourse as 'non-argument' [Thomas, 1977, p. 9].

Three years later, Michael Scriven provided the principle with what has become a widely attended to and well-received characterization.

> The principle of Charity requires that we make the best, rather than the worst, possible interpretation of the material we are studying [Scriven, 1976, p. 71].[3] The Principle of Charity requires that you be fair or just in your criticisms. They can be expressed in heated terms, if that is appropriate; they may involve conclusions about the competence, intellectual level, or conscientiousness of the person putting forward the argument, all of which may well be justified in certain cases. But your criticisms shouldn't be unfair; they shouldn't take advantage of a mere slip of the tongue or make a big point out of some irrelevant point that wasn't put quite right.[4]

Charity requires that we not be unfair in our disputes. Scriven's principle also implies:

> Interpret your opponent or your interlocutor correctly.

It is perhaps a bit odd that in virtually none of the discussions of the Charity Principle in theories of argumentation is there any recognition that something called the Principle of Charity is at the centre of a brisk contention and a huge literature in the adjacent precincts of philosophy of language and cultural anthropology. [5]

Given its earlier ancestry, it is appropriate to make a detour and to set up a problem in the theory of translation for whose solution a Charity Principle is invoked. It is convenient to secure this motivation by examining briefly the principal thesis of Chapter Two of Quine's *Word and Object* [Quine, 1960]. But a second reason is that if Quine is right about translation, this will affect importantly the logic of translation problems.

[3] See also [Baum, 1975; Johnson and Blair, 1983; Hitchcock, 1983; Govier, 1988; Johnson, 1981; Rescher, 1964; Parret, 1978].

[4] For cases in which it appears unreasonable to apply Charity see [Gabbay and Woods, 2001c; Gabbay and Woods, 2001a].

[5] Among the exceptions that come to mind are Rescher and Parret. According to Frans van Eemeren and Rob Grootendorst, Rescher acknowledges a deviation in his own treatment of charity from Davidson's, whereas Parret's is a Davidsonian notion [van Eemeren and Grootendorst, 1983, p. 189, n. 47].

9.3.1 Indeterminacy of Translation

In his attack on analyticity, one of Quine's alleged victims was the interdependent idea of synonymy or sameness of meaning. As with analyticity, Quine's charge against sameness of meaning is that it is defective in some way and that, whatever the details, being defective in this way precludes the use of synonymy in the derivation of scientific theories. But, if this is right, there is something wrong with translation as conceived by science and common sense alike.

'Chapter Two of Quine's *Word and Object* contains what may well be the most fascinating and most discussed philosophical argument since Kant's Transcendental Deduction of the Categories'; so says Hilary Putnam [1975b].[6] and he is right. Quine is talking there about 'radical translation and, as he proposes,

> There can be no doubt that rival systems of analytical hypotheses can fit the totality of dispositions to speech behaviour as well, and still specify mutually incompatible transactions of countless sentences insusceptible of independent control.

What does this mean?

Imagine that a field linguist is making the first attempt to translate into English the language of a people deemed to be culturally very different from the translator's own. The linguist will prepare a translation manual which Quine calls an 'analytical hypothesis'. Preparing such a manual in such circumstances is what Quine dubs 'radical translation'.

Quine assumes that the linguist knows how to recognize assent and dissent as speech acts in the host community. If this is right, he can locate truth functions in the host language and he can also discriminate occasion sentences, such as 'Here's a cup', for which certain stimulations will produce assent and others dissent. Standing sentences, such as 'Mayfair is in London', are such that once a speaker has been prompted to assent (or dissent) he prevails in it without 'further immediate stimulation from the environment' [Putnam, 1975b, p. 159].

Quine identifies the stimulus meaning of a sentence with the set of stimulations of a host's nerve endings which would induce assent to that sentence. Two sentences will have the same stimulus meaning if the same stimulations would induce assent to each. A sentence is stimulus analytic for a speaker if he assents to it no matter what the stimulation, and a sentence is stimulus contradictory for a speaker if he dissents from it no matter what the stimulation. Observation sentences that have intersubjective stimulus meaning.

Central to Quine's thesis is the notion of an analytical hypothesis. An analytical hypothesis is a general recursive function, f, from the set of all sentences of

[6] We follow Putnam's exposition, which is the best three-page account of the indeterminacy thesis we know of.

the host language to a subset of the set of all sentences of the visiting language. The function is constrained by three conditions.

(i) if s is an observation sentence of the host language, $f(s)$ is a visiting observation sentence and $f(s)$ has the same stimulus meaning for visiting speakers as does s for host speakers;

(ii) f commutes with the truth functions, i.e., $f(s \vee s')$ equals $f(s) \vee f(s')$, and so on;

(iii) if f is stimulus analytic (or stimulus contradictory) in the host language, then $f(s)$ is stimulus analytic (or stimulus contradictory) in the visiting language.

If the linguist happens to be bilingual in the host language, condition (i) is replaced by

(i*) if s is an occasion sentence of the host language, then $f(s)$ is an occasion sentence in the visiting language, and the stimulus meaning of s for the linguist is the same as the stimulus meaning of $f(s)$ for the linguist.

Conditions (i)–(iii) or (i*)–(iii) are Putnam's paraphrase of conditions (1)–(4) or (1*)–(4) of chapter two of *Word and Object*. Translations made in conformity with these conditions are made in circumstances that collectively exhaust the evidence for the translation in question.

The thesis of chapter two is this. Suppose that our linguist entertains two rival analytical hypotheses $f1$ and $f2$. There is, says Quine, no 'fact of the matter' as to whether the translations afforded by $f1$ are correct or as to whether the translations afforded by $f2$ are correct. Given that conditions (1)–(4) or (1*)–(4) exhaust all possible evidence for any translation, $f1$ and $f2$ could each fulfill (1)–(4) or (1*)–(4) and still 'specify mutually incompatible translations of countless sentences insusceptible of independent control [i.e., beyond the sway of *evidence*]'. *So there is no such thing as an objectively correct translation.* If this is right, it bears directly on the translator's abductive task. In conjecturing about how to interpret an interlocuter's utterances, the abducer is at liberty to conjecture that the meaning of the interlocuter's utterance is the same as some utterance of the abducer's language. But it is not, and cannot, be a condition of abductive success in such a case that the two utterances *acutally* mean the same.

But couldn't the translator become bilingual and from that position check to see which of $f1$ or $f2$ (or some other manual) *is* objectively correct? He can do the one, says Quine, but not the other. Becoming bilingual just means that the linguist's analytical hypothesis will be governed by (1*)–(4) rather than (1)–(4); but fulfilling (1*)–(4) will not block the specification of mutually incompatible translations of countless sentences beyond the sway of (1*)–(4). The fate of the bilingualist suggests to Quine that indeterminacy considerations apply equally to intralinguistic translation and interpretation.

How is the translator (or interpreter) to choose among rival analytical hypotheses? Suppose he specifies some further condition (5), and that conformity with conditions (1)–(5) produced a *unique* translation, that is, that there were one and only one translation that fulfilled those conditions. Such a translation would be rivalless, but it could not be said to be objectively correct, for there is no fact of the matter as to whether there is a fact of the matter that (5) is an objectively correct constraint.

Charity is one way of specifying condition (5). It is a way of discriminating rationally among rival analytical hypotheses, never mind that none could he said to be objectively correct. If we take Charity in Davidson's way we will hold that 'no sense can be made of the idea that the conceptual resources of different languages differ dramatically' [Davidson, 1984]. And

> Charity is forced on us; whether we like it or not, if we want to understand others, we must count them right in most matters.

Davidson concedes that this is a 'confused ideal'.

> The aim of interpretation is not agreement but understanding. My point has always been that understanding can be secured only by interpreting in a way that makes for the right sort of agreement. The 'right sort', however, is no easier to specify than to say what constitutes a good reason for holding a particular belief [Davidson, 1984, p. xvii].

The heart of Davidson's approach is that in interpreting another person it is necessary, if we are to make sense of him at all, to make an assumption of the following sort.

Proposition 9.8 (Charity) *Others are not noticeably deficient in holding true those things which we ourselves are adept at holding true. ('Oh, look, it's raining'.) But, by parity of reasoning, a corollary should be acknowledged.*

Corollary 9.8 (a) *Concerning those things that we notice we ourselves are not particularly adept at holding intersubjectively true (e.g., a theory in sharp contention), we should likewise suppose that the persons whom we are interpreting may not be significantly more adept about such matters than we. Where there is intersubjective disagreement at home it is not unreasonable to predict it elsewhere.*

This carries interesting consequences. It permits us to preserve *disagreement* in our interpretations when two conditions are met: first, when the interpreter takes heed of all available evidence and, thus, complies with (1)–(4) or (1*)–(4), and second, when he has reason to believe that the subject matter of the interpretation has had an intersubjectively bad track record here at home. (The theory of being *qua* being; or the morality of euthanasia).

There is a great deal to puzzle over in the modern history of, as we shall now say, Radical Charity (the Principle of Charity in the context of radical translation), ensuing from Wilson [1959]. We shall not persist with those puzzles here, for our prior question is what, if anything, does Radical Charity have to do with Charity as a *dialogical* maxim? The answer would appear to be 'Not much'.

If we consider the sorts of cases mentioned by Thomas and Scriven, it is plain that Dialogical Charity (as we shall now call it) is not the principle that fills for condition (5) on radical translation.

For if it *does* follow from Scriven's principle that the interpreter should furnish objectively correct interpretations, it is a presupposition of Radical Charity that it is precisely this that an interpreter cannot do for any sentence or discourse that outreaches the provisions of conditions (1)–(4) or (1*) to (4). Scriven's principle is not only not the same as Radical Charity but, if the implication we've been discussing holds, it is incompatible with it.

Suppose now that we have been radicalized by Quine's thesis about the possibility of objectively correct interpretation. In that case, we will say that our implication, *CorrectInt* ('Interpret your opponent or your interlocuter correctly'), does not hold. Will Radical Charity now serve as an explication of Dialogical Charity? We think not. Propositions 9.9–9.11 explain why.

Proposition 9.9 (Radical charity) *Radical charity has nothing to do with fairness. It does nothing to disenjoin the attribution of embarrassing or loony views to natives or others on grounds that this is a nasty, culturocentric thing to do.*

Radical Charity, applied at home, holds us to the view that

Proposition 9.10 *The community of our co-linguists is not in general to be counted as massively in error on those matters concerning which we are in relative epistemic serenity, that is, intellectually sincere and reflectively untroubled serenity. So Radical Charity gives no instructions for the individual case, except as noted below.*

Even so,

Proposition 9.11 (Negative charity) *Radical Charity does have a kind of negative providence in various kinds of contextually specifiable situations. Here the Principle is suspended, not applied. So, for example, the very fact that we are locked in a dispute with a co-linguist counts as some evidence that he is wrong, and our interpretation of what he says must preserve this fact or make a nonsense of our disagreement.*

Perhaps the charitable interlocuter will object that we got things muddled. He might point out that neither Radical Charity nor Dialogical Charity is a *carte blanche*, and that neither requires the suppression of disagreement in individual cases. This is quite right. Radical Charity posits uniformities in human nature that would justify the assumption of agreement in interpretation in the *general* case. But we want to suggest that not only does Dialogical Charity permit disagreement-preserving interpretations in particular cases, it also permits them and counsels against them in the general case as well or, more carefully, for the general run of cases for which the Dialogical Principle has apparently been designed. If this is right, the two Principles remain significantly different.

It is well past time for us to cease our preoccupation with enthymemes, and to concentrate our attention upon interpretation problems, which bear only contingent linkages to enthymeme resolution. Interpretation is the natural antidote for understatement. It also bears stronger apparent affinities to abductive reasoning. The general form of resolutions of understatement problems is this.

1. The interlocuter makes an utterance U, which in a certain respect or respects is not explicit.

2. The way in which U itself is not explicit is such as to move the interpreter to think that U is not what the utterer intended to communicate, or intended his interpreter to take from U.

3. Quine's view is that the interpreter then tries to put himself in the shoes of the utterer. He seeks empathy with him. He asks himself, "If I were to say what the utterer did say explicitly, and if I made this utterance from the perspective I believe him to have brought to bear on the making of that utterance,[7] what would I have likely intended to have had my interpreter take me as having said?"

In its simplest terms the solution under review is this: *What would I in his place have intended U to be taken for? That, defeasibly, is what the utterer of U also intended and what he said implicitly.*

[7] This clause leaves it open to an interpreter to interpret an argument successfully even though he himself from his own point of view would not be making the argument be attributes to interlocuter, but rather would have intended something else. Thus an evolutionist might be able to give a creationist spin to an argumental utterance, which, had he himself uttered it would have carried a Darwinian spin. So the factor of empathy might strike us as essential to this account. For an appreciation of the centrality of empathy in analogical reasoning, see [Holyoak and Thagard, 1995].

9.4 Is it Abduction (Again)?

It remains to say why the interpretation of understatements may involve an exercise in abduction. The problem is triggered by a phenomenon P in which the interpreter takes an utterance U of an interlocuter to be implicit or to contain an inarticulacy, knowledge of which is necessary to understand what the interlocuter was intending to say. Thus the phenomenon P is one in which the utterer intends to say something that doesn't get said by U, and yet the utterance of U is *not defective* on that account.

What needs explaining is how an utterance of U fails to say what its utterer intended to say, and yet in saying U the interlocuter is not simply guilty of not saying what he intended to say.

The general conjectural explanation is that implicit in the utterance of U there is an inarticulacy which, once repaired, would produce an utterance U^* which would, just so, say what the utterer intended. The specific conjectural explanation is the expansion of U produced by the empathetic thought-experiment that we have been discussing. Thus what explains that U fails to say what the utterer of U intended to say, and yet U is not a defective utterance in that very regard, is that the utterance that explicitly is U is also implicity U', and U' does, just so, say what the utterer intended

The presumed abduction we have been describing is subject to the regulatory control of various sublogics. The logics of generation and engagement attempt to say how conjectures which explicitize inarticulacies get chosen. In the example we have been discussing the selection test is by way of our empathetic thought experiment. It then falls to the logic of hypothesis-discharge to set forth conditions which would determine whether the explicitization of the inarticulacy had been done accurately or reasonably. For the kind of case we have been considering, we can give a rough indication of how this works. In his conversation with the utterer of U, the interpreter has interpreted U as U'. To test for accuracy, it suffices in what remains of their discussion or in subsequent additional discussions, for the interpreter to guide his future remarks by the assumption that the utterer's position is in fact U'. The straight way of determining this is by the interpreter's explicitly attributing U' and waiting to see what happens. ("Given that you hold that U', don't you think that so-and-so?".) If there is no move to reject the attribution, then, as Quine says, there is evidence of "successful negotiation and smooth conversation," which is evidence of communicative success, and hence is evidence of accuracy of interpretation [Quine, 1992, pp. 46–47]. Less directly, the interpreter can introduce into the conversation propositions which the utterer could be expected to accede to were it indeed true that U' is a correct interpretation. These propositions could be immediate or fairly obvious consequences of U' but not of U.

9.5 Constitutional Inarticulacies

Perhaps we should be struck by how difficult it is to shed light on how even rather simple looking understatement problems are solved. In the approach we have sketched the interpreter is an abductive agent. But in proposing that account we ourselves are abductive agents. Our account is highly conjectural, and it is offered for the explanatory contribution it makes to some genuinely puzzling phenomena. Of central importance is our ability to distinguish between genuinely defective utterances and those that are defective merely on their face and not otherwise. The other puzzling phenomenon is the comparative ease and accuracy with which individual cognitive agents are able to interpret understatements. Our conjecture is that the interpreter puts himself in the utterer's place and determines, so placed, how he himself would intend to be understood in uttering the original speaker's utterance. Underlying this conjecture are numerous further assumptions, which we have not had the time to explore. One is that the idea of a thinking agent's place is legitimate and theoretically load-bearing. Another is that empathy permits, within limits, the occupancy of alien places, of places that aren't the agent's own. A third is that one's own articulacies are self-interpretable in a fairly direct way, involving neither change of place nor empathic engagement.

9.5.1 Inarticulate Understanding

It is easy to see that each of these assumptions could run into the heavy weather of harsh challenge, especially the third. It is an assumption that tends to conflate understanding with articulacy. More promising is the idea that even in one's own case, understanding inarticulacy is not giving an articulation of it. This leaves the door open for a notion of understanding inarticulacy in ways that themselves are inarticulate. If so, such inarticulacies are constitutional rather than circumstantial. They are inarticulacies which we can't articulate, yet which we manage to understand.

Upon reflection, even if our conjectural sketch of how interpreters handle understatement is true as far as it goes, it is unlikely that the interpreter, as he negotiates the routines of placement and empathetic engagement, is able to articulate what he is doing. In these respects, the interpreter of understatements is like the maker of them. No matter how explicit their utterances, discussants do not in general have command of the precise conditions under which their utterances are given form and expression.[8]

[8] We might mention in passing visual factors in hypothesis formation in contexts of visual inference in archaeology. As detailed by [Thagard, 1995], such hypothesis can have highly complex structures, which are nonsentential. They are nonlinguistic and subconscious.

This suggests constitutional inarticulancy on a rather grand scale. What would account for it? It has to do with constraints imposed by our cognitive economy, especially by the hostility of consciousness toward informational abundance. Or so we conjecture.

In considering the question of the abductive character of linguistic understanding, it is necessary to return once again to the requirement that abduction problems originate and terminate in conditions of reasoner-ignorance — what we have been calling the nescience condition. In the beginning, an abduction problem is occasioned by the fact that the reasoner's cognitive target T cannot be hit with is present epistemic resources K. For this epistemic shortfall to exist, it is not necessary (and often isn't true) that the agent be in utter ignorance about what T's fulfillment might consist in. Presupposed by his problem is the central fact that hitting T requires doing so at a certain epistemic level and that K's own epistemic credentials preclude K from meeting T at that level. As we remarked earlier, the fact that the nescience called for by our condition is always a matter of degree, failing to achieve an abductive target can quite routinely be a near miss, what with the degree of K's epistemic virtue falling only slightly below that standard required for the attainment of T. Of course, the gap can be very large and, as in some of the more dramatic of the documented cases, is very large indeed. This allows us to say that the wider the gap, the more "originary" the solution is required to be. But unless "originary" just means "new" and "new" just means "$\notin K$", we take Peirce's insistence on originary thinking for all of abduction to have been a mistake.

The abductive status of understatement resolution turns directly on how we decide to understand the approximate universality of understatement, according to which it is not less than generically true that to state is to understate. So substantial a generality suggests that the mechanisms for the interpretation of understatement are largely those required for the interpretation of direct meaning. There is an abundance of empirical evidence suggesting that the competent interpreter achieves the bulk of his data-interpretation by instantiation of rules and conventions, of which he can be said to possess enough of an epistemic grasp for them to function satisfactorily in the ensuing interpretations. If this is so, then "John has children" would be a part of an interpreter's understanding of "All John's children have measles" by deployment of the requisite presupposition rule. Similarly, concluding that he is free to help himself to some of Harry's beer from Harry's utterance of "There's beer in the fridge, if you want some", is a matter of engaging the requisite social convention. Seen this way, these are resolutions that lie within the epistemic levels required for satisfactory interpretational outcomes. Accordingly, they are not abductive resolutions unless instantiation is itself abductive. In fact, it appears not to be the case that classifying an observed object as an observable of kind C is *intrinsically* an epistemically lesser task than noting that things of the C-kind are things of the D-kind. In any case, there are large ranges of cases in which it is epistemically no "harder" to see that Jasper is a crow than to be aware that crows caw quite loudly. Saying so certainly doesn't rule out exceptions that are rather well-understood. The generalities about Ewing's sarcoma are, tragically, well-known; but there are cases in which a diagnosis of this cancer can be rather tricky. Accordingly, contrary to what we have been assuming

Proposition 9.12 (Interpretation and abduction) *Determining the presumptive meanings of texts or utterances is not intrinsically an abductive undertaking.*

Our present reflections give us occasion to revisit the idea of common knowledge. When we discussed common knowledge in chapter 7, we emphasized the presence, and abductive importance, of generic and normalic propositions, and we conjectured that, for the most part, they arise by hasty generalization on natural kinds and also by hearsay. But, as our later discussion of Peirce's guessing instinct reminds us, our common knowledge is also well-stocked with what is known innately, including (to the extent that they *are* innate) the rules — transformational and otherwise — that undergird linguistic competence. To this should be added, on an individual to individual basis, the conventions, causal associations and other lawlike linkages that arise from his intercourse with the world and are retained in his memory. Accordingly,

♡ **Proposition 9.13 (Interpretation and common knowledge)** *The basic structure of an agent's identification of presumptive meaning (in languages in which he is fluent) is by instantiation on the contents of his common knowledge.*

If this right then, given Proposition 9.13, we have it that an agent's grasp of presumptive meanings is not by and large the the result of solving abduction problems. This helps us appreciate the conditions under which textual and discourse interpretations happen to require abductive resolution.

Proposition 9.14 (Abductive cases) *If an agent's agenda is to grasp the meaning of a text or of a speaker's utterance, and doing so is not possible by combining those data with what is instantiable from his common knowledge, then in the absence of particular indications to the contrary, the agent's interpretation problem will be abductive.*

No doubt it will not have gone unrecognized that for the most part a fluent's interpretation of linguistic data is a cut-to-the-chase achievement. In the general case, such interpretations are achieved very quickly and in the absence of phenomenological and behavioral indications of either conjecture or effort. In our previous discussion of cut-to-the-chase phenomena, we reflected on classes of responses to abductive triggers in which the shedding of both irrelevant and implausible possibilities was also achieved in the absence introspective or behavioural trace. Such cases aren't disqualified as abductive, provided resolution is by inference of a proposition asserting that a given such possibility may be conjectured. But it is easy to see that if homing in on the relevant and the plausible can be so typically cut-to-the-chase achievements, it is not unreasonable to suppose that going straight to a final answer is often achieved without prior conjecture, and hence without occasion or necessity to discharge it. In such cases, reading a situation is like reading a paragraph in the *Times*. They are situations in which abduction has no role. So it is necessary to acknowledge that

Proposition 9.15 (Non-abductive responses to abduction triggers) *Not every cognitively successful response to an abductive trigger is itself a case of abductive resolution.*

Corollary 9.15 (a) *Proposition 9.15 covers two kinds of case. One is when the cut-to-the-chase conclusion is derived with a force strict enough to place it in an* extension *K^* of the agent's original knowledge-set K (in which case the ignorance condition is failed). The other is one in which the selected hypothesis is drawn non-presumptively.*

Corollary 9.15 (b) *The second case of Corollary (a) reminds us, even so, that not every presumptively derived proposition is a case of abductive resolution.*

This would be a good place also to revisit our earlier suggestion that, since memory searches and searches of possibility spaces are in various contexts phenomenologically indistinguishable, it is reasonable to suppose that, rather routinely, the process of abductive resolution begins in a memory search, and that abduction proper is reserved for the juncture at which the memory search *fails*. (For it is then that the agent's cognitive system becomes "aware" that the requisite target cannot be reached via what the agent currently knows.) No doubt, there are substantial ranges of cases in which the activation of memory precedes the engagement of the mechanisms of abduction. But it would be a mistake to accord to memory either strict temporal precedence over abduction, or conceptual priority. Cognitive economies would be achieved if both memory searches and hypotheses searches happened concurrently.

Neither should it be naively assumed that remembering is intrinsically an easier task than abductive conjecture. There will be considerable variation here, but as the empirical record suggests, there are types of abduction problem, and kinds of circumstances in which they arise, in which imagination is more readily available than memory. Accordingly, we must amend our earlier suggestion of the precedence of memory over hypothesis.

Proposition 9.16 (Abduction as an aid to memory) *There are occasions in which an (apparently) successful resolution of an abduction problem $C(H)$ induces the memory that H is the case. Since the recollection of H implies that it was in the agent's K with respect to which the original problem arose, apparently abductive stimulation of such memories is* not *abduction.*

9.6 Visual Abduction

Notoriously, it is Peirce's view that "[a]bductive inference shades into perceptual judgement without any sharp demarcation between them" [Peirce, 1955, P. 304].[9] Peirce extends the same latitude to the experience of emotion.

> Thus the various sounds made by the instruments of the orchestra strike upon the ear, and the result is a peculiar musical emotion, quite distinct from the sounds themselves. This emotion is essentially the same thing as a hypothetic inference, and every hypothetic inference involved the formation of such an emotion [Peirce, 1931–1958, p. 2.643]; *cf.* [Oatley, 1996].

Oakley and Johnson-Laird [1987] develop a cognitive theory of the emotions on this same abductive model of Peirce. O'Rorke and Ortony in [1992] and [1995] put to the same purpose computational devices implemented in PROLOG, AbMa and the set-up of situational semantics. The question of whether semantic interpretation is abductive turns on whether an agent's knowledge of what some linguistic data mean or what the producers of them meant in giving them utterance attains a degree of epistemic virtue than that exhibited by the interpreter's command of the data themselves. Does a fluent speaker know less well that John has children than he does that all John's children have the measles (when he does indeed know this)? Does a competent speaker of English know less well that he is free to have some of Harry's beer than he knows the meaning of Harry's utterance, "There's beer in the fridge, if you'd like some."? If the answer is "No", the interpreter's drawing of these presumptive meanings is not an exercise in abduction. There are numerous cases of non-abductive presumptive-meaning interpretation, and also cases in which the opposite is true. When Mary pointed out to her son that there was no more wine, how likely is it that an over-hearer of that utterance would know that the mother of Jesus was asking for a *miracle*?

Is non-abductive semantic interpretation just like, or for the most part like, visual identification? For this to be so, it must be the case, at least for the most part, that knowing what one sees involves epistemic levels as high as those attained by his recognition of the raw data of sensory stimulation. In other words, is what psychologists call "visual cognition" as non-abductive as presumptive-meaning determination (or, for that matter, direct-meaning determination)?

This is a vexed and contentious issue, whose satisfactory development would take us well beyond the ambit of the present work. We shall, even so, give a sketch of how the problem is structured and of where we think the solutions are likely to lie. The basic question is whether, when an observer knows what he sees on the

[9] Other perceptual inferentialists include [Kant, 1929; Bruner, 1957; Fodor, 1983; Gregory, 1980], and [Josephson and Josephson, 1994].

basis of seeing it, what he knows is of a lesser grade than the knowledge possessed of the data interpreted by his visual mechanisms. Two notorious philosophical problems bear on this issue. One is the controversy over *sense-data.* The other is the controversy over the relevance of the *underdetermination* thesis. We turn to these issues now.

In the heyday of logical positivism, it was taken as given that perceptual knowledge involved two distinct levels of epistemic attainment. In the first phase, the percipient is presented with sense-data, which are the rudiments of sensory experience apprehended immediately and incorrigibly. In the second phase, these mental phenomena are given a semantic interpretation. Thus, when Harry sees the snow drifts piled up in the street below, he infers that this is so by conceptualizing the raw sensory data of his perception in the appropriate way. Since the sense-datum account is not itself phenomenologically confirmed in normal visual apprehension, the phase-one claim that the perceiver has an epistemically surer command of what he doesn't see than of what he does see is cause for hesitation.

A related consideration is the underdetermination thesis as applied to perception. What gives it a certain appeal is that it tries to make do without relying upon the hypothesis of sense-data. In their place, the perceiver is reckoned to be the subject of neural stimulations that the subject's perceptual apparatus then reconstructs as the perception of an object in the world (to speak loosely). The perceptual underdetermination thesis likens the construction of perceptual objects to the construction of scientific theories. In the case of theories of the observable, the underdetermination theorist notices that up to arbitrarily many different and possibly incompatible theoretical articulations are logically consistent with the (interpreted) observational data. So it can be said that those data underdetermine what their best theoretical articulations are. In its application to perception, the underdeterminationist construes neural stimulations as analogues of interpreted observable data, and the perceptual reconstruction of those stimulations as analogues of the theoretical articulation of interpreted observational data. Since it is assumed that the underdetermination theory holds for all observation-based theories, it is concluded that here, too, a subject's nerve-end hits underdetermine what the subject sees. Here, too, the reasoning would be that since possession of those nerve-end hits is logically consistent with the subject's seeing something different from, and even compatible with, what he does in fact see, it must be concluded that what the subject sees is not *determined* by his requisite neurological excitations.

If the underdetermination thesis held for visual perception, there would be occasion to ask whether the resulting perception is a matter of inference from neural turbulence. If we thought it were, there would also be room to consider whether such inferences were typically (or ever) abductive. The answer is "No". Even if the undetermination thesis were true, it doesn't follow that sensory stimulations don't structure observations causally. Consider a case. Let $s_1, \ldots, s_n$ be the neurological stimulation in which agent's observation of the falling apple is embedded. The fact that it is not a logical contradiction to say that those stimuli correlate with the observation of a balloon rising gives no comfort to the notion that seeing the apple fall is achieved by inference, still less that it is inference to a conjecture.

Peirce's thesis concerning the abductive character of perception makes no serious headway in the logic of "up above". Its sole chance lies in the logic of down below. This could happen in two ways, one of which is more plausible than the other. The implausible option would be one in which the observer of the falling apple has greater *tacit* knowledge of his neural stimulations than he has express knowledge of the apple's fall. A less implausible alternative would be one in which the physiology of the "down below" construction of observations from nerve-end hits is one that resembles the structure of abduction problems "up above". See here [Nersessian, 1995; Nersessian, 1994]. But even so, it would be a stretch to say that, in seeing the apple fall, Harry was solving an abduction problem, as opposed to saying that in their observation-generating behaviour, Harry's neuroperceptual devices *were* operating in a way that simulates abductive behaviour in beings like us.

A good treatment of visual and temporal abduction may be found in [Magnani, 2001a, chapter 5]

9.7 Empathy

We have been suggesting that empathy may play a role in the understanding of inarticulate utterance. We now examine this idea in somewhat greater detail. The concept of empathy is dominated by the metaphor of place. One party empathizes with another when he puts himself in her place (or her shoes). Doing so allows him to see things from her perspective (or frame of mind). Perhaps the best non-geometric characterization of empathy is this.

Definition 9.17 (Empathy) *A cognitive agent X empathizes with O to the degree that X knows what it is like to be an O.*

It is well to note the sheer range of things denoted by 'O', attesting in turn to the versatility of empathy. Harry may empathize with Sarah; he may do so with respect to some particular state or set of circumstances that Sarah is presently in, or he may do so quite generally. Harry may also empathize with women (or children, immigrants, untenured assistant professors, the Catholic Church, perhaps even some the forces that precipitated World War I). In each case, there is something that Harry knows what it is like to be — individuals in a situation s, individuals whatever their situation, groups, institutions, the energies of history. If some writers on analogy are to be believed, it was also possible for Bohr to form some kind of empathetic association with the interior of an atom, and thus, knowing what it's like to be the interior of an atom, came to grasp that being an electron in motion is like being a planet in motion, albeit differently (which makes empathies of this sort analogical).[10]

In foregoing chapters we emphasized the tie between what is characteristic of something and what something is like. An appreciation of what things are like we saw as playing a central role in hypothesis-selection. We proposed as a defeasible rule the rejection of possibilities that run counter to how things are — to what things are like. In rejecting the hypothesis that the reason the door is ajar is because it was made so by Sarah and her early return home, Harry depended on his knowledge of what Sarah is like. This is not the same as knowing what it's like to be Sarah. Knowing what Sarah is like is not empathy. Harry can know what Sarah is like as concerns her weekday homecoming schedule, and other things as well. But he need not have the slightest clue about what is is to be Sarah. "You just don't understand me!", Sarah might expostulate. As in confirmation of her complaint, Harry could well know that it is part of what Sarah is like to press this grievance endlessly, but not know what it's like to be someone who has occasion to press it. It is also true, however, that if Harry had *no* idea of what Sarah was like it is not very probable that he could know what it is like to be Sarah; but the converse reflects no such connection. Accordingly,

Proposition 9.18 (Empathy and characteristicness) *Knowing what is characteristic of (a) Y is standardly involved in knowing what it's like to be (a) Y, but is not a necessary or sufficient condition of it.*

The difference between knowing what Sarah is like and knowing what it's like to be Sarah establishes that it is one thing to know what is characteristic of a person and another to empathize with her. There is a difference, in turn, between empathizing with Sarah and sympathizing with her. There are two points on which this difference turns. Sympathy is intrinsically compassionate; empathy is not. Having compassion for someone does not require that it be known what it is to be him. A second point of difference is the range of objects of sympathy is a good deal more circumscribed than that of the objects of empathy. To say it as briefly as possible, Y can be an object of one's sympathy only if it is possible in principle to feel sorry for Y. Actors are adept at empathetic attachment. There is no such condition on empathy. A good historian may bring himself to know what it is like to be Hitler without a flicker of affection or fellow-feeling. It is said that what made Anthony Hopkins so convincing in the role of Hannibal Lecter was that Hopkins was able to get himself to appreciate what it is like to be Hannibal Lecter. This is commonly said of our best actors. They have the facility for putting themselves in the shoes of the characters they portray. But it is also commonly said, often

[10] No doubt there are limits. Perhaps no one knows what it is like to be the square root of 2, Caesar's right eyebrow or even, one daresay, a bat. (Concerning which, see [Nagel, 1974].)

by those very actors, that it is psychologically difficult — some say impossible — to put oneself in the shoes of even a monster like Lecter without having some tenderness of regard for what it is to be that way. If this is right, it appears that although

Proposition 9.19 (Empathy and sympathy) *Empathy and sympathy are conceptually disjoint*

nevertheless

Proposition 9.20 (Contingent links) *If Y is an object of empathy that is also a possible objection of compassion, empirical findings suggest that in empathizing with Y there is some tendency also to sympathize with Y, where Y's situation is fitting occasion for compassionate response.*

In the idea of putting oneself in another's place, the notion of place is metaphorical. Shall we say the same of its non-geometrical analogue, "knowing what it is to be (a) Y? If so, where in this form of words does the metaphor reside? Apart from the word "know", nothing else stands out as a plausible candidate. When Harry knows what it's like to be Sarah, it is hardly credible that the name of Sarah is a mere figure of speech, or that "like" is a *façon de parler*, or that "to be" is not the copula plain and simple. And yet if it is the knowledge-component in which the aspect of metaphor inheres, how does this come to be so? It is a difficult question, and yet one with respect to which the Hopkins – Lecter example may again prove helpful. When actors tell us that they have succeeded in putting themselves in the place of the characters they play, they often characterize their empathetic success by saying that they were able to *identify with* them.

This helps explain the extent to which nuances of sympathy so frequently attend empathetic attachment, since the extent to which one has made oneself another is also the extent to which one is able to feel for the other at least aspects of the regard one feels for oneself. The idiom of identification also bears on the metaphorical character of empathetic knowledge. In knowing what it is to be Hannibal Lecter, Anthony Hopkins identified with him. Actors sometimes say of such identifications things like, "And so, for two and a half hours eight times a week in the Royal Haymarket Theatre, I *was* Falstaff". These identity claims are clearly metaphorical. Accordingly,

Proposition 9.21 (Empathy and identity) *To the extent that empathy with Y involves identification with Y, knowing what it is to be (a) Y is knowledge with a metaphorical aspect.*

9.7.1 Discourse Empathies

Consider now the empathies postulated between parties to a discourse, each an interpreter of utterances of the other and of the other himself. Take the case of determining what Harry meant in uttering "There's beer in the fridge if you want some." The suggestion that I know by empathy what Harry meant requires us to make sense of the idea that I achieve this understanding by there being something or other that I know what it's like to be. But what? It is certainly too much to expect that in order to know what Harry meant I must know what is is to be Harry. What is required is that there is some aspect of what Harry is involved in that I know what it is to be. Since what Harry is involved in is the communication of what he means by the making of an utterance that does not mean it, the empathy thesis is correct only if we can see our way clear to agreeing that in understanding what Harry meant I knew what it was to be a conveyor of that meaning via an utterance of those words. In forging that connection, the element of identification would appear to be essential. For it is indeed essential to the thesis that we are now examining that in attributing to Harry the invitation to drink his beer, I reason autoepistemically, as follows: Since I know that I would have intended had I uttered "There's beer in the fridge if you want some", I may attribute the same to Harry because in the matter of this very utterance I *identify* with Harry. Since I know what I would mean by it, and since, in the requisite sense, I *am* Harry, then what I would mean by it is what Harry does mean by it.

No doubt it will be agreed by all that this goes much too far. In empathetic identification the flow is from empathizer to the object of his empathy. Anthony Hopkins becomes Hannibal Lecter. Anthony Hopkins allows himself to be taken over by Lecter. But in knowing what Harry meant, if there were *any* element of identification involved, it would flow in the other direction. Because Harry is, in the requisite sense, *me* and since if I were to utter what Harry uttered I would mean to invite my addressee to have some beer, that is what Harry also means. A little reflection shows that even this "reverse identification" is greatly overblown. Given that I do know what I would mean by it, it suffices for the attribution of it to Harry that, when Harry utters "There's beer in the fridge if you want some", Harry is behaving like me. Undergirding that assumption is nothing that resembles, or requires, that I identify with Harry or Harry identifies with me. What is required is the assumption that if Harry is speaking my language, we are bound by the same conventions by and large, and in knowing what I would have meant by it, I am thus enabled to know what Harry meant in uttering what he did utter. At the heart of it all there is a judgement of likeness, but it is not one of empathetic attachment. It is rather a judgement of what is characteristic. It is characteristic of an utterance of "There's beer in the fridge if you want some" for the utterer to be inviting the addressee to help himself to the beer. It is this knowledge of what such utterance is like that explains my self-knowledge and my knowledge of Harry. I know what I would have meant by it because I know that I am a speaker of English, hence that I know that since I know what English is like, I also know that in uttering that hence I would be extending that invitation. In this regard, I know no more or less about Harry. In knowing that he is a speaker of English, I know that he knows what English is like, I know that what Harry means by it is what I would mean by it. I know this because Harry and I are alike in knowing what English is like. Knowing this, I know that Harry is bound by the same conventions that I am.

Proposition 9.22 (Interpretation without empathy) *In the general case, the interpretation of utterance-meaning and utterer's-meaning is not a matter of empathy.*

That is the general case. There are, of course, special cases. Quine on analyticity comes readily to mind. Here it is not enough to determine either Quine's complaint or the case he makes for it simply to know what English is like and that Quine is speaking English. It undoubtedly helps to know what Quine is like. So, even here, the difference is not between interpretations that don't and interpretations that do require empathetic connection. The difference lies in the level and range of characteristicness the interpreter must have some command of. For not only must he know what is characteristic of English, he may also need to know what is characteristic of Quine. But this is not empathy.

It would be going too far to suggest that empathetic attachment is never a necessary ingredient *in* the successful negotiation of — especially intractable — interpretation problems. But what requires emphasis is the distance that separates such cases from the hermeneutical norm.

9.8 Semantic Space Interpretation of Texts

It is well-known among AI researchers that there are no easy ways in which to automatically unpack the propositional content of texts. This is a problem for the mechanization of textual interpretation, but it is also an indication of how small a part of the problem is encompassed by enythemematic interpretation. In the present section, we bring to bear on the general issue of interpretative abduction some recent developments in research on semantic spaces. It represents yet another attempt at what might pass for a logic of down below. We here follow results reported in [Bruza *et al.*, 2004].

The problem that motivates the semantic space approach is that when AI researchers build systems that are able to reason over substantial texts, the deployment of techniques of propositional representation of the knowledge embedded in the text fails to achieve the objective in a satisfactory way. Even so, on the assumption that texts embed knowledge of some sort, and on the further assumption that textual interpretation and inference somehow "gets at" that knowledge, it is reasonable to postulate a knowledge representation capability of some kind. In the approach of Bruza and his colleagues, highly dimensional semantic spaces are put into play. Semantic spaces have a good record in the cognitive interpretation of human information processing, and they offer attractive promise as a kind of computational wherewithall that simulates abductive behaviour in humans. Another attraction of semantic spaces is the size of their representational capacities. Bruza and his colleagues point out impressive degrees of knowledge representation success in relation to quite large texts. In one example, a substantial body of Usenet news (160 million words) responded well to the system's knowledge representation function [Lund and Burgess, 1996; Burgess *et al.*, 1998].

Abduction introduces something new. Any mechanical system adaptable for abduction must take this fact into account. To this end, two mechanisms are required. One uncovers implicit associations. The other computes them. A test case for this technology is the replication by automatic means of an abduction by Donald Swanson that fish oil is effective in the treatment of Raynaud's disease.

HAL (the Hyperspace Analogue to Language) constructs lexical representations in a high dimensional space that score well against humanly generated representations [Lund and Burgess, 1996; Burgess *et al.*, 1998]. HAL's space is an n x m matrix, relative to an n-word vocabulary. The matrix trains a window of length l on a text, at a rate of one word at a time. (Intuitively, windows function as *contexts*). Punctuation, and sentence and paragraph boundaries are ignored. All words in the window are assumed to be co-occurrent in degrees that vary proportionally to their distance from one another. In a HAL matrix, row and column vectors are combined to produce a unitary vector representation for the word in question. Table 1 displays part of the unified HAL vector for the word in question. Table 1 displays part of the unified HAL vector for the word "Raynaud". It is computed by directing HAL's attention to a set of medical texts got from the MEDLINE collection. In this representation, a word is a weighted vector with lexical alternatives serving as its dimensions. In this case, weights give the degree (or strength) of association between "Raynaud" and other lexical items caught by the moving window. The higher the value of the weight, the greater the degree of co-occurrence with "Raynaud", assuming contextual invariance (see table 9.1).

HAL is a type of computational model known as *semantic space* systems [Lowe, 2001, p. 200] and [Patel *et al.*, 1997; Sahlgren, 2002; Lowe, 2000]. Semantic spaces operate geometrically rather than propositionally. They are simplified adaptations of the notion of a conceptual space originated by Gärdenfors [2000]. In conceptual spaces knowledge is represented dimensionally. Colours for example, have a three-dimensional representation: hue, chromaticity and brightness. On this approach a colour is a three-dimensionally convex region of a geometric space. Red is one such region. Blue is another. In Gärdenfor's account there is a principled link between ontological items such as colour-*properties* and mental items such as colour-*concepts*. Integral to this mapping is the concept of domain. A domain is a class of *integral* dimensions, which means that a value in one dimension either fixes or affects the values in other dimensions. The colour dimensions are

Raynaud	
Dimension	Value
nifedipine	0.44
scleroderma	0.36
ketanserin	0.22
synthetase	0.22
sclerosis	0.22
thromboxane	0.22
prostaglandin	0.22
dazoxobin	0.21
E1	0.15
calcium	0.15
vasolidation	0.15
platelet	0.15
. . .	. . .
platelets	0.07
blood	0.07
viscosity	0.07
vascular	0.07
. . .	. . .

Table 9.1 Example HAL representation

integral in the sense that colour-brightness affects both chromaticity (or saturation) and hue.

The geometric orientation also figures prominently in theories of information flow in the manner of Barwise and others. [Barwise and Seligman, 1997]. In such accounts, inferential information content is defined for real-valued spaces. Brightness is represented as a real number between 0 (white) and 1 (black). Integral dimensions are construed as observation functions that specify how a value in a given dimension affects a value in another dimension. Here the represented items are points, whereas in Gärdenfor's approach they are regions.

We can now see that a HAL-representation approximates to a Barwise and Seligman state space in which dimensions are words. For example, a noun phrase is such a point. The point represents the state of the context of the passage that is under examination from which, in turn, the HAL space is computed. If the lexical sample changes, the state of that noun phrase may also be altered, which is something of a setback (but see below). Even so, HAL has a good track record for what Lowe and his colleagues call "cognitive compatability". Another virtue is that HAL spaces are algorithmic. "Yes," one might say, "but can HAL do abduction?"

In Gärdenfor's approach, inference need not be considered in exclusively symbolic terms. Symbolically represented inference is for the most part a linear process and, in most of the standard treatments, a deductive one. In a conceptual space model, inference is a matter of associations based on semantic similarity, where similarity is given a geometrical rendering within a n-dimensional space. Thus the conceptual approach to reasoning has an explicitly geometrical character. This is a promising candidate for what we have been calling reasoning "down below", typified by irrelevancy-evasion in general and cut-to-the-chase abduction in particular. The promise lies in prospects of a computationally tractable logic of hypothesis-generation. Bruza and his colleagues conjecture that hypotheses are generated computationally on implied associations in semantic spaces. The factor of implicity is thought of as counterpart of Peirce's notion of the "originary" aspect of abductive reasoning. New hypotheses are realized by way of computations of information flow in semantic space. In an example from [Song and Bruza, 2001; Song and Bruza, 2000; Song and Bruza, 2003], *penguin, books* $\vdash$ *publisher* expresses that the concept publisher is transmitted informationally by the conceptual composition of penguin and books. The concept of publisher thus flows informationally from those of penguin and books. These and other kinds of information flow are fixed by an undergirding state space produced by HAL, Information flow comes in degrees which are functionally related to the degree of the inclusion between the requisite information states (i.e., over the HAL vectors). When inclusions are total, information flow is maximal.

We now define information flow somewhat more formally. Let $i_1, \ldots, i_n$ be concepts. Then c_i is the HAL representation of concept i, and δ is a threshold value. $\oplus_c i$ is the composite of $c_i, \ldots, c_k$; it is therefore a combined mode of representing a composite concept. Inclusion is denoted by $\subset$.

Definition 9.23 (Information flow in HAL)

$$i_j, \ldots, i_k \vdash j \;\text{ iff degree }\; (\oplus_c i \subset c_j) > \delta$$

In our discussion here, information is computed from just one term. So $\oplus c_1 = c_1$.[11]

The degree of inclusion is got by normalizing the score which is computed of ratios of intersecting c_i and c_j to the number of properties in c_i. Accordingly,

Definition 9.24 (Degrees of inclusion)

$$\text{degree}\,(c_i \subset c_j) = \frac{\sum_{P_1 \in (QP_\delta(c_i) \wedge QP(c_j))}}{\sum_{P_k \in (QP_\delta(c_i))}}$$

Intuitively, the more an inclusion relation includes, the more it is an inclusion relation. Definition 9.24 takes note of this by requiring that most of the properties represented by c_i (the "source" concept) also crop up among the properties represented by c_j. The properties covered by the source concept are defined the threshold δ. So, for example, in texts in which query expansion terms are derived automatically by way of information flow determinations, best results were achieved by setting δ to the average dimension weight in c_i [Bruza and Song, 2002].

Consider a case in which j has zero weight in c_i. What this means intuitively is that i and j have no co-occurrence in any window in the construction of the semantic space. But this does not preclude information flow from c_i to c_j. In such a case, the flow of information from c_i to c_j is called *implicit information inference*, and is of obvious interest to abduction theorists.

Information flow models have had a good record in automating query expansion for document retrieval. Effective query expansion is a matter of inferring expansion terms relevant to the topic of the query. Bruza *et al.* [2004] suggests that query expansion can be understood abductively. The task is to abduce terms relevant to the topic of the query.

> Terms which exhibit high information flow from the given query can be considered collectively, as furnishing explanatory hypotheses with regard to the given query, modulo the underlying semantic space [Bruza *et al.*, 2004, p. 104].

[11] See here [Levy and Bullinaria, 1999].

9.8.1 The Raynaud-Fish Oil Abduction

In the 1980s a librarian named Don Swanson made a chance discovery by linking together two different on-line medical sites, one having to do with Raynaud's disease and the other dealing with fish oil. As Swanson subsequently observed, "the two literatures are mutually isolated in that authors and readers of one literature are not acquainted with the other, and vice versa" [Swanson and Smalheiser, 1997, p. 184]. Swanson's discovery turned on what we might call *intermediate terms* or B-terms. If we take A to represent "fish oil" and C to represent "Raynaud", then the implicit link between them was indicated by groups of *explicit* links A-B and B-C [Weeber *et al.*, 2001]. The B-terms used were "blood viscosity", "platelet aggression" and "vascular reactivity". (See again Table 9.1.) While A-B and B-C links were reported in the two disparate literatures, there is in neither any explicit link $A \rightarrow C$. $A \rightarrow C$ Swanson characterizes as "undiscovered public knowledge" [Swanson, 1986].

Swanson downloaded 111,603 MEDLINE journal articles published between 1980 and 1985. He confined his attention to the titles of the papers collected. Swanson constructed a HAL semantic space from a vocabulary containing all words in these titles, save for those excluded by a stop but in the ARROSMITH system. The resulting vocabulary contained 28,834, which is the dimensionality of the semantic space.

Swanson's experiments manipulated the size l of the window and the threshold δ, which fixes the properties comprehended by the source concept which would be involved in the information flow computations. The importance of window size lay in the likelihood that the bigger the window (i.e., the larger the context), the greater the number of B-terms spotted. The importance of heavily weighting the threshold parameter lay in the likelihood that the Raynaud representations would have desirable degrees of relevance if heavily weighted.

Using the Raynaud representation as the source concept in Definition 8.26, the 1500 most heavily weighted terms were computed. Although 1500 is arbitrary, it reflects the fact that computational costs vary proportionally with vocabulary size. Implicit information inferences were ranked according to information flows — the greater the flow the higher the ranking. Swanson wanted to compare his information flow computations with other kinds of outcome computed on the Raynaud representation. One such is cosine, which, when used in semantic spaces, measures the angle between representations, where the strength of the association varies inversely with the size of the angle. In HAL's space, the cosine can be got by multiplying respective representations and ranking them in descending order of cosine. This is possible since in the HAL space representations are given a remit length normalization.

Using the Minkowski distance metric; it is possible to measure the distance between concepts x and y in the n-dimensional HAL space. Accordingly

Raynaud	Cod	Liver	Oil	Fish
Information flow $(l = 50, \delta = \mu)$	0.12(484)	0.34(54)	0.12(472)	0.04
Cosine $(l = 50)$	0.13(152)	0.04	0.04	0.06
Euclidean distance $(l = 50)$	1.32(152)	1.38	1.38	1.37(1088)

Table 9.2 Implicit information inference and semantic association strengths based on the Raynaud representation

Definition 9.25

$$d(x, y) = \sqrt[r]{\sum_{i=1} (|w_{xp_i} - w_{yp_i}|)^r}$$

where $d(x, y)$ is the distance between representations of x and y.

When x corresponds to a Raynaud representation, both Euclidean distance (r = 2) and city-block distance (r = 1) can be computed, and the y-terms are rankable on increasing order of distance, where terms closer to x are taken as having higher levels of semantic *connection*. In Swanson's experiment the top 1500 y-terms were singled out for consideration.

Cosine and Minkowski distance metrics measure the semantic strength of the association of x and y in the HAL space. Information flow computation is different. It measures the level of information overlap in the target term relative to the source term.

For ease of exposition, we report only the results achieved in the best runs. Degree of information flow and strength of semantic association are represented by the numbers in the table's cells, with the requisite ranking in parentheses. Bolded values indicate terms occurring in the top 1500. City-block metrics produced unsatisfactory runs, and don't appear in the table.

What stands out is that, for three of the four tested terms, information flow through a semantic space managed to register their implicit association with "Raynaud". Also of significance is the comparative lowness of these rankings. It bears on this that the best results were achieved when above average weightings were given to the Raynaud representation. Even so, in that situation only one B-term had an above average weighting ("platelet": 0.15), and the other B-terms occurring in the representation had below average weightings. This means that they failed to part of the information flow. It is also interesting that *relevant* information flow was restricted to the same one B-term, "platelet". However, when the weights of the B-terms were increased manually and the Raynaud vector was set at unity, the runs are more encouraging. All four target terms receive information flow, and three of the four now place quite in the ranking.

Regardless, semantic space calculations provide an account of how implicit connections can be computed from a semantic space and interpreted from an abductive perspective. In an automatic setting, information flow computation through a high dimensional space is able to suggest the majority of terms needed to simulate Swanson's Raynaud-fish oil discovery, though the strength of suggestion is relatively small.

It is interesting to note that in HAL-based semantic space models there is no express capacity for seeking out or responding to considerations of relevance and plausibility. Likewise, there is no express role here for analogy. In the HAL model, semantic weight is dominantly a matter of the physical distance between and among co-occurring terms. Not only is this a not especially notion of semanticity, intuitively speaking, but the HAL runs also show that comparatively light semantic weightedness is all that is required for comparatively successful hypothesis-generations, in the manner of Swanson.

This is a highly admonitory turn of events. It suggests confirms what we have repeatedly claimed, namely, that hypothesis selection does not require the abducer to make *judgements* about what is relevant, plausible and analogous. But it also suggests that hypotheses that are selected need not satisfy *conditions* on relevance, plausibility or analogousness. In particular, it calls into question our claim that a wining hypothesis will always turn out to have a determinate place in a filtration-structure. But if that claim should fail, it may be that the best to be said for our intuitions about the relevant, the plausible and the analogous is that they issue forth in judgements made (albeit tacitly, for the most part) after the fact of hypothesis-selection. Part of the importance of research in mechanized abduction is the further light that it might throw on these suggestions. We also see in this a considerable rehabilitation of what we (not HAL) call topical relevance.

In the Swanson case, the implicit links between Raynaud's disease and fish oil, were carried by connecting terms of the form $A - B$ and $B - C$, where A terms are from the Raynaud lexicon and C terms are from the fish oil lexicon. Implicit inferential flow thus passes between $A-$ and $C-$ terms by way of intermediate $B-$ terms. The basic structure of this flow resembles consequence relations that admit of an Interpolation Theorem. It also reflects the presence of the Anderson–Belnap conception of topical relevance, viz., term overlap. All this is food for thought. A semantic space approach to computerized abduction employs a weak notion of semanticity. The implicit inferential flows that drive the task of hypotheses-generation and hypothesis-engagement are semantically modest. The theory's tacit responsiveness to relevance constraints is one that involves the crudest conception of topical relevance to be found in the entire relevance canon. Yet HAL produced the right answer for Swanson's abduction problem. The semantic essence of the HAL model is given jointly by a pair of factors which our intuitions would lead us to think of as syntactic. Semantic insight is lexical co-occurrence under a distance relation. Inferential flow is driven by term-sharing. We see in this an approach to semantics that philosophically-trained readers might well associate with Paul Ziff's *Semantic Analysis* of over forty years ago [Ziff, 1960].

HAL's computational abduction successes were transacted in semantically austere computational environments.[12] It is clear that such semantic austerities possess economic advantages. In what we have so far said about the logic of down below, we have postulated structures with capacities to economize with complexity. In our discussions of Peirce, we have remarked upon the emphasis he gives to the human instinct for guessing right. The two points come together in a suggestion that is highly conjectural, but far from unattractive. It is that in real-life cases, especially in cut-to-the-chase abductions, beings like us might well be running something like the abductive logic of *HAL*-semantic spaces. It is a suggestion that we pass on to the ongoing research programme in cognitive science, with a special nod to neurocomputation and neurobiology.

[12] This calls to mind the extent to which causal attributions are founded in the cue-interpretation paradigm of low-level associative processes, [Shanks and Dickinson, 1987; Waldmann and Holyoak, 1992], as contrasted with the more structured assumptions of causal model theory and the Rescorla–Wagner model [Rescorla and Wagner, 1972]. See here [Waldmann, 2000; Tangen and Allen, in press].

This Page is Intentionally Left Blank

Part III

Conceptual Models of Abduction

This Page is Intentionally Left Blank

Chapter 10

A Glimpse of Formality

> The closest thing to God is an original thought.
>
> Abraham Isaac Kook

10.1 Introduction

In the previous chapters, we have endeavoured to produce a conceptual explication of abduction. A conceptual explication involves a considerable clarification of basic data about what abduction is and how it operates. It thus functions as an informal theory . As we have already explained, the informal theory is input for a formal theory, which latter is the business of the rest of the book. First we want to help the reader orient himself.

We are going to explain the basic ideas involved in fomalising abduction and give some examples. A general formal theory and a general formal modelling of abudction is left for the final volume of this series of books. The main reason for the final volume is that since all the concepts we are studying, such as relevance (Volume 1 of this series), abduction (the present Volume 2), fallacies (forthcoming Volume 3) etc. are all connected, we will need to produce a common final comprehensive modelling in the last volume. Still, we need to give the reader an idea of what features are needed for formal models of abduction and give some tools and some case studies. This is what we are going to do in the rest of this book.

Our plan is to examine the popular *AKM* model of abduction mentioned in Section 3.7 above and to discuss its formal needs and shortcomings. The discussion will naturally lead us to a model of our own proposed *GW* abduction. This is done in Sections 10.1.1 and 10.1.2. Section 10.2 gives some schematic remarks and later sections of this chapter give some case studies.

By the end of this chapter it will be made abundantly clear that the actual process of abduction also depends on the exact formulation of the logics involved. Now although we expect to do the formal modelling in a later volume, some material should be given now for the benefit of the reader. So chapter 11 gives a bit of a general theory of logical systems. Subsequent chatpers give sample formal abductive processes for some host logical systems

The aim of this section is to give the reader an idea as to what sorts of issues formal models of abduction we could reasonably be expected to address and what such models might look like. We imagine a certain cognitive agenda T, and a sentence V such that if V is obtained in a certain way, T would be realised. V is therefore a *pay-off* proposition for T. Given its role in relation to T, V may also be considered informally as a *goal*. Agendas can also be likened to targets. The structure of agendas is examined in detail in [Gabbay and Woods, 2003a]. For the present it suffices to use the notion informally.

10.1.1 The AKM model

Our starting point is the AKM schema for abduction, of Section 3.7.[1] In the *AKM*-model we have the following:

AKM 1 V (is the pay-off for T).

AKM 2 $\sim (K \looparrowright V)$, ($V$ does not follow from our knowledge base K).

AKM 3 $\sim (H \looparrowright V)$, ($V$ does not follow from our proposed hypothesis H).

AKM 4 $K(H)$ is consistent, (adding H to K to obtain $K(H)$ gives a consistent result).

AKM 5 $K(H)$ is minimal, (otherwise why not always take $K(V)$).

AKM 6 $K(H) \looparrowright V$, (the new theory $K(H)$ does yield V).

AKM 7 Therefore we are provisionally justified in assuming H.

Let us examine what kind of mechanisms are required in our logic to be able to model (AKM 1)–(AKM 7) above.

1. First we need to specify the structures that our knowledge bases K can take and explain what it means for us to move from K and H to $K(H)$, i.e. explain the process of insertion of H into K.[2]

[1] Among its supporters are [Aliseda-LLera, 1997; Kuipers, 1999; Magnani, 2001a; Meheus *et al.*, forthcoming]. Hence the acronym 'AKM'.

[2] Note that we are talking about "insertion" rather than "revision". There is a difference which can be appreciated only after reading more about the general theory of logical systnes. The usual notation for revision of K by H is $K \circ H$.

2. We also need a notion of the consistency of a knowledge base.

3. We need a notion of $K(H)$ being minimal relative to adding H to K. This can be part of the process of constructing $K(H)$ or of the definition of $K(H)$.

The above is not enough. Suppose we have as a goal V_1 and we add the hypothesis H_1. Suppose the process continues and we consider $V_2, V_3, \ldots$ and add $H_2, H_3, \ldots$ We thus get a sequence of knowledge bases $K(H_1), K(H_1)(H_2)$. We need to put forward some theory about how to deal with such a sequence i.e. we need a theory of iterated abduction.

Note that if the agenda cannot be closed from our knowledge base K, then we abduce H and form $K(H)$. If later on we chance upon additional knowledge k_1 which, together with K can close the agenda, then we abandon H and move to $K \cup \{k_1\}$. But now note what has happened here. We formally have a database $K(H)$ and when we add k_1 to it, we decide to revise it (because of some internal cross provability relationships) even though it might be completely consistent.

Since multiple abductions are very common in in real life, our model should take them into account. Given our limited resources and limited time, a large chunk of our knowledge is abduced. Our knowledge bases are continuously updated with new hypotheses which are treated as action-enabling data. In fact, the area of iterated abduction is more central to modelling human behaviour than just one step abduction. This is not to minimise the importance of one step abduction. As in real life, so too in the logic of abduction, small steps precede large steps.

Abduction admits of complexities beyond the fact of iterability. In a more comprehensive model than we are presenting here, such additional complexities would, of course, be taken into account. Such forms include:

a. *Multiple target abduction*, in which the closure of an agenda by way of a conditional in the form $K(H) \looparrowright V$ is itself a state of affairs that closes a further target.

b. *Compound abduction*, in which for subsets K_1 and K_2 of K, and different payoff propositions V_1 and V_2, we have it that $K_1(H) \looparrowright V_1$ and $K_2(H) \looparrowright V_2$.

c. *Transitive closure abduction*, in which for certain (but not all) interpretations of $\looparrowright$, we have it that $K(H) \looparrowright V_1, V_1 \looparrowright V_2$, hence that $K(H) \looparrowright V_2$. For example, this would work when $\looparrowright$ is construed as causal implication , but not for all interprtations in which $\looparrowright$ is explanatory consequence.

Q1 Our present question is: Do we take into consideration when adding H_2 into $K(H_1)$ that H_1 is an abduced item of data, given that it does not have the same epistemic status as the rest of K? In other words, do we assume, as part of the structures of our knowledge bases, a variety of data with a variety of status degrees and do we let the insertion process take that into consideration?

The answer to (Q1) should be "yes", because intuitively a common sense reasoner is sensitive to different kinds of abduced data. If we accept that, then we have to ask question 2.

Q2 Shall we have different types of insertions reflecting different policies of handling the already abduced data in K? In other words, if π denote an insertion policy, should we look at $K_\pi(H)$?

It looks more and more as though the framework we need is that of a Labelled Deductive System (LDS) where data is structured and labelled and different insertion policies can easily be formulated [Gabbay, 1996]. If this is the case, then we can ask our next question.

Q3 Why should we insist on $K(H)$ being consistent? We know from LDS methodology that we can easily (in fact more conveniently) work with a general labelled database, in which there are several notions of inconsistency and where the notion of consistency, although definable, is not central at all.

A more likely notion is that *acceptability*, that we want the database to have a certain kind of structure. It may be inconsistent, but structured in such a way that we can handle it, or it may be consistent but structured in a way that is unacceptable. [Gabbay and Hunter, 1991; Gabbay and Hunter, 1993; Woods, 2005b]

We now ask a very simple question. How do we effectively find this H? Without a proof theoretic algorithm for $\looparrowright$ (i.e. for checking whether $K \looparrowright V$ holds for arbitrary K and V) we cannot find such candidates H. The proposed criteria (AKM 4)–(AKM 6) (namely $K(H)$ is consistent and minimal in satisfying $K(H) \looparrowright V$) is too general and implicit.

We therefore need to assume that some algorithm $\mathcal{A}$ is available for checking whether $K \looparrowright V$ holds and that using this algorithm we can determine that (AKM 2) $\sim (K \looparrowright V)$ and (AKM 3) $\sim (H \looparrowright V)$ hold.

Our strategy in modelling AKM abduction is to assume some general properties of this proof algorithm and define an abductive algorithm as a metalevel abductive mechanism $\mathcal{M}(\mathcal{A})$ which works on $\mathcal{A}$ trying to find a candidate H.

We now need to postulate some basic assumptions on $\mathcal{A}$.

$\mathcal{A}$ operates on data K and goals V. So it manipulates the pair $K \looparrowright ?V$. We can also reasonably assume that whatever we do next in this algorithm at a given point depends also on what we have done up to that point. This means that we also need to take into consideration the history of the algorithm. Call it $\mathbb{H}$. Thus our algorithm manipulates triples like

$$[K \looparrowright ?V; \mathbb{H}] \qquad (*)$$

This means our current data structure is K, our current goal to prove is V and the history of the computation up to this point is $\mathbb{H}$. For a discussion of how to model GW abduction, see subsection 10.1.2 below.

The algorithm must tell us at this point how to continue. The following are the options:

Option Fail
$\mathcal{A}$ says we cannot continue. No rule applies.

Option Succeed
$\mathcal{A}$ says we succeed (in this branch of the computation) (e.g. if $K = V$, $\mathcal{A}$ might say that we succeed).

Option Continue
$\mathcal{A}$ may have a stock of rules which might apply. A rule has the form below:

$$\begin{array}{c}\text{General form of a computation rule } \mathbb{R}\text{:}\\ [K \looparrowright ?V, \mathbb{H}] \\ \text{if} \\ \bigwedge [K_i \looparrowright ?V_i; \mathbb{H}_i]\end{array} \qquad (**)$$

In other words: to compute $[K \looparrowright ?V; \mathbb{H}]$ successfully we must succeed with all of $[K_i \looparrowright ?V_i; \mathbb{H}_i]$.

We must assume that according to some complexity measure $[K_i \looparrowright ?V_i, \mathbb{H}_i]$ are simpler tasks and that $\mathbb{H}_i$ is obtained from $\mathbb{H}$ by further recording of the rule we have just applied, and possibly abandoning some recorded history which is no longer needed (for example we might wish to remember only the last step!)

It is possible that several rules may apply. Success is assured if one of them can lead us to success.

What does it mean then that $\sim (K \looparrowright V)$ holds? This means that if we start our algorithm with the task $[K \looparrowright ?V; \varnothing]$, then no matter how we continue we always encounter a subtask of the form $[K' \looparrowright ?V'; \mathbb{H}']$ which should succeed (in order for the original task to succeed) but which is actually fails (option failure holds for this task).

How do we do abduction $\mathcal{M}(\mathcal{A})$ on top of an algorithm $\mathcal{A}$?

Abduction works as follows:

Given a task $[K_i \looparrowright ?V_i, \mathbb{H}_i]$ the abduction machine needs to tell us the following

(D1) Whether it is allowed to abduce at this point

(D2) What to abduce.

Assume that we are told to abduce H_i. It must be a simple and obvious choice (e.g. H_i can be V_i). At this point the intention is not to start a new process of abduction[3] but to choose something simple and immediate.[4]

Once this simple and immediate H_i is chosen, then it is immediately clear what $K_i(H_i)$ is supposed to be and it is also immediately clear that the algorithm $\mathcal{A}$ applied to $[K_i(H_i) \looparrowright V_i, \mathbb{H}_i]$ succeeds.

An obvious choice of rules is to allow abduction (D1) only if the sole option available is the failure option (why abduce now if the computation can continue?[5]) and then D2 can choose something which can be calculated in a deterministic way out of K_i, V_i and $\mathbb{H}_i$. It must be a calculation which is sure to terminate and yield something.

(D1)–(D2) are not enough for our purpose. Remember that the original problem was $[K \looparrowright ?V; \varnothing]$ and that the current piece of computation is only a cog in a big logical machine. We need to be able to reconstruct H out of all the H_i. How do we do that? The simplest way is to attach with every rule $\mathbb{R}$ an abduction rule $\mathcal{M}(\mathbb{R})$ going upwards. So if $\mathbb{R}$ has the form

$$\mathbb{R} : [K \looparrowright ?V; \mathbb{H}] \text{ If } \bigwedge_i [K_i \looparrowright ?V_i, \mathbb{H}_i]$$

then we need a rule $\mathcal{M}(\mathbb{R})$ allowing us to know what to abduce for $[K \looparrowright ?V; \mathbb{H}]$ provided we know what to abduce for each $[K_i \looparrowright ?V_i; \mathbb{H}_i]$.

It may be useful to say that we always abduce something, even when $K \looparrowright V$ is successful, in which case we abduce $\top$ (truth) or something harmless (a *unit* such that $K(\textit{unit}) = K$).

The reader may think that $\mathcal{M}(\mathbb{R})$ is just a technical rule, but actually there is a much deeper phenomenon here. The algorithm $\mathcal{A}$ is in practice a meaningful algorithm in the application area it addresses. Put differently, $\mathcal{A}$ follows natural lines of reasoning in the application area. This means that the backward propagation of abduction (the family of rules $\mathcal{M}(\mathbb{R})$) can also have a meaning. But let

[3] In more complex systems there may be several algorithms for abduction and a hierarchy of when one process hands over to another. The hierarchy must not be circular.

[4] In general there may be several possible H_{ij} to abduce, $j = 1, 2, \dots$ to make V_i succeed from K_i. So the general form of abductive options are sets of formulas $\mathcal{H}_{ij}$, such that for each j, $\mathcal{H}_{ij}$ enables V_i to succeed from K_i. We can also order this family of sets according to our preference as to what we consider better hypotheses. We can propagate this ordering as we acquire more sets. We need a special additional logic or choice function to choose one option and define the preference.

[5] If the system talks about component failure in some machine or a system, then some components may be known from experience to be weak and likely to fail. In this case we abduce immediately even though we can continue and explore further.

us stop and reflect on what we are saying here! We are saying that the abductive process is concerned not only with finding a hypothesis H when needed, but also with providing lines of reasoning of how to propagate this hypothesis backwards against the flow of deduction. This point is important and, so far as we know, new.

There is an interesting point to observe here. Our own algorithm $\mathcal{A}$ goes backwards, reducing one provability question to another. The abduction algorithm, therefore, goes forwards, in an opposite direction to the provability.

The tableaux method, to take a well known example, works like that. We try to falsify $K \looparrowright V$ and if the tableaux is closed then $K \looparrowright V$ is impossible to falsify. So the abduction process will close the endpoint tableaux and propagate backwards the additional assumptions.

Note that we can as easily (in the tableaux case) abduce on the goal rather than on the data. In other words, if $K \looparrowright V$ fails, we use the same process to abduce an V' such that $K \looparrowright V'$ will succeed! I.e. we change the goal posts. People do that a lot in real life.

The real life abduction changes both K and V in order to succeed. So, for example, if someone undertakes a project and cannot exactly meet the goal he may slightly misinterpret K and justify a slightly different V and hope he can get away with it!

Another point to discuss is the connection of abduction with inconsistency. If $K \looparrowright V$ does not hold, then $K(\neg V)$ is consistent namely it does not prove $\bot$. Find an H such that $K(\neg V)(H)$ is inconsistent (i.e. does prove the 'goal' $\bot$). Then this means $K(\neg V) \looparrowright \neg H$ holds and in many logics this implies that $K(H) \looparrowright V$ holds.[6]

So now conditions on acceptability of H become parallel to conditions on what kind of wffs we want to use to render a theory inconsistent.

Let us take stock of what we have so far. We start with $K \looparrowright ?V$ which does not succeed, i.e. our algorithm $\mathcal{A}$ definitely fails. We have a meta algorithm $\mathcal{M}(\mathcal{A})$ which follows the computation $\mathcal{A}$ and yields an H (or several of them) such that $K(H) \looparrowright V$ succeeds.

This is how we abduce. The reader may ask whether the abduction depends on the choice of $\mathcal{A}$? The answer is yes. Different $\mathcal{A}$'s with the same $\mathcal{M}$ will (or may) give different H's. We are not bothered by that. We think that part of the logic is its proof theory and so it makes sense that the abduction depends on the proof theory. In practice all (or let us say, almost all, to be safe) abduction algorithms for a logic are tagged to some proof theory for that logic.

Also note that according to Gabbay [Gabbay, 1996], a system like $\mathcal{M}(\mathcal{A})$ (i.e. a logic with abduction) is also considered a logic. In other words, part of the notion

[6] In practice one can use a theorem prover to generate $Y_1, Y_2, Y_3 \ldots$ such that $K(\neg V) \vdash Y_i, i = 1, 2, \ldots$. Then any $H_i = \neg Y_i$ can serve as a hypothesis to prove V. Conditions on what kind of H we want become conditions on what kind of Ys we generate.

of a logic is to have resident abduction algorithms. This opens the way for several abduction algorithms, a primary one $\mathcal{M}$ and a secondary μ one. We can apply the secondary abduction when the primary one does not yield an answer. Here is an example: Start with $K \looparrowright ?V$ failing. We apply $\mathcal{M}(\mathcal{A})$ to the problem and obtain H. We look at $K(H)$ and it is not acceptable (therefore the abduction fails). We now have a secondary target. Modify the abduction $\mathcal{M}$ to $\mathcal{M}' = \mu(\mathcal{M})$ so that an acceptable H' is obtained.[7]

Let us now rewrite (AKM 1)–(AKM 7) in view of the above discussion. The prefix 'N' stands for 'new'.

NAKM 1 V (should succeed).

NAKM 2 $\sim (K \looparrowright V)$, The algorithm $\mathcal{A}$ fails to succeed from K with V as a goal.

NAKM 2.5 We apply $\mathcal{M}(\mathcal{A})$ and get an H.

NAKM 3 $\sim (H \looparrowright V)$. To ensure this we need to look at how $\mathcal{M}$ works and prove it as a theorem.

NAKM 4 $K(H)$ is acceptable.

NAKM 5 $K(H)$ is minimal. Again we need to say what this means and prove it as a theorem.

NAKM 6 $K(H) \looparrowright V$. We prove this as a theorem on $\mathcal{M}$.

NAKM 7 Therefore, H. This statement makes sense in traditional logics where $K(H)$ is $K \cup \{H\}$. In our context we must say "Therefore we insert H into K". If we insert H into K as a structure, in what sense do we say that H holds?

Do we mean that $K(H) \looparrowright H$ holds? We can require it if we want as part of NAKM 7 but in general it does not need to hold.

NAKM 8 Provide machinery for the backward reasoning of the abduced hypotheses (against the forward deductive flow of H).

10.1.2 The GW Model

The backward propagation mechanism that we saw we need for the case of abduction is a special case of a general way of propagating metapredicates over proof algorithms. We can imagine that $\mathcal{M}$ gives any value (action, cost, time, etc) to the tasks involved in a rule $\mathbb{R}$ and that $\mathcal{M}(\mathbb{R})$ propagates these values backwards. In

[7] Going back to a previous footnote, each abduction process $\mathcal{M}$ may be itself a multiple algorithm.

fact, such metapredicates can apply to any kind of finitary algorithms (not necessarily having to do with logic) and the backward propagation can be an inductive definition of what the values are.

We invite the reader to the following different interpretation for the present notation. Imagine we have a target V which we want to achieve in some environment K. Present this as $K \looparrowright ?V$. We have at our disposal a variety of steps we can take to achieve our target. If we execute any one of these steps our target may be replaced by other auxiliary targets or subtargets. Of course the relevant environments for achieving these subtargets may be different. So we have several *options* (we can use Robin Milner's term *tactics*) for reducing our target to other subtargets. We can still have that the current choice of option depends on the history $\mathbb{H}$ of previous choices.

We thus write

General form of a tactic $\mathbb{T}$

$$\begin{array}{c} [K \looparrowright ?V, \mathbb{H}] \\ \text{if} \\ \bigwedge [K_i \looparrowright ?V_i, \mathbb{H}_i] \end{array}$$

Note that in this generality the process has nothing to do with consequence. It is a general algorithm with tactics which modifies the environement to achieve one's goals. It could be modelling lobbying in the Senate for some bill serving some interest group and the options are legal actions, donations, press releases, demonstrations, etc, etc.

Note that in this context we may want to make some targets fail as well as some other targets achieved.

A more practical representation is to write

$$[K \looparrowright ?V, \mathbb{H}] = x, x \in \{0, 1\}$$

where $x = 1$ signifies the goal achieved while $x = 0$ signifies the goal failing.

A tactic $\mathbb{T}$ becomes a pair of tactics, one for $x = 1$ and one for $x = 0$. So for example, for $x = 1$ we hve

$$\begin{array}{c} [K \looparrowright ?V, \mathbb{H}] = x \\ \text{if} \\ \bigwedge [K_i \looparrowright ?V_i, \mathbb{H}_i] = x_i \\ \text{where } x_i = 0 \text{ or } 1. \end{array}$$

In this context we can abduce any additional legitimate component which can help achieve or fail our targets!

So for example even in the context of consequence explanation, we can abduce not only new hypothsis H to form $K(H)$ but also a new proof rule or change the original consequence relation. Such a proof rule abduction does not fall under the AKM schema of abduction.

The reader should bear this different reading of our notation in mind while going through the discussion of subsection 10.1.1 again. This new reading of the notation does model GW abduction.

10.2 Some Schematic Remarks

Two well-known logicians Bjoern and Johannes (B and J) have accepted a visiting position for three months at the newly established Institute of Artificial Intelligence Technology in the Far East. The Institute takes pride not only in its high salaries for visitors (they negotiate their *net* take-home amount per month, all expenses paid!) but also in its support facilities, which are themselves run by advanced AI programs.

Our two visitors have been located in a large lab containing computers, printers, an especially fancy copying machine, automated windows and doors, in short, the latest in latex technology[8]

B and J are the only people in the lab. Life passed smoothly in the first two days until B complained to J that the photocopier does not seem to be responding. 'I keyed in my code but nothing is happening' he said. J advised him to switch it off and on again and re-try. To no avail. 'Let me try', J says. He switched the machine off and on again and keyed in his code. The machine worked. B happily resumed his photocopying.

In the days ahead, every now and then the photocopier would conk out again. Various attempts were made to start it, usually successful, after a number of tries.

It is now today, and whether the copier will work today (and how) has become a focus of their attention. After a while, B and J also noticed that on rainy days, when the photocopier is working, the door is open, and if they insist on shutting the door, the photocopier stops working.

B conjectured that there is some intelligence in charge of their lab and that there is some kind of program coordinating the various operations. Being logicians, they decided to build a model of the relevant operations and discover what the program does. To this end, they decided that three languages would be required. [9]

1. *State Language $\mathcal{L}$*
   capable of describing and reasoning on the state of the lab.

2. *An Action language $\mathcal{A}$*

[8] Including an electric pencil sharpener.

[9] If we are working in a logic which is general enough (say labelled deductive systems) then all three components can be part of this same logic.

3. *A meta language $\mathcal{M}$*
   coordinating the interaction between the other two languages.

It is in the metalanguage that B and J hoped to formulate the hidden rules of the program that was controlling the lab.

B and J understood that the state language and its logic can be any logical system in the sense discussed in [Gabbay, 1996]. For their present purposes the language chosen was a propositional language with atoms intended to mean:

b = B uses the photocopier
j = J uses the photocopier
c_1 = The photocopier accepts B's code
c_2 = The photocopier accepts J's code
w = The window is open
d = The door is open
r = It is raining outside.

Given this interpretation, there are some obvious rules. For example,

- $\neg(b \wedge j)$
- $\neg(c_1 \wedge c_2)$

The action language describes actions $\mathbf{a}_i$ of the form $\mathbf{a}_i(\alpha_i, \beta_i)$, where α_i is the precondition and β_i is the post condition. If Δ is a state and $\Delta \vdash \alpha_i$ then it may be that action a_i can be executed, in which case we move to a new state $\Delta \circ \beta_i$, obtained from Δ by the revision process $\circ$:

$$(\Delta, \beta) \mapsto \Delta \circ \beta.^{10}$$

The job of the metalanguage is to give expression to conditions that regulate the intersection of the actions, revisions and the logic of the theories Δ.

Following the example of B and J, we shall use an executable temporal logic as a metalanguage but will write the rules in semi-English in this introduction.

The expressions of $\mathcal{M}$ are as follows:

1. *Execute*(q), for $\pm q$ a literal of $\mathcal{L}$.

   For example, *Execute*(w) means: *close the window.*

2. Formulas of $\mathcal{L}$ are also atomic formulas of $\mathcal{M}$.

3. Temporal operators $Y\alpha$, yesterday α (or some state before α obtains), $P\alpha$ (*some past state α obtains*); and $S(\alpha, \beta)$ (*from a past state at which α was true, β has been true in all intermediate states.*)

[10] Compare this with the notation $\Delta(\beta)$ of the previous section, meaning the inserton of β into Δ.

4. $\Box\alpha$, (*always* α)

5. Rules of the form
$$\text{Past } \mathcal{M} \text{ wff } \Rightarrow \mathit{Execute}(\Pi).$$

There are examples of $\mathcal{M}$ rules in action.

1. If c_1 three times in the past, up to now, without any c_2 between them, $\mathit{Execute}(\sim c_1)$. That is, we cannot use the machine more than three times without someone else using it.

2. Always either the door or the window is open or the photocopier is not on.

3. If it rains, then close the window.

An abductive logic describes or mimics what an abductive agent does.[11] In its most basic sense, a formal abductive agent is faced with the following situation. There is a database (or a belief-set or a theory) Δ and a fact symbolized by a target wff V such that $\Delta \nvdash V$. The basic abductive task is to adjust Δ in ways that now yield ϕ, and to do so in compliance with the appropriate constraints (e.g. explanatory power, predictive accuracy, simplicity, etc.) The purpose of this final chapter is to explain the kind of logics that are involved in our study of abduction. A formal treatment of such logics is given in a later volume, in the context of formal models of cognitive systems. This chapter will take the reader through the fundamental concepts involved by following a few simple examples.

Consider a simple propositional language, with atoms $\{p, q, r, \ldots\}$ and the connectives $\looparrowright$ for implication, and $\bot$ for falsity. We can form the set of all formulas in the usual way.

How do we define a logical system on this language? We define the following two notions

(1) The notion of a *theory* Δ (database) of the logic

(2) The notion of consequence for the logic, i.e. the notion of the conditions under which we say the Δ *proves* A in the logic, i.e. $\Delta \vdash ?A$.

The usual notion of a theory is that Δ is a set of wffs. This notion works for many logics but not for all. Many logics require, in view of the intended application, additional structure in the theory. For example, a theory Δ can be considered abstractly as a list of wffs
$$\Delta = (A_1, \ldots, A_n)$$

[11] The discussion in this section is appropriate for any application area modelled by a logical theory. In the philosophy of science we have explanation/abduction in the context of much richer mathematical theories (mechanics, relativity, etc). The 'data' and 'inference mechanisms' are different.'

where the consequence relation holds between lists of wffs and a formula (i.e. $(A_1, \ldots, A_n) \vdash B$). We shall come back to this later. Meanwhile we examine what our options are for defining $\vdash$. We begin with the options for classical logic.

Option 1. Tableaux.
Define truth functions h:Wff$\to \{0, 1\}$ in the traditional manner and let

$$\Delta \vdash A \text{ iff} \forall h (\forall B) \in \Delta(h(B) = 1 \to h(A) = 1).$$

The proof theory is constructed from semantic tableaux.

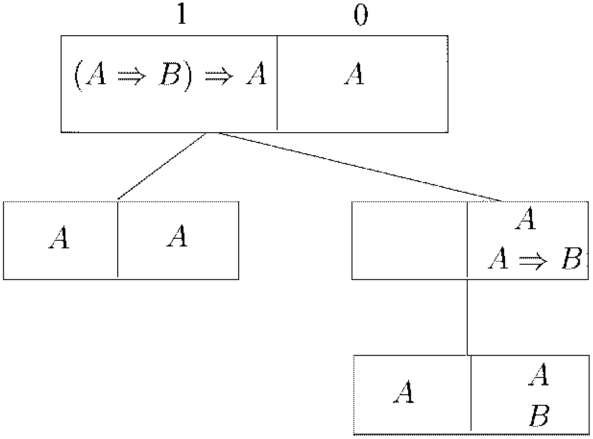
Option 2. Resolution.
Let $A \wedge B$ be defined as $(A \Rightarrow (B \Rightarrow \bot)) \Rightarrow \bot$. Let $A \vee B$ be defined as $(A \Rightarrow \bot) \Rightarrow B$. Let $\neg A$ be defined as $A \Rightarrow \bot$. Write every wff in conjunctive normal form using $\wedge, \vee$ and $\neg$. To check whether $\Delta \vdash A$ consider $\Delta \cup \{\neg A\}$ written in conjunctive normal form as clauses, and define and apply resolution.

Option 3.
Use a goal directed proof mechanism (see Definition 12.12)

These are three different ways of telling us when $\Delta \vdash A$. To illustrate the difference let us check the example of $(A \Rightarrow B) \Rightarrow A \vdash ?A$.

Option 1.
Try to find a falsifying h, using tableaux:

The proof goes as follows:



The tableaux is closed. So a countermodel h' does not exist.

Option 2.
To take the set $\{((A \Rightarrow B) \Rightarrow A), \neg A\}$ and rewrite it in clausal form as $\{A \vee A, A \vee \neg B, \neg A\} = \{A, A \vee \neg B, \neg A\}$.

$A, \neg A$ give the empty clause.

Option 3

To deploy the goal directed computation of Definition 12.12
Success $(\{(A \Rightarrow B) \Rightarrow A\}, A) = 1$
if
Success $(\{(A \Rightarrow B) \Rightarrow A\}, A \Rightarrow B) = 1$
if
Success $(\{(A \Rightarrow B) \Rightarrow A, A\}, B) = 1$
if (by restart)
Success $(\{(A \Rightarrow B) \Rightarrow A, A\}, A) = 1$
if Success!. . .

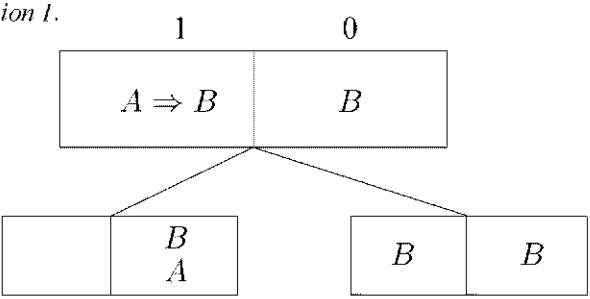
It is for us an important point of departure that a logical system is not just $\vdash$ (the consequence relation) but also the proof option chosen for $\vdash$. Accordingly, the above discussion yields three logical systems, ($\vdash$, Option 1), ($\vdash$, Option 2) and ($\vdash$, Option 3). They share the same consequence relation, but the proof theory is different.

The abductive mechanism which we will develop in detail in Chapter 13 will depend on the proof theory, and so as far as abduction is concerned, the three logics may give different results. Here is an example of an abductive problem:

$$A \Rightarrow B \vdash ?B$$

We abduce by applying the proof theory and adding assumptions at key moments of failure.

Option 1.



The left box is not closed. To close it, we can abduce A or abduce B.[12] *Option 2.*

[12] We can also abduce their disjunction $(A \Rightarrow \bot) \Rightarrow B$, which is weaker. However, our policy in these examples is to enable success immediately at the point at which we are stuck.

We run a resolution on

$$\neg A \vee B, \neg B$$

obtaining $\neg A$. From this we are not able to obtain the empty clause. So our obvious abductive option is to add A.

In fact, we can add the negation of any formula in the set, but it is preferable to follow our proof procedure as far as we can go, and abduce only at the last moment, when we are stuck.

Option 3.
Success(A$\Rightarrow$ B, B)=1
if
Success(A$\Rightarrow$ B, A)=1.
For this to succeed we must add A to the data.

We saw above that applying the proof mechanism and trying to succeed with $A \Rightarrow B \vdash ?B$ gives rise to several candidates for abduction. The next step is to choose which candidate to take. Do we add A? Do we add B? The choice actually requires a second logic, which we can call the *background logic of discovery*. We may include the abduction algoirthm itself (as defined using Π) as part of the logic of discovery.

We mentioned the possibility that the database may be a structure, say a list. Why do we need structure? The need arose from the interpretation of the logic and its applications. If, for example, $\Rightarrow$ is strict S4 temporal implication, then $A \Rightarrow B$ means that whenever A is true (in the future) then B is true. We might think of $A \Rightarrow B$ as an insurance policy (whenever you are disabled you get \$100,000). To apply modus ponens, we must have A come *after* $A \Rightarrow B$ not before. So

$$(A \Rightarrow B, A) \vdash B$$

but

$$(A, A \Rightarrow B) \nvdash B.$$

$C \Rightarrow (A \Rightarrow B)$ means whenever C holds then we get the payout of the insurance policy $A \Rightarrow B$.

$(A \Rightarrow B) \Rightarrow C$ means whenever we have a policy then we get C.

Let us now reconsider the abduction problem

$$A \Rightarrow B \vdash ?B$$

for our strict implication. The available proof theories require modification because the options so far considered yield classical logic. Indeed in the literature there are tableaux systems (option 1) for strict implication, and there are resolution systems (option 2) for strict implication; but they are too complex to describe here, and fortunately there is no need. It is sufficient for the reader to realize that the *abduction options depend on proof procedures*. For option 3 (goal directed) it is easy to see intuitively what needs to be abduced in our example. We need to abduce A, but we must be careful. We abduce A to be *positioned* after $A \Rightarrow B$ (not before); otherwise we cannot prove B. Again, think of $A \Rightarrow B$ as the insurance policy. I want to get B ($100,000). How do I do it? I need to have an accident *after* I get the policy, not before. To write this properly, we need to adopt *labelling* as part of our logic and to abduce A with a label indicating its position. To achieve this uniformly we need to label the data $t : A \Rightarrow B$ and label $s : B, \quad t \leq s$ and then abduce $s : A$. Our new database is

$$\begin{aligned} &t : A \Rightarrow B \\ &s : A \\ &t \leq s. \end{aligned}$$

Since $t : A \Rightarrow B$ is the original datum and $s : A$ is abduced (and hence possibly open to refutation) we may *double label* the data as:

(original, t):$A \Rightarrow B$
(abduced, s): A
$t < s$.

We see how easy it is to motivate moving into deductive systems involving structured labeled databases!

We can now take our insurance example a bit further. We hinted above that in order for B to obtain we need to assume that A obtained after $A \Rightarrow B$.

Suppose we have an insurance assessor checking the truth of A. Is there a disability?, he asks. The assessor can take *actions* to try and *refute* A—may he tempt the policy holder to a football game and video it? The actions taken by the insurance assessor will create an $r : X$ such that $s : A$ and our $r : X$ and $s < r$ yield a contradiction. The action has the form *action*=$(r_1 : X_1, r_2 : X_2)$. $r_1 : X_1$ is the precondition for taking the action. Preconditions are required to be reasonable because we need to have reasonable grounds for suspicion before we can be allowed to 'trip-up' the policy-holder. If the precondition can be obtained then we take the action and if successful we get $r_2 : X_2$. Thus our system must be a logic with actions, labels, inconsistency and abduction—and one more thing: a *refutation/revision process*.

$r_2 : X_2$ goes into the database.

Our inconsistency component tells us it is inconsistent. We must know how to deal with it. One possibility is to take $s : A$ out! If so, we must be able to recognize this option from the labels.

We summarize the logic components that are needed to deal with our simple insurance example:

1. Data are labelled and structured, forming databases Δ.[13]

2. A proof mechanism Π is available for proving formulae A with labels t. $\Delta \vdash_{\Pi} t : A$

3. A notion of inconsistency.

4. An abduction mechanism. Given $\Delta \nvdash_{\Pi} t : A$ we can abduce several Δ' such that $\Delta + \Delta' \vdash_{\Pi} t : A$. Note that we need to know how to add (input) Δ' to Δ. (If Δ and Δ' are structured how do we combine them?) Note that in non-monotonic logic and in general labelled deductive systems we may get $t : A$ to be provable by *deleting* from Δ. So we must equally allow for some Δ' such that $\Delta - \Delta' \vdash t : A$ or a combination of add and delete. Thus 'input' could mean a combination of add and delete.[14]

5. A notion of action along preconditions and postconditions which can be taken to check/refute an abduced Δ'

6. A revision process is needed to identify and take out refuted wffs in case of inconsistency.

7. A proof theory that takes account of actions.

8. We need one more component. Imagine in our example that the manager of the insurance company asks the assessor whether a \$100,000 (B) is to be paid out (i.e. he asks, does $A \Rightarrow B \vdash B$?). Of course, the answer is yes (from the above considerations). But what the manager is really asking is Have you (or can you) take action a to check/refute A? The assessor can say 'Yes, my action a is available to refute A (i.e. show $\neg A$)'. We write the above as $\Delta \vdash_a \neg A$, reading: After action a is taken, $\neg A$ follows.

10.3 Case Study: Defeasible Logic

We briefly revisit the account of individual cognitive agency. An individual agent manages his affairs with straited resources: information, time and computational power. Slightly over-stated, he doesn't know enough, hasn't time enough and hasn't the wherewithal for figuring things out that would enable him to meet the performance standards postulated by standard logics of rationality. Individuals

[13] The most general notion of database allows data items to have procedural 'patches' which affect the proof procedure Π once the item is used. Π itself can operate in several modes and the 'patches' can shift the mode. Component (8) below contains action as a mode.

[14] Implementing deletion in the object level by adding anti-formulas has been studied in [Gabbay *et al.*, 2002].

prosecute their cognitive agendas on the cheap. They take short cuts and they assume the concomitant risks. Among the economising measures of individual cognitive agents are what we have called their scant-resource compensation strategies. These include a disposition toward hasty generalisation, toward generic inference and the recognition of natural kinds. Also prominent is an inclination to stay with received opinion, and, in matters of new opinion, to be satisfied with the say so of others.

It is time to give a case study for our individual (low-resource) reasoning agents, in which we shall apply two of our low resource principles:

1. low resource individual agents perceive *natural kinds* around them

2. low resource individual agents are *hasty generalisers* of *generic rules*

and obtain immediately the well known and much studied *Nute's defeasible logic*

Imagine our agent looking into the world around him and dividing it into certain natural kind subsets. Let $A_i(x)$ be the predicate x *is of kind* A_i and let $\Rightarrow$ be logical implication (of some known logic) then our reasoning will have rules of the form

3. $A_i(b)$, where b is a constant name

4. $\forall x(A_i(x) \Rightarrow A_j(x))$ where A_i, A_j are natural kinds.

If we use the Prolog notation we can write (4) as

4*. $A_i(x) \Rightarrow A_j(x)$

and view (4*) as an *absolute* (*non-defeasible*) rule.

For example

5. Penguin $(x) \Rightarrow$ Bird(x)

6. Bird (Tweetie)

Our reasoner also has some positive and negative property predicates $B_j(x)$ and $\sim B_j(x), j = 1, 2, \ldots$ as well as a hierarchical understanding of them in the form of rules

7. $\forall x(B_i(x) \Rightarrow B_j(x))$,

or written in Prolog as

7*. $B_i(x) \Rightarrow B_j(x)$.

Again, (7) is an absolute non-defeasible rule.

Our agent is a hasty generaliser who deploys *generic* defeasible rules of the form

8. $A_i(x) \twoheadrightarrow B_j(x)$

For example

9. Bird $(x) \twoheadrightarrow$ Fly(x)

10. penguin $(x) \twoheadrightarrow\sim$ Fly(x).

Formally our agent's logic has predicates $A(x), B(x), \sim A(x), \sim B(x)$, *absolute rules* of the form $A(x) \Rightarrow B(x)$, *defeasible rules* of the form $C(x) \twoheadrightarrow D(x)$ and *facts* of the form $A(b), \sim A(b), b$ an individual name.

Here is a well known sample database Δ_1, t_i are used to name the data items and rules:

t_1: Bird$(x) \twoheadrightarrow$ Fly(x)
t_1: Penguin$(x) \twoheadrightarrow\sim$ Fly(x)
t_3: Penguin$(x) \twoheadrightarrow$ Bird(x)
t_4: Bird(Tweetie)
t_5: Penguin(Tweetie)

Δ_1 can prove both that Tweetie flies and that Tweetie does not fly. Defeasible logic allows certain proofs to *defeat* other roofs, and is thus an example of a genuine non-monotonic logic. The way in which this system one proof can defeat another proof is that it is *based* on a *more specific* body of facts than the other proof. The proof of Fly(Tweetie) is based on t_4. The proof of $\sim$Fly(Tweetie) is based on t_5. Item t_5 is more specific than item t_4 (t_4 can *prove* t_5 using *absolute rules* only). Therefore Δ_1 overall proves that Tweetie does not fly.

Consider $\Delta_0 = \Delta_1 - \{t_4\}$. Δ_0 can still prove that Tweetie Flies, as well as that Tweetie does not Fly. How can we show now that overall, Δ_0 proves that Tweetie does not Fly?

We know that $\sim$Fly is based on the more specific information that Tweetie is a penguin and this information is more specific because of Penguin $\Rightarrow$ Bird. However, we have a technical problem of how to bring this out formally in the system?

Let us try making use of the labels. We have

$$\Delta_0 \vdash_{t_5, t_3, t_1} \text{Fly(Tweetie)}$$

and

$$\Delta_0 \vdash_{t_5, t_2} \sim \text{Fly (Tweetie)}$$

Let $\alpha = \{t_5, t_3, t_1\}$ and $\beta = \{t_5, t_2\}$.

The absolute parts of α and β are $|\alpha| = \{t_5, t_1\}$ and $\beta = \{t_1\}$. $|\alpha|$ and $|\beta|$ are equivalent under the absolute part of Δ_0 and taken as they are, we get $|\alpha| \vdash |\beta|$ and not $|\beta| \vdash |\alpha|$, as we would have liked.

Obviously we have a technical problem of how to formalize the intuition that Penguin is more specific than Bird.

Let us look at the problem slightly differently.

Tweetie Flies because it is a Bird and t_1 says that Birds Fly. Tweetie does not Fly because it is a Penguin and t_2 says that Penguins don't Fly. Never mind how we show that Tweetie is a Bird and how we show that Tweetie is a Penguin. The fact is that in Δ_0, Penguin is more specific than Birds.

So let us view the labels α, β as sequences of proof steps. We do labelled modus ponens as follows:

$$\frac{l : A; t : A \to B}{(l, t) : B}$$

where '$\to$' can be either '$\Rightarrow$' or '$\twoheadrightarrow$'.

Thus we get

$$\begin{array}{l}
\Delta_0 \vdash_{((t_5,t_3),t_1)} \text{Fly(Tweetie)} \\
\Delta_0 \vdash_{(t_5,t_3)} \text{Bird(Tweetie)} \\
\Delta_0 \vdash_{(t_5,t_2)} \sim \text{Fly(Tweetie)} \\
\Delta_0 \vdash_{t_5} \text{Penguin(Tweetie)}
\end{array}$$

Now we can strip the last label (if it is a defeasible rule) and get the *fact* last proved and compare which fact is more specific.

To do this properly we need the machinery of Labelled Deductive Systems. We shall study this example in detail in a subsequent chapter. (See also our *Agenda Relevance*, chapter 13.) Abduction in defeasible logic.

To give the reader an idea of what is involved, note that a database Δ can prove A and $\sim A$ in many ways. This means that there are many labels α_i, β_j such that $\Delta \vdash_{\alpha_i} A$ and $\Delta \vdash_{\beta_j} \sim A$. The process in labelled deductive systems which decides whether overall we say $\Delta \vdash A$ or we say $\Delta \vdash \sim A$ or neither is called *Flattening*. This enables us to define a definite $\vdash$ for our case here.

We mentioned that we might not be able to decide whether A or $\sim A$ follows. This happens in our case if the facts are incomparable. Consider

$$\begin{array}{l}
A(x) \twoheadrightarrow C(x) \\
B(x) \twoheadrightarrow \sim C(x) \\
A(b) \\
B(b)
\end{array}$$

We have neither $A(b)$ nor $B(b)$ stronger than the other. So we cannot conclude $C(b)$ nor can we conclude $\sim C(b)$.

Suppose we factually get the information that Tweetie does fly. How can we explain it? We have to abduce. We can add Fly(Tweetie) to the database as a fact and this will defeat ~Fly(Tweetie) because the latter is based on a defeasible rule, but this is not explaining. If we think about it intuitively we 'explain' by saying, using our ability to perceive natural kinds, that Tweetie belongs to a subspecies of Penguins that do indeed Fly. Say Tweetie is an F-Penguin.

Thus we add

$$\begin{aligned}&t_6\colon \forall x(F\text{-Penguin}(x) \Rightarrow \text{ Penguin}(x))\\&t_7\colon F\text{-Penguin}(x) \twoheadrightarrow \text{ Fly}(x)\\&t_8\colon F\text{-Penguin(Tweetie)}\end{aligned}$$

This reminds us of John McCarthy's way of writing defeasible rules

$$A(x) \wedge \text{ not abnormal } (x) \Rightarrow B(x)$$

we will have

$$\text{Penguin}(x) \wedge \text{ not abnormal } (x) \Rightarrow\sim \text{Fly}(x)$$

Doing circumscription will minimise λx Abnormal(x) to be empty. But if Fly(Tweetie) is in the data it will contain Tweetie.

This Page is Intentionally Left Blank

Chapter 11

A General Theory of Logical Systems

Veniet tempus quo posteri nostri tam aperta nos nescisse mirentur.

Seneca

11.1 Introduction

In the previous chapter we said that a theory of abduction depends on several components. One if these is a **base logic** together with a **proof theory** Π for that logic.

The reader may have the impression that our proof theory is an opportunistic adjunct to the base logic (which would make the resultant abduction likewise opportunistic) rather than an essential part of the logic itself. We demur from this view. We have already put forward arguments that a proof theory should be considered part of the logic. (see later in this chapter for an account of this). There is another good reason. Abduction for logical systems based on proof theory take the general form similar to GW abduction. In the spirit of the GW model of Section 10.1.2, the corresponding analog of the base logic is the language describing the environment and the corresponding analog to the proof theory is the language of the algorithm. We leave the general GW abduction model to the last volume of this series. In this volume we concentrate on the logical approach. We saw in Section 10.1.1 that this approach can be very general anyway. The abduction algorithm deploys the proof theory Π to produce candidates which are inputs to the abduction selection mechanisms of the theory.

This chapter will achieve two objects. One will be to show that a base logical system is in fact a pair $\{\vdash, \Pi\}$; and secondly it will survey several candidates for selection as the base logic of our approach to induction.

The chapter to follow this one develops a Labelled Deductive System (LDS) logic with which to model various abductive mechanisms.

This chapter studies the notion of a logical system. The structure of the chapter is such that it leads the reader from the traditional notion of a logic as a *consequence relation* to the more complex notion of what we call a *practical reasoning system*

In general, to specify a logical system in its broader sense we need to specify its components and describe how they relate to each other. Different kinds of logical systems have different kinds of components which bear differnet kinds of relationships to each other.

The following components are identified.

1. *The Language*
   This component simply defines our stock of predicates, connectives, quantifiers, labels, etc; all the kinds of symbols and syntactical structures involved in defining the basic components of the logic.

2. *Declarative Unit*
   This is the basic unit of the logic. In traditional logical systems (such as classical logic, modal logics, linear logic, etc.) the declarative unit is simply a well formed formula. In more complex logics (such as *Labelled Deductive Systems*) it is a labelled formula or a database and a formula, etc.

3. *Databases*
   This notion is that of a family of declarative units forming a *theory* representing intuitively the totality of our *assumptions*, with which we reason. In classical and modal logics this is a set of formulas. In linear logic it is a multiset of formulas. In the Lambek calculus it is a sequence of formulas. In *Labelled Deductive Systems* it is a structured labelled family of formulas.

   In general we need a notion that will include as a database a single declarative unit (compatibility criterion) and allow for more complex structures.

   Other notions need also be defined for databases. Among these are:

   - *Input* and *deletion*, i.e. how to add into and take out declarative units from a database. In classical and modal logic this is union and subtraction.

   - How to *substitute* one database Δ for a declarative unit φ inside another database containing φ. We need this notion to define *cut*.

These notions are purely combinatorial and not logical in nature.[1]

4. *Consequence*

   Now that we have the notion of a database, we can add a notion of consequence. In its simplest form it is a relation between two databases of the form $\Delta_1 \mid\!\sim \Delta_2$. $\mid\!\sim$ needs to satisfy some conditions. It means that Δ_2 *follows* in the logic from Δ_1. Of special interest is the notion $\Delta \mid\!\sim \varphi$, between a database Δ and a declarative unit φ.

   $\mid\!\sim$ can be specified set theoretically or semantically. Depending on its properties, it can be classified as monotonic or non-monotonic.

   As part of the notion of consequence we also include the notions of consistency and inconsistency.

5. *Proof theory, algorithmic presentation*

   One may give, for a given $\mid\!\sim$, an algorithmic system for finding whether $\Delta \mid\!\sim \varphi$ holds for a given Δ and φ. Such an algorithm is denoted by a computable metapredicate $S^{\mid\sim}(\Delta, \varphi)$. Different algorithmic systems can be denoted by further indices, i.e.

   $$S_1^{\mid\sim}(\Delta, \varphi), \ldots, S_i^{\mid\sim}(\Delta, \varphi).$$

   One view is to regard $S_i^{\mid\sim}$ as a mere convenience in generating or defining $\mid\!\sim$ and that the real logic is $\mid\!\sim$ itself.

   However, there are several established proof (algorithmic) methodologies that run across logics, and there are good reasons to support the view that at a certain level of abstraction we should consider any pair $(\mid\!\sim, S^{\mid\sim})$ as a logic. Thus classical logic via tableaux proofs is to be considered as a different logic from classical logic via Gentzen proofs.

6. *Mechanisms*

   The previous items (1)–(5) do not exhaust our list of components for a practical logical system. We need different mechanisms such as *abduction, revision, aggregation, actions*, etc. Such mechanisms make use of the specific algorithm S (of the logic $(\mid\!\sim, S^{\mid\sim})$) and define metalevel operations on a database. The particular version of such operations are considered as part of the logic. Thus a logic can be presented as $(\mid\!\sim, S^{\mid\sim}, S_{abduce}, S_{revise}, \ldots)$.

[1] To explain what we mean, take the Lambek calculus. A declarative unit is any formula φ. A database is any sequence of formulas $\Delta = (\varphi_1, \ldots, \varphi_n)$. Input of φ into Δ can be either at the end of the sequence or the beginning, forming either $(\varphi_1, \ldots, \varphi_n, \varphi)$ or $(\varphi, \varphi_1, \ldots, \varphi_n)$. Similarly the correspondng deletion. Substitution for the purpose of Cut can be defined as follows:

The result of substituting $(\beta_1, \ldots, \beta_m)$ for α_i in $(\alpha_1, \ldots, \alpha_n)$ is the database $(\alpha_1, \ldots, \alpha_{i-1}, \beta_1, \ldots, \beta_m, \alpha_{i+1}, \ldots, \alpha_n)$.

Note that no logical notions are involved, only sequence handling.

> The notion of a database needs to be modified to include markers which can activate these mechanisms and generate more data. The language of the logic may include connectives that activate or refer to these mechanisms. Negation by failure is such an example.

We shall see in lateer that it is convenient to present the metapredicate $S^{\mid\sim}$ as a three-place predicate. $S(\Delta, \varphi, x)$, where $x \in \{0, 1\}$ or $S(\Delta, \varphi) = x$. $S(\Delta, \varphi, 1)$ means that the computation succeeds and $S(\Delta, \varphi, 0)$ means that the computation finitely fails. The definitions of some of the $S_{mechanism}$ will make use of this fine tuning of the S predicate.

The rest of this chapter will motivate and give examples for the above notions.

11.2 Logical Systems

We begin by presenting general answers to:

> *What is a logical system?*
> *What is a monotonic system?*
> *What is a non-monotonic system?*
> *What is a (formal) practical reasoning system?*

and related questions.

Imagine an expert system running on a personal computer, say the *Sinclair QL*. You put the data Δ into the system and ask it queries Q. We represent the situation schematically as:

> $\Delta?Q$ = yes/no depending on the answer.

We can understand the expert system because we know what it is supposed to be doing and we can judge whether its answers make reasonable sense. Suppose now that we spill coffee onto the keyboard. Most personal computers will stop working, but in the case of the QL, it may continue to work. Assume however that it now responds only to the symbol input and output; having lost its natural language interface. We want to know whether what we got is still 'logical' or not. We would not expect that the original expert system still works. Perhaps what we have now is a new system which is still a logic.

We are faced with the question of:

> *What is a logic?*

All we have is a sequence of responses:

> $\Delta_i?Q_i$ = yes/no

How do we recognize whether we have a logic at all?

This question was investigated by Tarski and Scott and they gave us an answer for monotonic logic systems. If we denote the relation

$$\Delta ? Q = \text{yes} \qquad \text{by } \Delta \vdash Q$$

then this relation must satisfy three conditions to be a *monotonic logic*:

1. *Reflexivity*:
   $\Delta \vdash Q$ if $Q \in \Delta$.

2. *Monotonicity* :
   If $\Delta \vdash Q$ then $\Delta, X \vdash Q$.

3. *Transitivity (Lemma Generation, Cut)*:[2]
   If $\Delta \vdash X$ and $\Delta, X \vdash Q$ then $\Delta \vdash Q$.

To present a monotonic logic you have to mathematically define a relation '$\vdash$' satisfying conditions 1, 2, and 3. Such a relation is called a *Tarski consequence relation.* Non-monotonic consequence relations are obtained by restricting condition 2 to be the following condition:

2*. *Restricted Monotonicity:*
If $\Delta \vdash Q$ and $\Delta \vdash X$ then $\Delta, X \vdash Q$.

This condition is discussed later in this section. In such a case we use the symbol '$\mid\!\sim$' instead of '$\vdash$'.

Example 11.1 *Let our language be based on atomic formulas and the single connective '$\Rightarrow$'. Define $\Delta \vdash_C Q$ to hold iff by doing classical truth tables for the formulas in Δ and Q, we find that whenever all elements of Δ get* truth, *Q also gets* truth. *We can check that conditions 1, 2, 3 hold. If so, we have therefore defined a logic. In this particular case, we also have an algorithm to check for a given Δ and Q, whether $\Delta \vdash Q$. In general, consequence relations can be defined mathematically without an algorithm for checking whether they hold or not.*

Example 11.2 *For the same language (with '$\Rightarrow$' only) define $\vdash_I$ as the smallest set theoretical relation of the form $\Delta \vdash Q$ which satisfies conditions 1, 2, 3 together with the condition DT (*Deduction Theorem*):*

[2] Cut has many versions. In classical logic they are all equivalent. In other logics they may not be. Here is another version:

4. *Second Version of Cut* :
   If $\Delta_1 \vdash X$ and $\Delta_2, X \vdash Q$ then $\Delta_1, \Delta_2 \vdash Q$.

We must be careful not to take a version of cut which collapses condition 2* to condition 2.

DT: $\Delta \vdash A \Rightarrow B$ *iff* $\Delta \cup \{A\} \vdash B$.

We have to prove that this is a good definition. First notice that '$\vdash$' is a relation on the set $Powerset(Formulas) \times Formulas$ where $Formulas$ is the set of all formulas. Second we have to show that the smallest consequence relation '$\vdash$' required in the example does exist.

Exercise 11.3 *(a) Prove that $\vdash_I$ of the previous example 11.2 exists. It actually defines intuitionistic implication.*

(b) Let $\Delta \vdash Q$ hold iff $Q \in \Delta$. Show that this is a monotonic consequence relation. (I call this civil servant logic, Beamten Logik*)*

(c) Similarly for $\Delta \vdash Q$ iff $\Delta = \{Q\}$. (This is the literally minded civil servant logic.)

The difference between Examples 11.1 and 11.2 is that Example 11.2 does not provide us with an algorithm of when $\Delta \vdash_I Q$. The $\vdash_I$ is defined implicitly. For example how do we check whether (see Example 11.9.)

$$(((b \Rightarrow a) \Rightarrow b) \Rightarrow b) \Rightarrow a \vdash_I ?a$$

This motivates the need for algorithmic proof procedures.

We thus have the relationships depicted in Figures 11.1, 11.2 and 11.3.

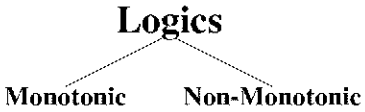


Figure 11.1

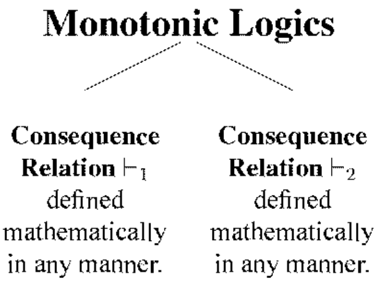

Monotonic Logics
Consequence Relation $\vdash_1$ defined mathematically in any manner.
Consequence Relation $\vdash_2$ defined mathematically in any manner.

Figure 11.2

$S_i^{\vdash}(\Delta, Q)$ is an algorithm for answering whether $\Delta \vdash Q$. Two properties are required:

Soundness If $S_i^{\vdash}(\Delta, Q)$ succeeds then $\Delta \vdash Q$.

Completeness If $\Delta \vdash Q$ then $S_i^{\vdash}(\Delta, Q)$ succeeds.

We assume of course that the algorithmic system is a recursive procedure for generating all pairs (with repetition) (Δ, Q) such that $\Delta \vdash Q$ holds.

There can be many algorithmic systems for the same logic. For example for classical logic, there are resolution systems, connection graph systems, Gentzen systems, semantic tableaux systems, Wang's method, etc.

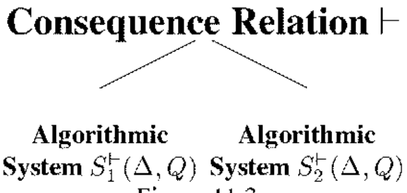
An algorithmic system $S_i(\Delta, Q)$ may not be optimised in practice. It may be, for example, double exponential in complexity, etc. There are several heuristic ways of optimising it. If we try these optimising methods we get into different automated deduction systems for the algorithmic system S_i, denoted by: $O_1 S_i, O_2 S_i, \ldots$

In this case we require only soundness:

if $O_1 S_i(\Delta, Q)$ succeeds then $S_i(\Delta, Q)$ succeeds.

But we do not necessarily require completeness (i.e. the automated system may loop, even though the algorithmic system does not). (In fact, if the relation $\vdash$ is not recursively enumerable (RE), we may still seek an automated system. This will be expected to be only sound).

So far we were talking about monotonic systems. What is a non-monotonic system? Here we have a problem: Can we give conditions on $\Delta \mid\!\sim Q$, to characterize $\mid\!\sim$ as a non-monotonic system? (We use $\vdash$ for monotonic systems and $\mid\!\sim$ for non-monotonic systems). To seek this answer we must look at what is common to many existing non-monotonic systems. Do they have any common features, however weak these common features are? Before we proceed, let us give the main recognisable difference between monotonic and non-monotonic systems. Consider a database (1),(2), (3) and the query $?B$



Consequence Relation $\vdash$

Algorithmic System $S_1^{\vdash}(\Delta, Q)$

Algorithmic System $S_2^{\vdash}(\Delta, Q)$

Figure 11.3

(1) $\neg A$ $\qquad$ $?B$
(2) $\neg A \Rightarrow B$
(3) *other data.*

B follows from (1) and (2). It does not matter what the other data is. We do not need to survey the full database to verify that B follows. This is because of monotonicity. In non-monotonic reasoning, however, the deduction depends on the entire database. Thus if we put in more data, we get a new database and the deduction may not go through. Suppose we agree to list only positive atomic facts in the database. Then negative atomic facts are non-monotonically deduced simply from the fact that they are not listed. Thus a list of airline flights Vancouver–London which does list an 11:05-flight Monday to Saturday, would imply that there is no such flight on Sunday. Thus following this agreement, clause (1) of the database can be omitted provided the database is consistent, i.e. A is not listed in (3). We can deduce B by first deducing $\neg A$ (from the fact that it is not listed) and then deducing B from (2). To make sure A is not listed we must check the entire database.

Our original question was what are the conditions on $\mid\!\sim$ to make it into a non-monotonic logic?

We propose to replace condition 2. on $\vdash$ of monotonicity by condition 2* already mentioned, namely:

2*. *Restricted Monotonicity*:
If $\Delta \mid\!\sim X$ and $\Delta \mid\!\sim Q$ then $\Delta, X \mid\!\sim Q$.

Its meaning is that if X, Q are expected to be true by Δ (i.e. $\Delta \mid\!\sim X$ and $\Delta \mid\!\sim Q$) then if X is actually assumed true, then Q is still expected to be true (i.e. $\Delta, X \mid\!\sim Q$).

Given a non-monotonic system $\mid\!\sim$, we can still ask for algorithmic systems $S_i^{\mid\sim}(\Delta, Q)$ for $\mid\!\sim$. In the non-monotonic case these are rare. Most non-monotonic systems are highly non-constructive and are defined using complex procedures on minimal models or priority of rules or non-provability considerations. So the metatheory for $\mid\!\sim$ is not well developed. There is much scope for research here. Later on we will introduce the notion of *Labelled Deductive Systems* (*LDS*) which will yield in a systematic way a proof theory for many non-monotonic systems.

We are now ready to answer provisionally the question of what is a logical system. We propose a first answer which may need to be modified later on.

Definition 11.4 (Logical System version 1) *A logical system is a pair* $(\Vdash, S^{\Vdash})$, *where* $\Vdash$ *is a consequence relation (monotonic or non-monotonic, according to* *whatever definition we agree on) and $S^{\Vdash}$ is an algorithmic system for $\Vdash$. $S^{\Vdash}$ is sound and complete for $\Vdash$.*[3]

Thus different algorithmic systems for the same consequence relation give rise to different logical systems. So classical logic presented as a tableau system is not the same logic as classical logic presented as a Hilbert system.

Here are some examples of major proof systems:

- Gentzen
- Tableaux
- Semantics (effective truth tables)
- Goal directed methodology
- Resolution
- Labelled Deductive systems, etc.

Definition 11.4 is supported by the following two points:

1. We have an intuitive recognition of the different proof methodologies and show individual preferences to some of them depending on the taste and the need for applications.
2. Slight variations in the parameters of the proof systems can change the logics significantly.

Figure 11.4 is an example of such a relationship.

In the truth table methodology classical logic and Łukasiewicz logic are a slight variation of each other. They resemble intuitionistic logic less closely. In the Gentzen approach, classical and intuitionistic logics are very close while Łukasiewicz logic is a problem to characterise.

This evidence suggests strongly that the landscape of logics is better viewed as a two-dimensional grid.

The reader may ask why we need $\Vdash$, if we have $S^{\Vdash}$ from which $\Vdash$ can be obtained? The answer is that the definition of $\Vdash$ is needed. $\Vdash$ is introduced via a mathematical definition which reveals the intended meaning of $\Vdash$ as separate from the algorithmic means of computing it. So by saying a logical system is a pair $(\Vdash, S^{\Vdash})$ we do not intend only a set theoretical definition of $\Vdash$ and an algorithm $S^{\Vdash}$, but also an expression of the intended meaning for $\Vdash$ as well.

[3] Of course we also need many additional notions such as the notion of what is a theory (database), input into a theory, consistency/inconsistency, etc.

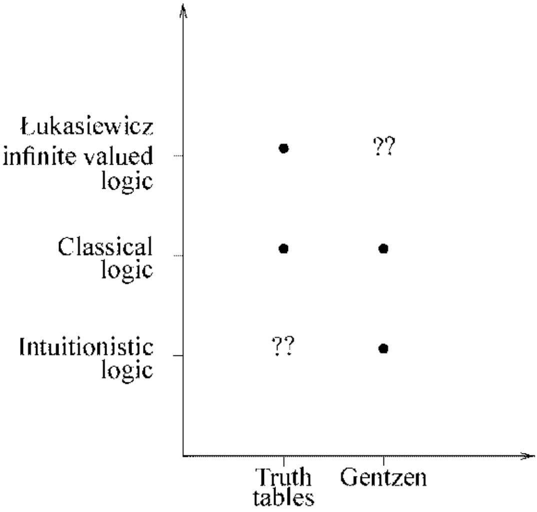


Figure 11.4 Logics landscape

The reader should note that Definition 11.4 is a central definition and a serious departure from current conceptual practice. It will be properly motivated throughout these chapters. Let us use the notation $\tau = (\Vdash_\tau, S^{\Vdash}_\tau)$ for a logical system τ. $\Vdash_\tau$ is its consequence relation and $S^{\Vdash}_\tau$ is an algorithmic system for $\Vdash_\tau$. We also note that in case $\Vdash_\tau$ is not RE, Recursively Enumerable, (as may happen often in non-monotonic logics) we will be satisfied with $S^{\Vdash}_\tau$ which is only sound for $\Vdash$.

Here is a further example:

Example 11.5 *(Modal logic* **S4***):*

1. *Consider a language with atoms p, q, r, … the classical connectives $\neg$, $\wedge$ and the unary connective $\Box$. Let h be a function assigning to each atom a set of points in the Euclidean plane $\mathcal{R}^2$. Let:*

   - $h(A \wedge B) = h(A) \cap h(B)$.
   - $h(\neg A) =$ complement of $h(A)$.
   - $h(\Box A) =$ topological interior of $h(A)$.

   Let $\models A =_{df} \forall h[h(A) = \mathcal{R}^2]$.

   Let $A_1, \ldots, A_n \models B$ *iff* $\forall h(\bigcap_i h(A_i) \subseteq h(B))$.

   Then $\models$ *is a consequence relation.*

Let $A \Rightarrow B$ be defined as $\neg(A \wedge \neg B)$ and let $A \leftrightarrow B$ be defined as $(A \Rightarrow B) \wedge (B \Rightarrow A)$. Then we have for example: $\models \Box(A \wedge B) \Leftrightarrow \Box A \wedge \Box B$.

2. *Let * be a translation from the previous language into classical logic. For each atom q_i associate a unary predicate $Q_i(t)$, with one free variable t. Let R be a binary relation symbol. Translate as follows (note that the translation function depends of t):*

   - $(q_i)^*_t = Q_i(t)$.
   - $(A \wedge B)^*_t = A^*(t) \wedge B^*(t)$.
   - $(\neg A)^*_t = \neg A^*(t)$.
   - $(\Box A)^*_t = \forall s(tRs \Rightarrow (A)^*_s)$.

   Let $A_1, \ldots, A_n \Vdash A$ hold iff in predicate logic one can prove:

   $$\textit{classical logic} \vdash [\forall x(xRx) \wedge \forall xyz(xRy \wedge yRz \Rightarrow xRz)] \Rightarrow \forall t(\textstyle\bigwedge_i (A_i)^*_t \Rightarrow (A)^*_t).$$

   The two consequence relations $\models$ and $\Vdash$ are defined in a completely different way. They are the same, however, from the mathematical point of view. i.e. $\Delta \Vdash A$ iff $\Delta \models A$ holds. Their meaning is not the same.

   To define an algorithmic system for $\Vdash$ or $\models$ we can modify the system in Example 12.7. We can also use any theorem prover for classical logic, if we want, and obtain yet another algorithmic system.

3. *It is possible to give a Hilbert formulation for this consequence relation with the following axioms together with modus ponens (necessitation is derivable):*

   (a) $\Box A$, *where A is an instance of a truth functional tautology.*

   (b) $\Box(\Box(A \Rightarrow B) \Rightarrow (\Box A \Rightarrow \Box B))$

   (c) $\Box(\Box A \Rightarrow \Box\Box A)$

   (d) $\Box A \Rightarrow A$

   (e) $\Box(\Box A \Rightarrow A)$

We have $A_1, \ldots, A_n \vdash B$ iff $\bigwedge_i A_i \Rightarrow B$ is a theorem of the Hilbert system.

This example illustrates our point that even with the same S, $(\models, S)$ and $(\Vdash, S)$ as defined are not the same logic*!*

Let us revisit monotonicity's three levels of presentation. These were:

Mathematical definition of $\vdash$	$\Delta \vdash Q$
Algorithmic procedures (recursively enumerable)	$S_i^{\vdash}(\Delta, Q)$
Optimising automated systems (polynomial time)	$OS_i^{\vdash}(\Delta, Q)$

Experience shows that if we take $S_i^{\vdash}$ and change the computation slightly, to S_i^*, we may get another logic. For example $S(\Delta, Q)$ is an algorithmic system for intuitionistic logic. Change S a little bit to S^*, and $S^*(\Delta, Q)$ gives you classical logic. These type of connections are very widespread, to the extent that we get a much better understanding of a logic $\vdash$, not only through its own algorithmic systems, but also (and possibly even chiefly) through its being a 'changed' version of another automated system for another logic.

It is difficult to appreciate fully what is being said now because we are talking in the abstract, without examples - we will consider some examples later on. There is no choice but to talk abstractly *in the beginning*.

Another important point. This is the role of failure. Assume we have an algorithmic system for some logic. The algorithmic system can be very precise, in fact, let us assume it is an automated system. Suppose we want to ask $\Delta ? Q$. We can look at the rules of the automated system and see immediately that it is looping (a loop checker is needed; we can record the history of the computation and be able to detect that we are repeating ourselves) or possibly finitely failing (i.e. we try all our computation options and in each case we end up in a situation where no more moves are allowed). We add new connectives to the logic, denoted by $loop(Q)$ and $fail(Q)$ and write formulas like $loop(Q) \Rightarrow R$ or $fail(Q) \Rightarrow P$. This is similar to Prolog's negation by failure. We get *new* non-monotonic logics out of old monotonic logics. Here we are connecting the monotonic hierarchy of logics with the non-monotonic one. The abduction mechansim can also be viewed in this way, as $fail(q) \Rightarrow q$. However, this kind of metalevel/object level mixing deserves a full chapter and is postponed to the final volume of our book series.

Our final point here concerns the general nature of theorem proving (or automated reasoning). So far we talked about $S_i^{\vdash}(\Delta, Q)$. All our automated rules have the form e.g. $S_i^{\vdash}(\Delta, Q \Rightarrow R)$ if $S_i^{\vdash}(\Delta \cup \{Q\}, R)$; in other words, the success of $\Delta \vdash Q \Rightarrow R$ reduces to that of $\Delta, Q \vdash R$ *in the same logic* $\vdash$. This suggests the need for taking the logic as a *parameter*, in view of the fact that we get things like $S(\Delta, Q, \text{logic } 1)$ if $S(\Delta', Q', \text{logic } 2)$.

We thus have automated system defining several logics by mutual recursion.

Example 11.6 *Consider a language with* $\wedge$, $\Rightarrow$, $\neg$, $\vee$ *and* $\Box$. *Define a Hilbert system for a modal logic* **K** *for modality* $\Box$. **K** *has the axioms and rules as follows:*

1. *Any instance of a truth functional tautology.*

2. $\Box(A \Rightarrow B) \Rightarrow (\Box A \Rightarrow \Box B)$

3. *The rules:*

$$\frac{\vdash A, \vdash A \Rightarrow B}{\vdash B} \quad \textit{and} \quad \frac{\vdash A}{\vdash \Box A}$$

Define $A_1, \ldots, A_n \vdash_{\mathbf{K}} A$ *iff* $\vdash \bigwedge A_i \Rightarrow A$. *It is complete for all Kripke structures.*

Consider the additional condition on Kripke structures that only the actual world is reflexive (i.e. we require aRa *but not generally* $\forall x(xRx)$*). The class of all models with* aRa *defines a new logic). Call this logic* **K1**. *We cannot axiomatise this logic by adding the axiom* $\Box A \Rightarrow A$ *to* **K** *because this will give us the logic* **T** *(complete for reflexivity of every possible world, not just the actual world). We observe, however, that* $\vdash_{\mathbf{K}} A$ *implies* $\vdash_{\mathbf{K1}} \Box A$. *If we adopt the above rule together with* $\vdash_{\mathbf{K1}} \Box A \Rightarrow A$, *we will indeed get, together with modus ponens an axiomatization of* **K1**.

The axiomatization of **K1** *is obtained as follows:*

Axioms:

1. *Any substitution instance of a theorem of* **K**.
2. $\Box A \Rightarrow A$.

Rules: *Modus ponens only.*

We can define an extension of **K1**, (call it $\mathbf{K1}_{[2]}$), by adding a $\Box^2$ necessitation rule

$$\frac{\vdash A}{\vdash \Box^2 A}.$$

Example 11.7 (Classical logic with restart) *The following is an algorithmic (possibly a phantom algorithms) presentation of classical propositional logic. (See [Gabbay, 1998a] for more details). We choose a formulation of classical logic using* $\wedge, \rightarrow, \bot$. *We write the formulas in a 'ready for computation' form as clauses as follows:*

1. q *is a clause, for* q *atomic, or* $q = \bot$

2. *If* $B_j = \bigwedge_i A_{ij} \rightarrow q_i$ *are clauses for* $j = 1, \ldots, k$ *then so is* $\bigwedge_j B_j \rightarrow q$, *where* q *is atomic or* $q = \bot$.

3. *It can be shown that every formula of classical propositional logic is classically equivalent to a conjunction of clauses.*

We now define a goal directed computation for clauses.

4. *Let $S(\Delta, G, H)$ mean the clause G succeeds in the computation from the set of clauses Δ given the history H, where the history is a set of atoms or $\bot$.*

   (a) *$S(\Delta, q, H)$ if $q \in \Delta$ for q atomic or $\bot$.*

   (b) *$S(\Delta, q, H)$ if $\Delta \cap H \neq \varnothing$.*

   (c) *$S(\Delta, \bigwedge_j(\bigwedge_i A_{ij} \to q_j) \to q, H)$ if $S(\Delta \cup \{\bigwedge_i A_{ij} \to q_j\}, q, H)$.*

   (d) *$S(\Delta \cup \{\bigwedge_{j=1}^{k}(\bigwedge_{i=1}^{m_j} A_{ij} \to q_j) \to q)\}, q)$ if $\bigwedge_j S(\Delta \cup \{A_{ij} \mid i = 1, \ldots, m_j\}, q_j, H \cup \{q\})$*

5. *The computation starts with $S(\Delta, G, \varnothing)$, to check whether $\Delta \vdash G$.*

We have the following theorem:

$$S(\Delta, q, H) \textit{ iff } \Delta \vdash q \vee \bigvee H.$$

The next example is a challenging example of a modal intuitionistic algorithmic system defined for the modality $\Diamond$, $\Rightarrow$, $\vee$, and $\bot$. To understand the intuition behind this example, imagine a family of possible worlds of the form (T, R, now), where T is the set of worlds, R the accessibility relation and now is the actual world. Syntactically we write $t : A$ to mean that A is assumed to hold at world t. A theory is a set $\Delta = \{t_i : A_i\}$ and some relation R among the labels $\{t_i\}$ of Δ. So for example (Δ, R), with $\Delta = \{t_1 : A_1, t_2 : A_2\}$ and $R = \{(t_1, t_2)\}$ is a theory which says that there are two worlds t_1 and t_2, t_2 accessible to t_1 and A_1 holds at t_1 and A_2 at t_2.

For a given world t the local reasoning is intuitionistic. Thus $S(\Delta, T, R, t, Q)$ reads that the theory (Δ, R) based on labels from T and accessibility R has the property that at the point t the wff Q can be proved.

In our example $S(\Delta, T, R, t_1, A_1)$ holds.

Exercise 11.8 *Define a Horn clause with negation by failure as any wff of the form $\bigwedge a_i \wedge \bigwedge \neg b_j \Rightarrow q$, where a_i, b_j, q are atomic; q is called the head of the clause. Define a computation as follows:*

1. *$\Delta?q = success$ if $q \in \Delta$.*

2. *$\Delta?q = failure$ if q is not head of any clause in Δ.*

3. *$\Delta?q = success$ if for some $\bigwedge a_i \wedge \bigwedge \neg b_j \Rightarrow q$ in Δ we have that $\Delta?a_i = success$ for all i and $\Delta?\neg b_j = success$ for all j.*

4. *$\Delta?\neg b = success$ (respectively failure) if $\Delta?b = failure$ (respectively success).*

5. *$\Delta?q = failure$ if for each clause $\bigwedge a_i \wedge \bigwedge \neg b_j \Rightarrow q \in \Delta$ we have either for* some *i, $\Delta?a_i = failure$ or for some j, $\Delta?\neg b_j = failure$.*

(a) *Show that propositional Horn clause Prolog with negation by failure is a non-monotonic system, i.e. if we define $\Delta \mid\!\sim Q$ iff (definition) Q succeeds in Prolog from data Δ, then $\mid\!\sim$ is a non-monotonic consequence relation according to our definition. (To show (3) and (2*) assume X is positive.)*

(b) (Challenge) *Prove in Prolog $\mid\!\sim$ that:*

$$\textit{if } \Delta \mid\!\sim q \textit{ and } \Delta, d \mid\!\sim \neg q \textit{ then } \Delta \mid\!\sim \neg d \textit{ (for } d \textit{ atomic).}$$

Similarly,

$$\textit{if } \Delta \mid\!\sim \neg q \textit{ and } \Delta, d \mid\!\sim q \textit{ then } \Delta \mid\!\sim \neg d.$$

Example 11.9 *Recall the definition of $\Phi \vdash_I A$ as the smallest Tarski relation on the language with $\Rightarrow$ such that the equation DT holds.*

DT: $\Phi \vdash A \Rightarrow B$ *iff* $\Phi, A \vdash B$.

According to Exercise 11.3 from this chapter, $\vdash_I$ exists.

Our algorithmic problem is how do we show for a given $\Phi \vdash_I ?A$ whether is holds or not.

Take the following (Hudelmaier):

$$(((b \Rightarrow a) \Rightarrow b) \Rightarrow b) \Rightarrow a \vdash_I ?a$$

We can only use the means at our disposal, in this case DT. Certainly, for cases of the form $\Phi \vdash_I A \Rightarrow B$ we can reduce to $\Phi, A \vdash_I B$. This is a sound policy, because we simplify the query and monotonically strenghten the data at the same time.

When the query is atomic we cannot go on. So we must look for patterns.

Out of a, b we can make the following possible formulas with at most one $\Rightarrow$:

$a \Rightarrow b$
a
b
$b \Rightarrow a$
$a \Rightarrow a$
$b \Rightarrow b$

The following are possible consequences:

$a, b \vdash ?a \Rightarrow b$
$a, a \Rightarrow b \vdash ?b$
$a \Rightarrow b \vdash ?a \Rightarrow b$
$a \Rightarrow a \vdash ?b \Rightarrow b$

Shuffling around and recognizing cases of reflexivity we can get:

1. $a, a \Rightarrow b \vdash b$ *(from reflexivity and DT)*

2. $\vdash a \Rightarrow a$

3. $b \vdash a \Rightarrow b$

Back' to our example:

$(((b \Rightarrow a) \Rightarrow b) \Rightarrow b) \Rightarrow a$	$\vdash_I ?a$
same	$\vdash_I ?((b \Rightarrow a) \Rightarrow b) \Rightarrow b$
add $(b \Rightarrow a) \Rightarrow b$	$\vdash_I ?b$
same	$\vdash_I ?b \Rightarrow a$
add b	$\vdash_I ?a$
same	$\vdash_I ?((b \Rightarrow a) \Rightarrow b) \Rightarrow b$
same	$\vdash_I ?b$ *(*success*: you already have b)*

11.3 Refining the Notion of a Logical System

If we look at the kind of applications studied in this book we see that we need a more refined notion of a logical system, to enable us to cope with the needs of the applications. This section surveys our options.

11.3.1 Structured consequence

The next move in the notion of a logical system is to observe that part of the logic must also be the notion of what the logic accepts as a theory. Theories have structure, they are not just sets of wffs. They are structures of wffs. Different notions of structures give rise to different logics, even though the logics may share the same notion of consequence for individual wffs.

We need the following:

1. A notion of structure to tell us what is to be considered a theory for our logic. For example, a theory may be a multiset of wffs or a list of wffs or a more general structure. The best way to define the general notion of a structure is to consider models M of some classical structure theory τ and a function $\mathbf{f} : M \mapsto$ wffs of the logic. A theory Δ is a pair $(M, \mathbf{f})$. We write $t : A$ to mean $\mathbf{f}(t) = A$.

2. A notion of insertion into the structure and deletion from the structure. I.e. if $t \in M$ and $s \not\in M$ we need to define $M' = M + \{s\}$ and $M'' = M - \{t\}$. M' and M'' must be models of τ. When theories were sets of wffs, insertion and deletion presented no problems. For general structures we need to specify how it is done.

3. A notion of substitution of one structure M_1 into another M_2 at a point $t \in M_2$. This is needed for the notion of cut. If $(M_2, \mathrm{f}_2) \mid\!\sim A$ and for some $t \in M_2, (M_1, \mathrm{f}_1) \mid\!\sim \mathrm{f}_2(t)$, we want to 'substitute' '(M_1, f_1)' for 't' in M_2 to get (M_3, f_3) such that $(M_3, \mathrm{f}_3) \mid\!\sim A$.

Example 11.10 Let the structure be lists. So let, for example, $\Delta = (A_1, \ldots, A_n)$. We can define $\Delta + A = (A_1, \ldots, A_n, A)$ and $\Delta - \{A_i\} = (A_1, \ldots A_{i-1}, A_{i+1}, \ldots, A_n)$. Substitution of $\Gamma = (B_1, \ldots, B_k)$ for place i in Δ (replacing A_i) gives $\Delta[i/\Gamma] = (A_1, \ldots, A_{i-1}, B_1, \ldots, B_k, A_{i+1}, \ldots, A_n)$.

The cut rule would mean

- $\Delta \mid\!\sim C$ and $\Gamma \mid\!\sim A_i$ imply $\Delta[i/\Gamma] \mid\!\sim C$.

We need now to stipulate the minimal properties of a structured consequence relation. These are the following:

Identity $\{t : A\} \mid\!\sim t : A$

Surgical Cut $\dfrac{\Delta \mid\!\sim t : A; \Gamma[t : A] \mid\!\sim s : B}{\Gamma[t/\Delta] \mid\!\sim r : B}$

Typical good examples of structured consequence relations are algebraic labelled deductive systems based on implication $\to$.

The basic rule is modus ponens.

- $\dfrac{s : A \to B; t : A; \varphi(s, t)}{f(s, t) : B}$

t, s are labels, φ is a compatibility relation on labels and $f(s, t)$ is the new label.

In terms of databases the rule means as in Figure 11.5

The labels can be resource, time, relevance, strength/reliability, complete file. See [Gabbay, 1996].

11.3.2 Algorithmic structured consequence relation

Just as in the case of ordinary consequence relations, different algorithms on the structure are considered as different logics. The general presentation form of most (maybe all) algorithms is through a family of rules of the form

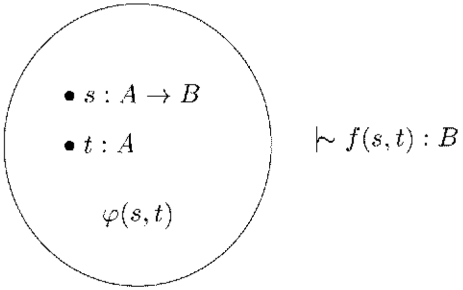


Figure 11.5

- $\Delta \mid\!\sim ?\Gamma, \Pi$ reduces to $\Delta_i \mid\!\sim ?\Gamma_i, \Pi_i, i = 1, \ldots, n$;

if we use a traditional method of display we will write:

- $\dfrac{\Delta_1 \mid\!\sim \Gamma_1, \Pi_1; \ldots; \Delta_n \mid\!\sim \Gamma_n, \Pi_n}{\Delta \mid\!\sim \Gamma, \Pi}$.

We need the notion of a formula $t : A$ in the structure Π *being used* in the rule. We further need to have some complexity measure available which decreases with the use of each rule. $\Delta, \Gamma, \Delta_i, \Gamma_i$ are structured theories and Π, Π_i are parameters involved in the algorithm, usually the history of the computation up to the point of application of the rule. There may be side conditions associated with each rule restricting its applicability.

Some rules have the form

- $\Delta \mid\!\sim ?\Gamma, \Pi$ reduces to *success*

or in a traditional display

- $\dfrac{}{\Delta \mid\!\sim \Gamma, \Pi}$.

These are axioms.

Having the rules in this manner requires additional (decidable) metapredicates on databases and parameters. We need the following:

1. $\Psi_0(\Delta, \Gamma, \Pi)$ recognising that $\Delta \mid\!\sim \Gamma, \Pi$ is an axiom.
2. $\Psi_1(\Delta, \Gamma, \Pi)$ recognising that $\Delta \mid\!\sim \Gamma, \Pi$ is a failure (no reduction rules apply).
3. $\Psi_2(\Delta, \Delta_1, \Delta_2, \Pi, \Pi_1, \Pi_2)$ says that Δ is the result of inserting Δ_2 into Δ_1 (also written as $\Delta = \Delta_1 + \Delta_2$, ignoring the parameters).

4. $\Psi_4^n(\Delta, \Delta_1, \ldots, \Delta_n, \Pi, \Pi_1, \ldots, \Pi_n)$ says that Δ is decomposed into $\Delta_1, \ldots, \Delta_n$. The decomposition is not necessarily disjoint. So for example, the reduction rule $\Delta \mid\!\sim ?\Gamma$ if $\Delta_i \mid\!\sim ?\Gamma_i, i = 1, \ldots, n$ may require that $\Psi_4^n(\Delta, \Delta_1, \ldots, \Delta_n)$ and $\Psi_4^n(\Gamma, \Gamma_1, \ldots, \Gamma_n)$ both hold, (ignoring the parameters).

There may be more Ψs involved.

The Ψs may be related. For example $\Psi_4^n(\Delta, \Delta_1, \ldots, \Delta_n)$ may be $\Delta = (\Delta_1 + (\Delta_2 + (\ldots + \Delta_n)\ldots)$.

The reader should note that with structured databases, other traditional notions associated with a logic change their relative importance, role and emphasis. Let us consider the major ones.

Inconsistency

In traditional logics, we reject inconsistent theories. We do not like having both A and $\neg A$ among the data. In structured databases, we have no problem with that. We can have $\Delta = \{t : A, s : \neg A\}$ and what we prove from Δ depends on the logic. Even when we can prove everything from a database, we can use labels to control the proofs and make distinctions on how the inconsistency arises. A new approach to inconsistency is needed for structured databases. Inconsistency has to be redefined as *acceptability*. Some databases are not acceptable. They may be consistent or inconsistent. Consistency is not relevant, what is relevant is their acceptability.

We have, for example, the notion of integrity constraints in logic and commercial databases. We may have as an integrity constraint for a practical database that it must list with each customer's name also his telephone number. A database that lists a name without a telephone number may be consistent but is unacceptable. It does not satisfy integrity constraints. See [Gabbay and Hunter, 1991] for more discussion.

Deduction theorem

If a theory Δ is a set of sentences we may have for some A, B that $\Delta \not\mid\!\sim B$ but $\Delta \cup \{A\} \mid\!\sim B$. In which case we can add a connective $\twoheadrightarrow$ satisfying $\Delta \mid\!\sim A \twoheadrightarrow B$ iff $\Delta \cup \{A\} \mid\!\sim B$.

If Δ is a structured database, we have an insertion function $\Delta + A$, adding A into Δ and a deletion function, $\Delta - A$, taking A out of Δ. We can stipulate that $(\Delta + A) - A = \Delta$.

The notion of deduction theorem is relative to the functions $+$ and $-$. Let $\twoheadrightarrow_{(+,-)}$ be the corresponding implication, then we can write $\Delta + A \mid\!\sim B$ iff $(\Delta + A) - A \mid\!\sim A \twoheadrightarrow_{(+,-)} B$.

We can have many $\twoheadrightarrow$s and many deduction theorems for many options of $(+, -)$ pairs.

Take, for example, the Lambek calculus.

The databases are lists, $(A_1, \ldots, A_n)$ of wffs. We can have two sets of $+$ and $-$, namely we can have $+_r, -_r$ (add and delete from the right hand side) and $+_l, -_l$ (add and delete from the left). This gives us two arrows and two deduction theorems

$$(A_1, \ldots, A_n) \mid\!\sim A \rightarrow_r B \text{ iff } (A_1, \ldots, A_n, A) \mid\!\sim B$$

and

$$(A_1, \ldots, A_n) \mid\!\sim A \rightarrow_l B \text{ iff } (A, A_1, \ldots, A_n) \mid\!\sim B.$$

In fact the smallest bi-implicational logic with $\rightarrow_r, \rightarrow_l$ only and databases which are lists which satisfy the two deduction theorems is the Lambek calculus.

11.3.3 Mechanisms

Our notion of logical systems so far is of pairs $(\mid\!\sim, S^{\mid\sim})$, where $\mid\!\sim$ is a structured consequence relation between structured databases and $S^{\mid\sim}$ is an algorithm for computing $\mid\!\sim$. The data items in any database $\Delta = (M, \mathbf{f})$ are wffs, i.e. for $t \in M, \mathbf{f}(t)$ is a wff. We know from many applications that items residing in the database need not be the data itself but can be mechanisms indicating how to obtain the data. Such mechanisms can be sub-algorithms which can be triggered by the main $S^{\mid\sim}$ algorithm or in the most simple case just a link to another database.

We regard such mechanisms as part of the logic.[4] They are ways of extending our databases Δ with more data without having to put them explicitly into the structure of Δ.

Among the well known mechanisms are

- abduction
- default
- other non-monotonic mechanisms (circumscription, negation as failure, etc).

Figure 11.6 shows what a database looks like. Formally a database $\Delta = (M, \mathbf{f})$ can be such that for $t \in M, \mathbf{f}(t)$ is a mechanism. Of course these mechanisms depend on $S^{\mid\sim}$.

These mechanisms are well studied in the literature but the traditional way of perceiving them is that they depend on $\mid\!\sim$ and not necessarily on the proof method $(S^{\mid\sim})$. Our view is that

[4] Note that we have already said that we need the notion of a logic to model abduction and here we say that abduction is a mechanism which can be part of a logic. Obviously this is an interactive recursive process.

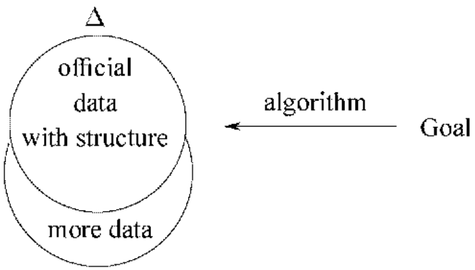


Figure 11.6 Mechanisms extend data

1. they are part of Δ;

and that

2. they depend on $S^{\mid\!\sim}$.

   Thus a database may be a list of the following form.

$$\begin{array}{rcl}\Delta & = & (A_1, A_2, \text{use abduction algorithm } \mathbf{Ab}_1, A_4, \\ & & \text{use default algorithm } \mathbf{D}_1, A_6)\end{array}$$

3. In fact each item of data in Δ may come with a little *algoithmic patch*, giving the main algorithm $S^{\mid\!\sim}$ extra or less freedom in using this item of data. A well known example of such a 'patch' is the exclamation mark in linear logic. $!A$ means 'you can use A as many times as you need'. The patch can interact with the *mode* (see subsection 11.3.4 below) and change it to a new mode. See footnote 5 p. 388.

The algorithm $S^{\mid\!\sim}$ may approach the third item in the list with a view of trying to succeed. Instead of a data item it finds an algorithm $\mathbf{Ab}_1$. It exchanges information with $\mathbf{Ab}_1$ and triggers it. What $\mathbf{Ab}_1$ does now depends on the state of the $S^{\mid\!\sim}$ algorithm when it 'hits' $\mathbf{Ab}_1$. $\mathbf{Ab}_1$ returns with some additional data, say $(B_1, \ldots, B_k)$. The $S^{\mid\!\sim}$ algorithm can now continue *as if* the database were $(A_1, A_2, (B_1, \ldots, B_k), A_4, \mathbf{D}_1, A_6)$. There are two ways of looking at mechanisms. One way is that they are some sort of algorithmic shorthand for additional data, like links to other places where data can be found. Thus according to this view $\Delta +$ *mechanism* $= \Delta + \Delta'$, where $\Delta' =$ data obtained by *mechanism* applied to Δ. This is a traditional view, since a theory Δ is still a (structured) set of data. The second view, which is the one we want to adopt is that procedures are themselves data. This view is better because of two reasons.

1. The mechanisms may yield different results depending on when in the computation they are being used (‘hit’), thus making it difficult to say what is the declarative content of the theory.

2. Since we accept the proof theory as part of the logic, we can go all the way and accept additional mechanisms and patches to the proof theory as part of the data. Thus different databases may include as part of themselves additional rules to be used to prove more (or prove less) from themselves. In fact each item of data (declarative unit) can carry as part of its label a patch on the computation $S^{\mid\sim}$.

This idea is pretty revolutionary because it also gives up the current received view that theories (data) must have a declarative content.

11.3.4 Modes of Evaluation

When we present logics semantically, through say Kripke models, there are several features involved in the semantics. We can talk about Kripke frames, say of the form (S, R, a), where S is a set of possible worlds, $R \subseteq S^{n+1}$ is the accessibility relation (for an n-place connective $\sharp(q_1, \ldots, q_n)$) and $a \in S$ is the actual world. We allow arbitrary assignments $h(q) \subseteq S$ to atomic variables q and the semantical evaluation for the non-classical connective $\sharp$ is done through some formula $\Psi_\sharp(t, R, Q_1, \ldots, Q_n)$, $Q_i \subseteq S$ in some language. We have:

- $t \vDash \sharp(q_1, \ldots, q_n)$ under assignment h iff $\Psi_\sharp(t, R, h(q_1), \ldots, h(q_n))$ holds in (S, R, a).

For example, for a modality $\Box$ we have

- $t \vDash \Box q$ iff $\forall s(tRs \to s \vDash q)$.

Here $\Psi_\Box(t, h(q)) = \forall s(tRs \to s \in h(q))$. Different logics are characterised by different properties of R.[5]

We can think of $\Psi_\sharp$ as the *mode of evaluation* of $\sharp$. The mode is fixed throughout the evaluation process.

In the new concept of logic, mode shifting during evaluation is common and allows for the definition of many new logics. We can view the mode as the recipe for where to look for the accessible points s needed to evaluate $\Box A$.

Consider the following:

- $t \vDash \Box A$ iff $\forall n \forall s(tR^n s \to s \vDash A)$
  where $xR^n y$ is defined by the clauses:

[5] Some logics presented axiomatically, cannot be characterised by properties of R alone but the family of allowed assignments h needs to be restricted. This is a minor detail as far as the question of ‘what is a logic’ is concerned.

- xR^0y iff $x = y$
- $xR^{n+1}y$ iff $\exists z(xRz \wedge zR^ny)$.

Clearly $\{(t, s) \mid \exists ntR^ns\}$ is the transitive and reflexive closure of R.

Thus in this evaluation mode, we look for points in the reflexive and transitive closure of R.

We can have several evaluation modes available over the same frame (S, R, a).

Let $\rho_i(x, y, R), i = 1, \ldots, k$, be a family of binary formulas over (S, R, a), defined in some possibly higher-order mode language $\mathcal{M}$, using R, a and h as parameters.

We can have a mode shifting function $\varepsilon : \{1, \ldots, k\} \mapsto \{1, \ldots, k\}$ and let

- $t \vDash_i \Box A$ iff for all s such that $\rho_i(t, s, R)$ holds we have $s \vDash_{\varepsilon(i)} A$.

Example 11.11 *Consider now the following definition for* $\vDash$ *for two modes* ρ_0 *and* ρ_1 *and* $x = 0$ *or* 1*:*

- $t \vDash_x q$ *for* q *atomic iff* $t \in h(q)$
- $t \vDash_x \neg A$ *iff* $t \nvDash_x A$
- $t \vDash_x A \wedge B$ *iff* $t \vDash_x A$ *and* $t \vDash_x B$
- $t \vDash_x \Box A$ *iff for all* s *such that* $\rho_x(t, s)$ *holds we have* $s \vDash_{1-x} A$.

We see that we have a change of modes as we evaluate.

We are thus defining $\vDash_0$ not independently on its own but together with $\vDash_1$ in an interactive way.

We repeat here Example 1.5 of [Gabbay, 1998b].

Example 11.12 *We consider two modes for a modality* $\Box$.

$$\begin{array}{rcl} \rho_1(x, y) & = & xRy \vee x = y \\ \rho_1(x, y) & = & xRy. \end{array}$$

Define a logic $\mathbf{K1}_{[2]}$ *as the family of all wffs* A *such that for all models* (S, R, a) *and all assignments* h *we have* $a \vDash_0 A$.

In such a logic we have the following tautologies.

$$\begin{array}{l} \vDash \Box A \to A \\ \nvDash \Box(\Box A \to A) \\ \vDash \Box^{2n}(\Box A \to A) \end{array}$$

It is easy to see that this logic cannot be characterised by any class of frames. For let (S, R, a) be a frame in which all tautologies hold and $\Diamond(\neg q \wedge \Box q)$ holds. Then aRa must hold since $\Box A \to A$ is a tautology for all A. Also there must be a point t, such that aRt and $t \vDash \neg q \wedge \Box q$. But now we have $aRaRt$ and this falsifies $\Box^2(\Box A \to A)$ which is also a tautology.

This logic, $\mathbf{K1}_{[2]}$, can be axiomatised.

The idea of a mode of evaluation is not just semantical. If we have proof rules for $\Box$, then there would be a group of rules for logic $\mathbf{L}_1$ (say modal **K**) and a group for $\mathbf{L}_2$ (say modal **T**). We can shift modes by alternating which group is available after each use of a $\Box$ rule.

This way we can jointly define a family of consequence relations $\mid\!\sim_\mu$ dependent on a family of modes $\{\mu\}$.

We believe mode evaluation and mode shifting is an important concept in proof theory and semantics.

The notion of mode can be attached to data items. Since mode means which proof rules are available, data items can carry mode with them and when they are used they change the mode. This is fully compatible with our approach that procedures are part of the data.

11.3.5 TAR-Logics

We are now ready to introduce TAR-logics. First let us summarise what we have got so far:

Logical Systems

- Structured data;
- algorithmic proof theory on the structure;
- mechanisms make use of data and algorithms to extend data;
- inconsistency is no longer a central notion. It is respectable and is most welcome; acceptability is the right notion;
- the deduction theorem is connected with cut and the insertion and deletion notions.

Notice the following two points about the current notion of a logical system:

- time and actions are not involved;
- proofs and answers are conceptually instantaneous.

Our new notion of logic and consequence shall make the following points:

- proofs take time (real time!);
- proofs involve actions and revisions;
- logics need to be presented as part of a mutually dependent family of logics of various modes.

We explain the above points.

In classical geometry, we have axioms and rules. To prove a geometrical theorem may take 10 days and 20 pages, but the time involved is not part of the geometry. We can conceptually say that all theorems follow instantaneously from the axioms.

Let us refer to this situation as a *timeless* situation.

Our notion of a logic developed so far is timeless in this sense. We have a structured database Δ, we have a consequence relation $\mid\!\sim$, we have an algorithm $\mathcal{S}^{\mid\sim}$, we have various mechanisms involved and they all end up giving us answers to the timeless question: does $\Delta \mid\!\sim ?\Gamma$ hold?

In practice, in many applications where logic is used, time and actions are heavily involved. The deduction $\Delta \mid\!\sim ?\Gamma$ is not the central notion in the application, it is only auxiliary and marginal. The time, action, payoffs, planning and strategic considerations are the main notions and the timeless consequence relation is only a servant, a tool that plays a minor part in the overall picture. In such applications we have several databases and queries $\Delta_i \mid\!\sim ?\Gamma_i$ arising in different contexts. The databases involved are ambiguous, multiversion and constantly changing. The users are in constant disagreement about their contents. The logical deductive rules are non-monotonic, commonsense and have nuances that anticipate change and the reasoning itself heavily involves possible, alternative and hypothetical courses of action in a very real way. The question of whether $\Delta_i \mid\!\sim ?\Gamma_i$ hold plays only a minor part, probably just as a question enabling or justifying a sequence of actions.

It is therefore clear that to make our logics more practical and realistic for such applications, we have no alternative but to bring in these features as serious components in our notion of what is a logic.

Further, to adequately reason and act we need to use a family of different logics at different times, $\Delta_i \mid\!\sim_i ?\Gamma_i$, and their presentation and proof theory are interdependent. We anticipate a heavy use of modes in the deduction.

Let $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ be a family of actions of the form $\mathbf{a}_i = (\alpha_i, \beta_i)$, where α_i is the precondition and β_i is the post condition. Let $*$ be a revision operation such that for a given Δ and α $\Delta * \alpha$ is the result of inputting α into Δ and then revising the new theory into an *acceptable* new theory $\Delta * \alpha$.

We define by induction on n the notion $\Delta \mid\!\sim_{(\mathbf{a}_1, \ldots, \mathbf{a}_n)}$ as follows:

- $\Delta \mid\!\sim_{\varnothing} A$ iff $\Delta \mid\!\sim A$.

- $\Delta \mid\!\sim_{(\mathbf{a}_1,\ldots,\mathbf{a}_{n+1})} A$ iff $\Delta \mid\!\sim_{\varnothing} \alpha_1$ and $\Delta * \beta_1 \mid\!\sim_{(\mathbf{a}_2,\ldots,\mathbf{a}_{n+1})} A$.

In $\mid\!\sim_{(\mathbf{a}_1,\ldots,\mathbf{a}_n)}$ the sequence $(\mathbf{a}_1,\ldots,\mathbf{a}_n)$ acts as a mode of provability. See Gabbay [2001b].

11.3.6 Relevance

Relevance is a central concept for the general theory of logical systems. In fact Volume 1 of this series of books [Gabbay and Woods, 2003a] is devoted to the notion. A logical system should be presented as ($\mid\!\sim$, $S^{\mid\!\sim}$, *Relevance*, *Mechanisms*). We would not go into details here but only hint. In non-monotonic logic, the fact that a Δ proves (or does not prove) A can be changed by adding (or deleting) items of data. In a real life argument involving a database, one can argue that more data is 'relevant' and hence by adding the relevant data the provability of A can change. Without a proper notion and algorithmic criteria for relevance in the logic any A can be proved or not proved by simply extending the data at will.

11.4 Discussion and Further Reading

For more about logical systems and consequence relation see [Gabbay, 1996]. For the goal directed computation see [Gabbay, 1998a] and [Gabbay and Olivetti, 2000]. For an approach to the dynamics of practical reasoning see [Gabbay and Woods, 2003a]. For an account of actions as premisses see [Gabbay, 2001a].

Chapter 12

A Base Logic

"I presume nothing!"

Sherlock Holmes

12.1 Formal Abduction: An Overview

In earlier chapters we saw that a formal logic of abduction needs, among other things, the following components.

1. A **base logic** $\mathbf{L}_1$, with proof procedures Π.
2. An **abductive algorithm** which deploys Π to look for missing premisses and other formulas to be abduced.
3. A further logic $\mathbf{L}_2$ for deciding which abduced formulas to choose, which criteria of selection apply, etc. Components 2 and 3 together form **the logic of discovery**.

It is clear that our account of abduction must give a principled description of each component. In the previous chapter we discussed base logics. In the present chapter we turn our attention to the **logic of discovery**, that is to the abductive mechanisms together with the $\mathbf{L}_2$ logic.

Here is the motivating insight as to how abductive mechanisms work. Logical agents have goals. Given a goal G and database Δ we mimic the agent's pursuit of G by deploying Π for the purpose of proving G. In its normal operation, the proof mechanism will branch into various alternative strategies. If one is not successful there will be various junctures, J, of local failure. At those points, because of the algorithmic nature of the proof theory, it is always clear what is locally needed to carry on. An abductive mechanism for the logic should tell us how to construct the abduction options to be added to Δ from what is (locally abduced as) missing at local failure points J.

For ease of exposition the account here developed of the interaction between the abductive mechanism and the $\mathbf{L}_2$ logic is highly explanary. But it is easy to see how the example generalizes.

Our base logic is a goal-directed labelled deductive system for implication. Such a logic is rich enough to include a large variety of human logics; and that for us is an attractive feature. This chapter introduces this logic and introduces the features needed for the operation of the abductive mechanism. The mechanism itself is described in the chapter to follow.

So we begin.

Given a logical system $\mathbf{L}$ ($\mathbf{L}$ may be a monotonic or non-monotonic logic) a theory Δ in $\mathbf{L}$ and a wff Q there are several possible interactions to consider between $\mathbf{L}, \Delta$ and Q.

1. *Consequence*
   We can ask whether $\Delta \vdash_{\mathbf{L}} Q$

2. *Retraction*
   If $\Delta \vdash_{\mathbf{L}} Q$ holds, we can ask how do we revise Δ to a theory Δ' (if $\mathbf{L}$ is a monotonic logic we require that $\Delta' \subseteq \Delta$) such that $\Delta' \nvdash_{\mathbf{L}} Q$.

3. *Consistency/Acceptability*
   If $\Delta \nvdash_{\mathbf{L}} Q$, we can ask whether $\Delta \cup \{Q\}$ is consistent or acceptable in $\mathbf{L}$.

4. *Revision*
   If $\Delta \cup \{Q\}$ is not consistent or not acceptable, we can ask how do we revise $\Delta \cup \{Q\}$ to a Δ' preferably satisfying $\Delta' \subseteq \Delta \cup \{Q\}$, such that $\Delta' \vdash_{\mathbf{L}} Q$ and Δ' is consistent/acceptable.

5. *Abduction*
   If $\Delta \nvdash_{\mathbf{L}} Q$, we can ask what X can we add to Δ to get $\Delta' = \Delta \cup \{X\}$ such that $\Delta \cup \{X\} \vdash_{\mathbf{L}} Q$. (If $\mathbf{L}$ is non-monotonic we may delete X from Δ to get a Δ' such that $\Delta' \vdash Q$).

   We would, of course, like X to be as logically weak as possible.[1]

A possible answer to (1) is a proof theory system for $\mathbf{L}$. Such a system would be an algorithm $\Pi_{\mathbf{L}}$ which for a given Δ and Q can be activated and which,

[1] The notion here defined should be more precisely defined as *abduction for consequence*, i.e. we abduce X for the purpose of making Q provable from Δ. There can be other purposes for abduction; for example, we may wish to abduce to make the *interpolation theorem* hold. This kind of abduction is GW abduction. Recall the discussion in Section 10.1.2.

if terminating (it might loop), may give an answer yes ($\Pi_{\mathbf{L}}(\Delta, Q) = 1$) or no ($\Pi_{\mathbf{L}}(\Delta, Q) = 0$).

The answer to (2) may be through a contraction mechanism which tells us what to take out to make Q unprovable.

The answer to (3) can be a semantical interpretation or a model building process.

The answer to (4) may be a revision mechanism which tells us how to revise a theory to accept an input Q.

We Note that the cases (2) and (4) can be quite complex in the case of non-monotonic theories.

The answer to (5) is an abduction mechanism, which is the concern of the following chapters of this book.

The above considerations used the notions 'logic', (the notion of 'logic' includes the notions of 'consistency' and 'consequence') 'theory', and 'formula'. To answer any of the above questions in a general setting, we need to have general settings for these notions.

Our view of logic is that in order to present a logic **L** it is not enough to specify its consequence relation (what proves what in **L**), but a specific proof procedure $\Pi_{\mathbf{L}}$ must also be given. Thus, according to this view, classical logic with the tableaux proof theory is not the same 'logic' as classical logic with resolution proof theory. (See [Gabbay, 1996], Chapter 1.) This view is particularly crucial for abduction. If the logic is used to model some real application, the proof methods of logic must be meaningful for the application (the proof moves are recognisable as meaningful in the semantics of the application). Hence, if the applications demands that $\Delta \vdash Q$ and the logic **L** does not give the result ($\Delta \not\vdash_{\mathbf{L}} Q$) then the abduction procedures carried out through the use of the proof procedures can be expected to be meaningful for the application!

We seek a logical framework which can accommodate many application areas and in which many traditional logics can be uniformly presented. Such a framework is also a good starting point for our formal theory of abduction. The framework is LDS (Labelled Deductive Systems).

Our view of the notion of a theory (or database) is that of a structured family of wffs (data items). The structure of the data can be sets, multisets, sequences, or even a general algebraic structure. In such structured data both a formula and its negation may reside in different locations. The appropriate notion of 'consistency' for structured databases is that of *acceptability*. Some databases are acceptable to us and some are not.[2] A database Δ is not just a family of structured data items.

[2] In standard logics, a theory Δ is acceptable iff it is consistent, i.e. $\Delta \not\vdash \bot$. However, the more general notion of acceptability may be employed for traditional logics as well. We may find some theories unacceptable even though they are consistent, and vice versa. (See Gabbay and Hunter[Gabbay and Hunter, 1993], and Woods [2005b].)

Part of the notion of a database is a recipe of how the database receives input. Since databases are structured, there may be several options concerning where the input can be placed.[3] Our logic $\mathbf{L}$ must have a clear cut notion for each database and each input formula as to where the formula is to be placed. Similarly we require a clear concept of how to take a formula out of the database and what the resultant database is to be. Other required notions are how to join together two databases and how to decompose a database into several databases. The proper set-up for a general database and logic is that the framework of *Labelled Deductive System*. In *LDS* the notion of a formula becomes the notion of *Declarative unit* of the form $t : A$, t is a label from an algebra and A a wff from a logic.

We now describe how we view the nature of the abductive mechanism in general.

The new ideas are:

- *Metalevel properties*:
  abduction is a *metalevel* notion and should be understood in the context of object level/metalevel interaction.

- *Dependence on proof theory*:
  abduction for a logic $\mathbf{L}_1$ depends on proof procedures $\Pi_{\mathbf{L}_1}$ for $\mathbf{L}_1$. Different proof procedures may give different abductive mechanisms.

- *Abduction mechanisms are data*:
  abductive principles can be part of the data. In other words, a declarative item of data can be either a (labelled) formula or a principle of abduction.[4]

- *Abduction can change the logic*:
  abductive principles can be a new principle of proof. In other words new rules can be abduced which can be used to prove more, i.e. the abductive mechanism for $\mathbf{L}_1$ can modify $\Pi_{\mathbf{L}_1}$. The effect of such modification can enrich the logic $\mathbf{L}$.

- *Abduction requires a second background logic of discovery*:
  Assume we have a logic $\mathbf{L}_1$ together with a proof theory $\Pi_{\mathbf{L}_1}$ for it. In order to give an abductive mechanism for $\mathbf{L}_1$ we need to provide two independent components: a procedure for the abduction (which will make use of the proof theory of $\mathbf{L}_1$, and can be thought of as part of the logic of discovery) together with an additional, possibly completely different, logic $\mathbf{L}_2$, (which can be thought of as including a relevance justification and the *plausibility* component). The abductive procedure determines, using the logic $\mathbf{L}_1$ and

[3] In standard logics, where the theories are sets of formulas, input is done by set union.

[4] We have argued elsewhere (Gabbay [1998b; 2001a], Gabbay and Woods [1999; 2000]) that databases should contain the other mechanisms, such as *actions* as part of the data.

its proof theory, possible *candidate* formulas to be abduced. The logic $\mathbf{L}_2$ is used to decide which candidates it is plausible to choose for abduction. The logic $\mathbf{L}_2$ itself may involve its own abduction process, and so on ....

- *Abduction may not succeed:*
  There may not be anything acceptable to abduce, (adding the goal as abducible may not be an option), or if something is added, it may be withdrawn. We need a logic of hypothesis-testing which includes a revision process for when abductions are falsified. (R-revision).

We illustrate the above ideas in a simple example.

Example 12.1 *Consider the database comprising of* $\{a \Rightarrow q, b \Rightarrow q\}$ *and suppose our goal is* $?q$.

Let us examine how the six properties of abduction just mentioned present themselves in this example.

1. Abduction is metalevel
   The very idea of what we want to do is metalevel. We want to add data to make q *provable.*

2. Dependence on proof theory
   We find what options we have to add to the database by following the proof theory and extracting the additional data at the point of failure. Possibilities for our case are $\{a\}, \{b\}, \{a \vee b\}, \{a \wedge b\}, \{q\}$.
   If we adopt a goal directed apporach our options are $\{a\}, \{b\}$.

3. Abduction mechanisms are data
   This is clear since they yield additional data.

4. Abduction requires a background logic
   Given the choice of whether to add $\{a\}$ *or* $\{b\}$, *a background plausibility logic may help us choose.*

5. Abduction may not succeed
   The plausibility logic may tell us not to add anything or the abduction algorithm may yield nothing.

The structure of the chapter is as follows. Section 2 provides an *LDS* formulation for logics with $\Rightarrow$ only. This kind of *LDS* is general enough to cover in a uniform way a variety of implications: classical, intuitionistic, strict (modal), resource/substructural, many-valued and conditional. This system can therefore serve as a general case study for abduction. The later part of the section will give many examples of proofs from resource logics.

Section 3 develops a goal-directed proof theory for the $\Rightarrow$ of Section 2. Although we are presenting a specific *LDS* and its proof theory, the example is general enough to illustrate how abduction can be done in general, for any logic.

Section 4 discusses intuitively various options of how to do abduction.

12.2 Introducing LDS

We have proposed that the best way of describing the abduction mechanism in its general form is to present it within the framework of *Labelled Deductive Systems* (*LDS* [Gabbay, 1996]). This framework is general enough to contain as special cases most, if not all, well-known logics, whether classical, non-classical, monotonic and non-monotonic.

We begin by introducing a typical *LDS* formulation for implication $\Rightarrow$, within which we discuss the principles of abduction.

12.2.1 LDS for $\Rightarrow$

To present an *LDS* for implication $\Rightarrow$ we require an algebra $\mathcal{A}$ of labels of the form $(\mathbb{A}, f, \varphi)$, where $\mathbb{A}$ is a set of labels, f is a binary operation on $\mathbb{A}$ and φ is a binary compatibility relation on $\mathbb{A}$. This is not the most general definition, but it is sufficient for having a general system in which abduction principles are explained. f, φ are necessary predicates in any *LDS*. Different logics may have additional relations and functions on $\mathbb{A}$, besides the compulsory f and φ. The most common additions are an ordering $<$ and the constant d. φ is in the language with $f, <$ and d.

Our notion of a *well-formed declarative unit* is defined as a pair $\alpha : B$, where α is a label, i.e. $\alpha \in \mathbb{A}$, and B is a traditional wff in the propositional language with $\Rightarrow$ and with atomic propositional variables $\{q_1, q_2, q_3, \ldots\}$.

Using the above algebra we can present the (f, φ) $\Rightarrow$-elimination rule as follows:[5]

[5] 'A more general rule is where we have a set of proof *modes* (see Section 11.3.4) and labels which contain information in them which influences the change of mode. Let $\mathbb{M}$ be a set of modes and $\mathbb{R}$ be a set of rules. For each mode $\mu \in \mathbb{M}$ let $\mathbb{R}_\mu$ be the set of rules available to use by the algorithm Π in the mode μ. Thus modus ponens becomes

$$\frac{\mu; \alpha : A; \beta : A \Rightarrow B, \varphi(\mu, \beta, \alpha)}{g(\mu, \beta, \alpha); f(\mu, \beta, \alpha) : B}$$

where μ is the mode before the application of the rule, $g(\mu, \beta, \alpha)$ is the new mode after the application of the rule, $\varphi(\mu, \beta, \alpha)$ is the compatibiltiy predicate and $f(\mu, \beta, \alpha)$ is the new label of B.

- $(f, \varphi) \Rightarrow E$ Rule:

$$\frac{\alpha : A; \beta : A \Rightarrow B; \varphi(\beta, \alpha)}{f(\beta, \alpha) : B}$$

We need to assume that $\varphi(\beta, \alpha)$ is a decidable predicate over $\mathbb{A}$ and that f is a total recursive function on A. This makes the rule effective. This rule is a general labelled *Modus Ponens* rule. We shall see later in Definition 12.2 how to use it.

We need one more notion. Let $t(x)$ be a term of the algebra built up using the function symbol f and the variable x ranging over $\mathbb{A}$. The function $\lambda x t(x)$ is a unary function over $\mathbb{A}$ which may or may not be definable in A.

We say that $\lambda x t(x)$ is *φ-realisable* by an element $\alpha \in \mathbb{A}$ iff the following holds

$$\forall x[\varphi(\alpha, x) \Rightarrow f(\alpha, x) = t(x)].$$

We denote this α, if it exists, by $(\eta x)t(x)$.

We must assume that our algebra is such that it is decidable to check whether $(\eta x)t(x)$ exists and if it does exist we have algorithms available to effectively produce it.

Given an $(f, \varphi) \Rightarrow$-elimination rule, we can define the $(f, \varphi) \Rightarrow$-introduction rule as follows:

- To show $\gamma : A \Rightarrow B$, assume $x : A$ and $\varphi(\gamma, x)$ and prove $f(\gamma, x) : B$,[6] where x is a new atomic variable.

The traditional way of writing this rule is as follows:

Show $\gamma : A \Rightarrow B$ from subcomputation:

	show $f(\gamma, x) : B$
Assume	$x : A; \varphi(\gamma, x)$
$\vdots$	
$f(\gamma, x) : B$	

Exit with $\gamma : A \Rightarrow B$

Definition 12.2

1. *A database Δ is a set of declarative sentences (labelled formulas) together with a set of compatibility relations of the form $\varphi(\alpha_i, \beta_i)$ on the labels of Δ. Part of the notion of a database is the notion of where to place additional inputs. Thus, if we want to add the declarative unit $x : B$ to the database Δ*

[6] In practice one proves $t(x) : B$ for some term $t(x)$, and then shows that $\gamma = (\eta x)t(x)$.

to get a new database Δ', we need an insertion operation. We can assume that with the database there is a metapredicate $\Psi_2(\Delta, \Delta', x : B)$ saying Δ' is the result of inserting $x : B$ into Δ. We also need a notion of deleting $x : B$ from Δ. We can use the metapredicate $\Psi_3(\Delta, \Delta', x : B)$. Of course Ψ_3 and Ψ_2 satisfy some obvious properties. These predicates need not be definable in the algebra $\mathcal{A}$, but could be higher order. We can write $\Delta' = \Delta + (x : B)$ for insertion and $\Delta' = \Delta - (x : B)$ for deletion, provided we understand $+$ and $-$ as referring to Ψ_2 and Ψ_3. Ψ_2 and Ψ_3 must be computable. In practice we can express Ψ_3 and Ψ_2 in the algebra.[7]

2. *We define the notion $\Delta \vdash \alpha : A$ by induction on the number of nested uses of $\Rightarrow$ Introduction.*

   (a) *$\Delta \vdash_0 \alpha : A$ iff there exists a sequence of labelled wffs $\alpha_1 : A_1, \ldots, \alpha_k : A_k$ such that each $\alpha_i : A_i$ is either an item of data in Δ or is obtained from previous two elements*

   $$\alpha_m : A_m, \alpha_n : A_n, m, n < i$$

   via the $(f, \varphi) \Rightarrow$-elimination rule. This means that $\varphi(\alpha_n, \alpha_m)$ can be proved from Δ, $A_n = A_m \Rightarrow A_i$ and $\alpha_i = f(\alpha_n, \alpha_m)$. We also have $\alpha_k : A_k$ is $\alpha : A$.

   (b) *We say $\Delta \vdash_{n+1} \alpha : B \Rightarrow C$ iff $\Delta + \{x : B; \varphi(\alpha, x)\} \vdash_m f(\alpha, x) : C$, where x is a new variable and $m \leq n$.*

3. *We write $\Delta \vdash_\infty \alpha : A$ if for some n, $\Delta \vdash_n \alpha : A$.*

4. *Let Ψ be a decidable meta-predicate on pairs of the form (S, α) where S is a data structure of labels and α is a label.*[8] *Given a database Δ let $S_\Delta = \{\alpha \in \mathbb{A}$ s.t. for some formula $B \mid \alpha : B \in \Delta\}$.*

   We write $\Delta \vdash_\Psi B$ if for some α, $\Delta \vdash_\infty \alpha : B$ and $\Psi(S_\Delta, \alpha)$ holds.[9]

5. *We say B is a theorem of the logic if $\varnothing \vdash_\Psi B$.*

Example 12.3 *Here is a simple example. Let $\mathcal{A}$ be a free non-commutative non-associative semigroup with operation $*$, i.e. it has the form $\mathcal{A} = (\mathbb{A}, *)$. $*$ will be our f. Consider the database Δ with*

[7] We let $\Delta' = \Delta \cup \{x : B\}$, but by definition we also add some formulas about x, relating x to the labels y of Δ. For example, the labels in Δ may be linearly ordered by $<$. We may agree that all inputs $x : B$ receive the highest priority in the ordering, in which case the formula to add is $\forall y(y \leq x)$. We write $\Psi_2(\Delta, \Delta', x : B)$ to say Δ' is the result of proper input of $x : B$ into Δ. See Section 13.1 below.

[8] Although Ψ is a predicate on the algebra it need not be first order. It need only be recursive. S is usually a sequence of labels.

[9] Note that the logic $\vdash_\Psi$ can turn out to be either monotonic or non-monotonic. It depends on Ψ and the properties of the input function.

1. $a_1 : A \Rightarrow (A \Rightarrow B)$
2. $a_2 : A$

We can prove B as follows

3. $(a_1 * a_2) : A \Rightarrow B$, *from (1) and (2)*
4. $(a_1 * a_2) * a_2 : B$, *from (3) and (2)*
5. *Thus we have* $\Delta \vdash (a_1 * a_2) * a_2 : B$.

   We have several options for Ψ.

   (a) $\Psi(S_\Delta, \alpha)$ *holds if all atoms in* Δ *appear in* α. *(This is* relevance logic *for* $\Rightarrow$. *All assumptions are used.)*
   (b) $\Psi(S_\Delta, \alpha)$ *holds if all atoms in* Δ *appear exactly once, (This is* linear logic *for* $\Rightarrow$. *All assumptions in* Δ *are used exactly once).*
   (c) $\Psi(S_\Delta, \alpha) = \top$ *(This is* intuitionistic logic. *The labels are ignored.).*
   (d) *Assuming* $*$ *is associative, we can have* $\Psi(S_\Delta, \alpha)$ *holds iff* $x_1 * x_2 * \ldots * x_n = \alpha$, *where* $S_\Delta = (x_1, \ldots, x_n)$. *This yields a Lambek calculus implication.*

12.2.2 Examples of Resource LDS

We now illustrate the use of the labels in performing computations in substructural logics. This will show how different labelling algebras can give different implicational logics, and will justify the choice of the *LDS* for the development of general abduction principles. Our labels are multisets as defined in Definition 12.4 below. Multisets are like sets except that elements may appear in them more than once. The use of multisets allows us to track the frequency with which an assumption is used in *modus ponens*.

Definition 12.4 (Multisets) *Let $\mathbb{A}$ be a set of atoms. A multiset based on $\mathbb{A}$ is a function α on $\mathbb{A}$ giving for each element $a \in \mathbb{A}$ a natural number $\alpha(a) \geq 0$. $\alpha(a)$ tells us how many copies of a we have in the multiset α. Let $\gamma = \alpha \cup \beta$ be defined as the function $\gamma = \alpha + \beta$ (i.e. for each a, $\gamma(a) = \alpha(a) + \beta(a)$). Let $\gamma = \alpha \dot{-} \beta$ be the function defined for each a by $\gamma(a) = 0$, if $\alpha(a) \leq \beta(a)$, and $\gamma(a) = \beta(a) - \alpha(a)$, otherwise.*

Example 12.5

1. *The simplest example is the set resource labelling. The labels are sets $\mathbb{A}$ of atomic labels and $f(\beta, \alpha) = \beta \cup \alpha$. φ and Ψ depend on the logic under consideration. In many logics, however, we have it that $\varphi(\beta, \alpha) = \top$, i.e. that φ always holds.*

2. *Consider the following database* Δ_1*:*

$$\begin{array}{ll} \{a_1\}: & A \\ \{a_2\}: & A \\ \{a_3\}: & A \Rightarrow B \end{array}$$

The database can prove

$$\begin{array}{ll} \Delta_1 \vdash_1 \{a_3, a_2\}: & B \\ \Delta_1 \vdash_1 \{a_3, a_1\}: & B \end{array}$$

Let $\Psi(S, \alpha)$ *be* $\alpha = \bigcup_{\beta} \beta \in S$. *Then we do* not *have that* $\Delta_1 \vdash_\Psi B$.

3. *Consider the database* Δ_2

$$\begin{array}{ll} \{a_1\} & (A \Rightarrow B) \Rightarrow (A \Rightarrow B) \\ \{a_2\} & A \Rightarrow B \\ \{a_3\} & A. \end{array}$$

We have (using the last two items)

$$\Delta_2 \vdash \{a_2, a_3\} : B$$

We can also first use $\{a_1\}$ *and* $\{a_2\}$ *and derive*

$$\Delta_2 \vdash \{a_1, a_2, a_3\} : B.$$

So we do have in this case

$$\Delta_2 \vdash_\Psi B.$$

Note the following three conventions:

1. Each new assumption is labelled by a new atomic label. An ordering on the labels can be imposed, namely $a_1 < a_2 < a_3$. This is to reflect the fact that the assumptions arose from our attempt to prove $A \Rightarrow (A \Rightarrow ((A \Rightarrow B) \Rightarrow B))$ and not for example from $(A \Rightarrow B) \Rightarrow (A \Rightarrow (A \Rightarrow B))$ in which case the ordering would be $a_3 < a_1 < a_2$. The ordering can affect the proofs in certain logics. Some logics allow us to bring in, anywhere in the proof, theorems of the logic with the empty label and allow their use in the proof. Other logics do not allow this.

2. If in the proof A is labelled by the multiset α and $A \Rightarrow B$ is labelled by β then B can be derived with a label $\alpha \cup \beta$ where $\cup$ denotes multiset union.

3. If B was derived using A, as evidenced by the fact that the label α of A is a singleton $\{a\}$, a atomic and a is in the label β of B ($\alpha \subseteq \beta$) then we can derive $A \Rightarrow B$ with the label $\beta \dot{-} \alpha$ ('$\dot{-}$' is multiset subtraction).

In case our labels are sets, we use $\beta - \alpha$, where '$-$' is set subtraction.

Example 12.6 *Show* $(B \Rightarrow A) \Rightarrow ((A \Rightarrow B) \Rightarrow (A \Rightarrow B))$

$\underline{\{a_2, a_1\} : (A \Rightarrow B) \Rightarrow (A \Rightarrow B)}$

(1) $\{a_1\} : B \Rightarrow A$
(2) $\{a_2, a_1\} : (A \Rightarrow B) \Rightarrow (A \Rightarrow B)$

> $\underline{\{a_2, a_1, a_2\} : A \Rightarrow B}$
>
> *(2.1)* $\{a_2\} : A \Rightarrow B$
> *(2.2)* $\{a_2, a_1, a_2\} : A \Rightarrow B$
>
> > $\underline{\{a_2, a_1, a_2, a_3\} : B}$
> >
> > *(2.2.1)* $\{a_3\} : A$
> > *(2.2.2)* $\{a_2, a_3\} : B$
> > *(2.2.3)* $\{a_1, a_2, a_3\} : A$
> > *(2.2.4)* $\{a_2, a_1, a_2, a_3\} : B$

The above is the box method of representing the deduction. Note that in leaving the inner box for $\{a_2, a_1, a_2\} : A \Rightarrow B$, *multiset subtraction was used and only one copy of the label* a_2 *was taken out. The other copy of* a_2 *remains and cannot be cancelled, so that the entire computation finishes with the label of* $\{a_2\}$. *We therefore have scope to define different logics by saying when a labelled proof is acceptable. For linear logic, the final label at the end of the computation must be empty, signifying that formulae have only been used once. Hence this formula is not a theorem of linear logic because the outer box does not exit with label* $\varnothing$. *In relevance logic, the discipline uses sets and not multisets. Thus the label upon leaving the inner box in this case would be* $\{a_1\}$ *and that upon leaving the outermost box would be* $\varnothing$.

Note that different conditions on labels correspond to different logics, given informally in the following table:

condition Ψ:	**logic:**
ignore the labels	intuitionistic logic
accept only the derivations which use all the assumptions	relevance logic
accept derivations which use all assumptions exactly once	linear logic

The conditions on the labels can be translated into reasoning rules.

The following are examples of further proofs. The examples speak for themselves.

Example 12.7

(1) $\varnothing : (A \Rightarrow (A \Rightarrow B)) \Rightarrow (A \Rightarrow B)$ *subcomputation*

> $\underline{\{a_1\} : (A \Rightarrow B)}$
>
> *(1.1)* $\{a_1\} : A \Rightarrow (A \Rightarrow B)$ *assumption*
>
> *(1.2)* $\{a_1\} : A \Rightarrow B$ *subcomputation*
>
> > $\underline{\{a_1, a_2\} : B}$
> >
> > *(1.2.1)* $\{a_2\} : A$ *assumption*
> >
> > *(1.2.2)* $\{a_1, a_2\} : A \Rightarrow B$ *by (1.1), (1.2.1)*
> >
> > *(1.2.3)* $\{a_1, a_2\} : B$ *by (1.2.1), (1.2.2)*

The proof is acceptable in relevance logic, but not in linear logic, because if labels are multisets we obtain (1.2.3) $\{a_1, a_2, a_2\} : B$ *in the innermost box, which provides only* (1.2) $\{a_1, a_2\} : A \Rightarrow B$ *in the next box, and then* (1) $\{a_2\} : (A \Rightarrow (A \Rightarrow B)) \Rightarrow (A \Rightarrow B)$ *in the outermost box. This happens because the assumption (1.2.1) is used twice in this proof.*

It is clear that for linear logic the labelling becomes more efficient if we introduce the condition $\varphi(\beta, \alpha) = \beta \cap \alpha = \varnothing$ (instead of $\varphi(\beta, \alpha) = \top$). This new condition prevents assumptions from being used more than once. Ψ then makes sure that all assumptions are used.

Example 12.8

(1) $\varnothing : (C \Rightarrow A) \Rightarrow ((B \Rightarrow C) \Rightarrow (B \Rightarrow A))$ *subcomputation*

$\underline{\{a_1\} : (B \Rightarrow C) \Rightarrow (B \Rightarrow A)}$

(1.1) $\{a_1\} : C \Rightarrow A$ *assumption*

(1.2) $\{a_1\} : (B \Rightarrow C) \Rightarrow (B \Rightarrow A)$ *subcomputation*

$\underline{\{a_1, a_2\} : B \Rightarrow A}$

(1.2.1) $\{a_2\} : B \Rightarrow C$ *assumption*

(1.2.2) $\{a_1, a_2\} : B \Rightarrow A$ *subcomputation*

$\underline{\{a_1, a_2, a_3\} : A}$

(1.2.2.1) $\{a_3\} : B$ *assumption*

(1.2.2.2) $\{a_2, a_3\} : C$ *by (1.2.2.1), (1.2.1)*

(1.2.2.3) $\{a_1, a_2, a_3\} : A$ *by (1.2.2.2), (1.1)*

The proof is acceptable in linear logic. In fact, the proof is acceptable in Lambek logic, where a database is a sequence of formulas, input into the database is done by adding the input to the end of the sequence and modus ponens is done by order adjacent formulas, where the implication comes first, i.e. '$X \Rightarrow Y, X$' yields 'Y' and the result of the modus ponens replaces the participants in the position of the ordering.

Example 12.9

(1) $\varnothing : (A \Rightarrow (B \Rightarrow C)) \Rightarrow ((A \Rightarrow B) \Rightarrow (A \Rightarrow C))$ *subcomputation*

> $\underline{\{a_1\} : (A \Rightarrow B) \Rightarrow (A \Rightarrow C)}$
>
> *(1.1)* $\{a_1\} : A \Rightarrow (B \Rightarrow C)$ *assumption*
> *(1.2)* $\{a_1\} : (A \Rightarrow B) \Rightarrow (A \Rightarrow C)$ *subcomputation*
>
> > $\underline{\{a_1, a_2\} : A \Rightarrow C}$
> >
> > *(1.2.1)* $\{a_2\} : A \Rightarrow B$ *assumption*
> > *(1.2.2)* $\{a_1, a_2\} : A \Rightarrow C$ *subcomputation*
> >
> > > $\underline{\{a_1, a_2, a_3\} : C}$
> > >
> > > *(1.2.2.1)* $\{a_3\} : A$ *assumption*
> > > *(1.2.2.2)* $\{a_1, a_3\} : B \Rightarrow C$ *by (1.1), (1.2.2.1)*
> > > *(1.2.2.3)* $\{a_2, a_3\} : B$ *by (1.2.1), (1.2.2.1)*
> > > *(1.2.2.4)* $\{a_1, a_2, a_3\} : C$ *by (1.2.2.2), (1.2.2.3)*

The proof is acceptable in relevance logic. It is not acceptable in linear logic, because the assumption (1.2.2.1) is used twice (cf. the solution to Example 12.7).

Example 12.10

(1) $A \Rightarrow (B \Rightarrow A)$ *subcomputation*

> $\underline{B \Rightarrow A}$
>
> *(1.1)* A *assumption*
> *(1.2)* $B \Rightarrow A$ *subcomputation*
>
> > $\underline{A}$
> >
> > *(1.2.1)* B *assumption*
> > *(1.2.2)* A *by (1.1)*

The proof is acceptable in intuitionistic logic. It cannot be transferred to relevance logic by introducing labels because the label for B in (1.2.1) does not disappear in the outermost box.

Example 12.11

> *(1)* $\varnothing : ((A \Rightarrow A) \Rightarrow A) \Rightarrow A$ *subcomputation*
>
> > $\underline{\{a_1\} : A}$
> >
> > *(1.1)* $\{a_1\} : (A \Rightarrow A) \Rightarrow A$ *assumption*
> >
> > *(1.2)* $\varnothing : A \Rightarrow A$ *subcomputation*
> >
> > > $\underline{\{a_2\} : A}$
> > >
> > > *(1.2.1)* $\{a_2\} : A$ *assumption*
> > >
> > > *(1.2.2)* $\{a_2\} : A$ *by (1.2.1)*
> >
> > *(1.3)* $\{a_1\} : A$ *by (1.1), (1.2)*

This proof is not *in relevance logic, because (1.2) is proved as a 'lemma' with label $\varnothing$. Relevance logic does not allow for such labelling: you should not forget assumptions while making a proof, so $A \Rightarrow A$ should be proved with a label $\{a_1\}$, which is impossible.*

Of course, by deleting all labels we get an intuitionistic proof.

12.3 Goal Directed Algorithm for ⇒

One of our general abduction principles for general logics is that abductive procedures do not depend only on the consequence relation (i.e. on $\Delta \vdash A$ alone) but also on the proof procedure given for finding whether $\Delta \vdash A$ or not. In other words we abduce something in the context of our algorithmic search. To make this concrete, we offer, in this section a general goal directed proof procedure for our ⇒ *LDS*. We will then show how abduction can hinge upon it.

In fact, we will show how abduction can be dependent on any general type proof theoretic procedure, since our goal directed algorithm is a typical reduction algorithm. It makes good sense to illustrate abduction using a goal directed system because the very idea of abduction is goal directed. We have $\Delta \nvdash A$ and we are looking for ways to prove A. A is our goal!

12.3.1 The Algorithm

We now give a goal directed computation which seeks to find for a given Δ and B whether $\Delta \vdash_\Psi B$, namely, whether there exist any labels β such that $\Delta \vdash_\infty \beta : B$ and $\Psi(S_\Delta, \beta)$ hold.

We begin by observing that each declarative formula in the database has the form $\alpha : A_1 \Rightarrow (A_2(\Rightarrow \cdots \Rightarrow (A_n \Rightarrow q)\ldots))$ where q is atomic. q is called the *head* and $(A_1, \ldots, A_n)$ is called the *body* (sequence). Consider the question of whether $\Delta \vdash_k \delta : A$. A can have one of two forms:

1. A can be an atom q

2. A can have the form $A = (B \Rightarrow C)$.

Also recall that

3. $\Delta \vdash_\Psi A$ if for some $\delta, k, \Delta \vdash_k \delta : A$ and $\Psi(S_\Delta, \delta)$ holds.

In the first case, if $\Delta \vdash_k \delta : q$, then there must exist an $\alpha : A_1 \Rightarrow (A_2 \Rightarrow \cdots (A_n \Rightarrow q)\ldots)$ in Δ and labels $\alpha_1 \ldots, \alpha_n$ such that $\Delta \vdash_k \alpha_i : A_i$ and labels $\delta_1, \ldots, \delta_n = \delta$ such that the following all hold:

$$\begin{aligned}
\delta_1 &= f(\alpha, \alpha_1); \varphi(\alpha, \alpha_1) \\
\delta_2 &= f(\delta_1, \alpha_2); \varphi(\delta_1, \alpha_2) \\
&\vdots \\
\delta = \delta_n &= f(\delta_{n-1}, \alpha_n); \varphi(\delta_{n-1}, \alpha_n)
\end{aligned}$$

Let us abbreviate the above as

$$\delta = f[\alpha, (\delta_1, \ldots, \delta_n)]$$

and

$$\varphi[\alpha, (\delta_1, \ldots, \delta_n)].$$

Of course it may be the case that there is no item of data in the database with head q. In this case we are stuck and the computation clearly cannot continue. We say that we have *immediate failure* at this point.

Another possibility of *immediate* failure is that although a clause does exist, φ cannot be satisfied. Again we cannot go on.

In the second case, if $\Delta \vdash_k \delta : B \Rightarrow C$ holds, then if we add $x : B$ to Δ for a new variable x, we must have that

$$\Delta + \{x : B\} \vdash_m f(\delta, x) : C$$

for some $m \leq k - 1$.

Given the above, we can write a constraints logic programming program (the algebra of constraints is our algebra $\mathcal{A} = (\mathbb{A}, f, \varphi)$) for the metapredicate, *Success* $(\Delta, \delta : A, \mathit{constraints}, \theta) = 0$ or 1.

where Δ is a labelled database (with labels containing variables), $\delta : A$ is the current goal, *constraints* is a set of constraints on the labels and θ is a possibly partial substitution for the label variables.

Success = 1 means the computation succeeds, and
Success = 0 means the computation finitely fails.

Of course the computation can loop or just continue forever for whatever reasons. So the two options are not exhaustive. The consequence relation meaning of *Success* = 1 means that $\Delta\theta \vdash \delta\theta : A$ and $\mathcal{A} \vdash \mathit{constraints}\ \theta$.

We now give a recursive definition. The definition is formulated using metapredicates $\Psi_1, \Psi_2, \Psi_3, \Psi_4$ and Ψ. This way we can change the computation for different logics by varying the Ψs. (See [Gabbay and Olivetti, 2000].) In general the Ψs need to satisfy some conditions discussed in Section 5. The meaning of the Ψs is as follows. Ψ_1 tells us when an atomic query can succeed or fail from a database in *one step* (*immediate success* or *failure*). Ψ_2 is an insertion predicate. Ψ_3 is a deletion predicate and Ψ_4 is a decomposition predicate; given a database Δ, and databases $\Delta'_1, \ldots, \Delta'_n$, it may be that Δ'_i are a decomposition of Δ in some way. We write $\Psi_4(\Delta, \Delta'_1, \ldots, \Delta'_n)$ to express this relation. Note that n may vary and that the decomposition may not be disjoint. Ψ tells us when a proof is acceptable.

The choice of Ψs proposed below is for the simple resource logics case.

Definition 12.12 *We define the metapredicate* (Immediate) Success $(\Delta, \delta : A$, constraint, $\theta)$ = *1 or 0, where θ is a substitution to label variables, as follows:*

1. Immediate Success Case: (Immediate) Success$(\Delta, \delta : q$, constraints, $\theta) = 1$ *if q is atomic and $\delta\theta = \alpha\theta$, for some label α such that $\Psi_1(\Delta\theta, \delta\theta : q)$ holds and the constraints are provable in the algebra $\mathcal{A}$.*

   In the case of resource logics, $\Psi_1(\Delta, \delta : q)$ can be, for example, $\alpha : q \in \Delta$.

2. Implication case: Success$(\Delta, \delta : B \Rightarrow C$, constraints, $\theta) = x$ *if for $\Delta' = \Delta + (a : B)$ (i.e. Δ' such that $\Psi_2(\Delta, \Delta', a : B)$) we have* Success$(\Delta', f(\delta, a) : C$, constraints, $\theta) = x$, *for a new atomic constant a. Recall that Ψ_2 says that Δ' is the (usually unique) result of inserting $a : B$ into Δ.*

   In the case of resource logics, $\Psi_2(\Delta, \Delta', a : B)$ is $\Delta' = \Delta \cup \{a : B\}$.

3. Immediate Failure case: (Immediate) Success$(\Delta, \delta : q$, constraints, $\theta) = 0$ *for q atomic if either the constraints are not provable under the substitution θ or there is no clause in the database Δ with head q, such that $\Psi_1(\Delta\theta, \delta\theta : q)$ holds.*

4. Cut Reduction case: Success$(\Delta, \delta : q$, constraints, $\theta) = 1$ *(resp.* $= 0$*) if for some (resp. any)* $E = (\alpha : A_1 \Rightarrow (A_2 \Rightarrow \ldots (A_n \Rightarrow q)\ldots))$ *in* Δ *and for some (resp. any) new variables* $\delta_1, \ldots, \delta_n$ *and some (resp. all) choices of* $\Delta'_1, \ldots, \Delta'_n$ *and* $\Delta'_0 = \{\alpha : A_1 \Rightarrow (A_2 \Rightarrow \ldots (A_n \Rightarrow q)\ldots)\}$ *such that* $\Psi_4(\Delta, \Delta'_0, \ldots, \Delta'_n)$ *holds and substitution* θ' *extending* θ *to* $\delta_1, \ldots, \delta_n$ *we have for each* $1 \leq i \leq n$ *(resp. for some i)* Success$(\Delta'_i, \delta_i : A_i$, constraints$', \theta') = 1$ *(resp. =0) where* constraints$'$ = constraints$\cup\{\delta\theta' = f[\alpha\theta', (\delta_1\theta', \ldots, \delta_n\theta')]\} \cup \{\varphi[\alpha\theta', (\delta_1\theta', \ldots, \delta_n\theta')]\}$.[10]

5. Consequence: *We have* $\Delta \vdash A$ *if for some variable* δ *and some substitution* θ *for* δ *we have* Success$(\Delta, \delta : A, \{\Psi(S_\Delta, \delta)\}, \theta) = 1$.

Note that the computation may not be decidable unless we assume we have recursive algorithms for all the conditions and metapredicates mentioned in it and all ranges of choice are computably finite.

The computation may loop, so a historical loop checker may be needed.

The above algorithm is a typical reduction algorithm. We start with an initial state involving $(\Delta, A,$ *parameters*$)$ and rewrite or reduce it to other states. Some states (Γ, B, P) are reduced to *success* and some states may be such that no reduction rules apply, in which case we can say they are reduced to (*immediate*) *failure*.

The typical abduction principle will consider the history of the reduction up to a point of failure and tell us what to do at that stage.

Of course the process may loop but in the propositional case a historical loop checker can take care of that and so we may assume that the parameters include a device that eliminates looping.

12.3.2 Examples

Example 12.13 *In any logic, consider* $q \Rightarrow q \vdash ?q$.

Success($q \Rightarrow q, q$, no constraints*)* $= x$, *simply loops.*

Example 12.14 *Let us do Example 12.7 in a goal-directed way, for linear logic. The labels are multisets. Note the way the computation is displayed, as*
Data $\vdash ?$ **Goal; constraints.**

1. $\varnothing \vdash ?\delta : (A \Rightarrow (A \Rightarrow B)) \Rightarrow (A \Rightarrow B); \delta = \varnothing$

2. $\{a_1\} : A \Rightarrow (A \Rightarrow B) \vdash ?\delta \cup \{a_1\} : A \Rightarrow B; \delta = \varnothing$

3. $\{a_1\} : A \Rightarrow (A \Rightarrow B), \{a_2\} : A \vdash ?\delta \cup \{a_1, a_2\} : B; \delta = \varnothing$

[10] Δ_0 appears in Ψ_4 as a parameter containing the formula we are using in this rule.

4. *We split here into two parallel computations, listed below as (5a) and (5b). Using the clause* $\{a_1\} : A \Rightarrow (A \Rightarrow B)$.

5. *(a)* $\{a_1\} : A \Rightarrow (A \Rightarrow B), \{a_2\} : A \vdash ?x_1 : A$

   (b) $\{a_1\} : A \Rightarrow (A \Rightarrow B), \{a_2\} : A \vdash ?x_2 : A$

The constraints for both 5a and 5b (done in parallel) are

$$\{a_1\} \cup \{x_1\} \cup \{x_2\} = \delta \cup \{a_1, a_2\} \textit{ and } \delta = \varnothing$$

5a can succeed by substituting $\theta(x_1) = a_2$ *and similarly 5b can succeed by substituting* $\theta(x_2) = a_2$. *Thus we get the constraints to be* $\{a_1, a_2, a_2\} = \{a_1, a_2\}$.

Since for linear logic we are dealing with multisets, this computation fails, because the constraints cannot be satisfied and there are no other options for the computation. In the case of relevance logic, the labels are sets and the constraints are satisfied and the computation succeeds.

The above computation can be made more efficient. Our first optimisation move is to throw out of the database any clause 'used' by rule 3. This saves us time because we are allowed to use each clause at most once.

So condition 4 of the definition of the *Success* predicate can be modified by changing Ψ_4 and requiring it to say that $\Delta'_i \subseteq \Delta \dot{-} \{\alpha : (A_1 \Rightarrow \ldots \Rightarrow (A_n = q) \ldots)$ for all i.

Since also all the clauses in the database must be used, we can further modify Ψ_4 to be the conjunction of the two following conditions:

- $\bigcup_{i=1}^{n} \Delta'_i = \Delta \dot{-} \{\alpha : (A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q) \ldots)\}$
- $\Delta'_i \cap \Delta'_j = \varnothing$ for $i \neq j$.

We will need to change clause 1 of the algorithm to be

- *Success*$(\Delta, \delta : q,$ *constraints*$, \theta) = 1$ if $\{\alpha : q\} = \Delta$ and $\alpha\theta = \delta\theta$, i.e. change Ψ_1 to the condition $\{\alpha : q\} = \Delta$.

The computation of our example will now be short and quick as follows (we don't even need to use labels):

$$\varnothing \vdash ?((A \Rightarrow (A \Rightarrow B)) \Rightarrow (A \Rightarrow B)$$

if

$$A \Rightarrow (A \Rightarrow B) \vdash ?A \Rightarrow B$$

if

$$A; A \Rightarrow (A \Rightarrow B) \vdash ?B$$

if

$$A \vdash ?A \text{ and } \varnothing ?A$$

if

success and *failure*.

We take this opportunity to indicate options for abduction. Our purpose is to succeed. The simplest option in the case of linear logic is to *duplicate* A (as we are missing one copy of A).

This would normally be done at the stage where $\{a_2\} : A$ was put into the database. At that stage $A \Rightarrow (A \Rightarrow B)$ was already in the database, so we need to add to the database the wff $(A \Rightarrow (A \Rightarrow B)) \Rightarrow A$, (rather than adding just A).

So in order for $(A \Rightarrow (A \Rightarrow B)) \Rightarrow (A \Rightarrow B)$ to succeed from the empty database, we need to abduce $(A \Rightarrow (A \Rightarrow B)) \Rightarrow A$ into the database. This will not help in our case, where the logic is linear logic, because to get A we need another copy of $A \Rightarrow (A \Rightarrow B)$. So we are better off putting A in the database. This can be more readily seen in the case of the more optimised computation where failure results from $\varnothing \vdash ?A$.

Another option for abduction is to regard the constraints as satisfied if they hold for the labels when regarded as sets. This means changing the proof procedures to those of relevance logic. [Gabbay and Olivetti, 2000] shows how to formulate many logics in a goal directed way.

12.4 Intuitive Theory of Labelled Abduction

This section will introduce our intuitive theory of abduction within the framework of Labelled Deductive Systems, and give some simple examples.

The basic situation we are dealing with can be presented as

$$\begin{array}{ll} \Delta & \vdash ?!Q \\ \text{data} & ?\text{query or } ! \text{ input} \end{array}$$

It is a relationship between a database and a (labelled) formula. The relationship is either declarative (i.e. $?Q$, Q a query) or imperative ($!Q$, Q is an input or a demand to perform abduction or a demand for explanation etc). In the imperative case there is an interaction between Δ and Q and a new database Δ' emerges.

We have put forward in previous sections that the most general and useful database is the one where the data is labelled and the proof procedures use the labelling.

In this set up, the abduction rules are extra moves that help answer the query or help change the database as a result of the query or input. We include as part of our concept of abduction the task of finding out what to delete from the database to make Q not provable. We write $Abduce^{+}(\Delta, Q)$ for the task of making $\Delta \vdash Q$ and $Abduce^{-}(\Delta, Q)$ for making $\Delta \not\vdash Q$.

So to do abduction we need more precise proof procedures or update procedures for labelled databases and then on top of that we can define the extra abductive rules.

The exact proof procedures can be conveniently formalised in the framework of *LDS* and the previous section gave one example procedure, namely the goal directed proof procedure. We shall give a sample abduction algorithm for the goal directed procedure in a later section. The reason why abduction should depend also on the proof procedures can be intuitively motivated by an example. Suppose I am looking for an object in my home which I need for some purpose. I search the house and keep track of where I looked of what I found along the way. If I fail to find the object, I look for a substitute. This too can be seen as a kind of abduction. I may use something else I found during the search or I may add a new principle of supplementary searching in the hope of finding either the object or a new possible substitute. If I find several possible candidates for a substitute I may use other unrelated reasoning to decide which one to take.

We illustrate these ideas in a series of examples.

12.4.1 Abduction in Knowledge Bases

Example 12.15 The database below is a Horn clause database. It is labelled in the sense that each clause is named. The query is D. The query does not follow from the database as it is. We are going to use it to illustrate principles of abduction.

	Data	**Query**
a_1:	$I \Rightarrow (T \Rightarrow D)$	$?\,D$
a_2:	$L \Rightarrow I$	
a_3:	$L \Rightarrow (S \Rightarrow T)$	
a_4:	$O \Rightarrow (P \Rightarrow T)$	
a_5:	L	

The database literals have no meaning. Let us give them a meaning. In the Stanford University English Department, there are two main ways of getting a PhD degree. One can either put forward a thesis, stay in the department for 4-5 years acquiring and displaying an immense breadth of knowledge and pass an interview, or one can write a very good publication and get a job offer from another university in the top ten in the country. The database then becomes:

	Data
a_1:	Interview $\Rightarrow$ (Thesis $\Rightarrow$ Degree)
a_2:	Lecture $\Rightarrow$ Interview
a_3:	Lecture $\Rightarrow$ (Scholarly Survey $\Rightarrow$ Thesis)
a_4:	(Job) Offer $\Rightarrow$ (Publications $\Rightarrow$ Thesis)
a_5:	Lecture

Query

? Degree

Another interpretation for the same database is a component interpretation. To do the laundry (D) one needs a washing machine (T) and washing powder (I). For washing powder one can use dishwashing soap (L). For a washing machine one may use a dishwasher (S) and dishwashing soap or one may handwash (P) but then one at least needs a spinner (O).

This gives the following:

Data

a_1: Washing Powder ⇒ (Washing Machine ⇒ Laundry)

a_2: Dishwashing Soap ⇒ Washing Powder

a_3: Dishwashing Soap ⇒ (Dishwasher ⇒ Washing Machine)

a_4: Spinner ⇒ (Handwash ⇒ Washing Machine)

a_5: Dishwashing Soap.

Query

? Laundry

We now list several possible abductive principles for the query $?D$. The principles depend on the computation, so let us suppose that we compute the query prolog like, where the pointer always starts at the top clause (assume $a_1 > a_2 > a_3 > a_4 > a_5$.)

We note that in logic programming [Eshghi and Kowalski, 1990] abduction for Horn clause programs is done via a system of the form (Δ, I, A), where Δ is the program, I is a set of integrity constraints and A is a set of literals which are *abducible*. Whenever an abducible literal is encountered in the computation (e.g. $?D$) it is immediately added to the database provided it does not violate the integrity constraints.

Let us now examine our options:

Possible Principles of Abduction

1. The first option is to abduce on anything as soon as needed. This corresponds in our case, to no integrity constraints and every literal is abducible. In this case we add D to the database, i.e. the Abduction principle yields D. In the component example such abduction makes no sense. I want to know which parts are missing so that we can get them and wash our clothes.

2. The second option is to abduce on literals which are not heads of clauses. In this case, we add S. This is because S is the first literal encountered in the top down order of execution. Note that we do not use here a set of abducibles. The structure of the database determines what we add.

3. If our underlying logic is not classical logic but some other resource logic, we will not succeed by adding S to the database because that would require the 'use' of L twice: once to make I succeed in clause a_2 and once to make T succeed in clause

   a_3. In the component example we need more dishwashing soap if we use a dishwasher, and we have only one lot of it (i.e. a_5).

   Note that the database is structured and thus we can add
   $a_6: \quad L$
   and $\{a_1, \ldots, a_5\}$ is *not* the same database as $\{a_1, \ldots, a_6\}$.

   Anyway, if the underlying logic is a resource logic, the result of our abduction will be O and P, unless we are prepared to add another copy of L.

4. If we require the weakest logical assumption (in classical logic) which makes the goal succeed then we must add $S \vee (O \wedge P)$. This abduction principle is independent of the computation.

5. In co-operative answering, the abduction principle takes the top level clause. In this case the answer is T. To the query '$?D$' we answer 'yes if T'. Think of the thesis example. If an ordinary student wants to know what is missing to get a PhD, the obvious answer is 'a thesis' and not 'a paper and a job offer from Harvard'.

6. The power of our labelling mechanism can be easily illustrated by a more refined use of the labels. If atoms are labelled, for example, by cost (laundry example) the abduction principle can aim for minimal cost. One can also 'cost' the computation itself and aim to abduce on formulas giving maximal provability with a least number of modus ponens instances.

Example 12.16 To show that the abduction depends on the computation let us change the computation to forward chaining or Gentzen like rules. From

$$\text{Data} \vdash ?D$$

we get

$$I \Rightarrow (T \Rightarrow D), I, S \Rightarrow T, O \Rightarrow (P \Rightarrow T) \qquad \vdash ?D$$

which reduces to

$$T \Rightarrow D, S \Rightarrow T, O \Rightarrow (P \Rightarrow T), \qquad \vdash ?D$$

which reduces to the following by chaining:

$$S \Rightarrow D, O \Rightarrow (P \Rightarrow D), \qquad \vdash ?D$$

As we see, not many abduction possibilities are left!

So far we have discussed the possibilities of abduction principles being added to proof rules. We now come to our second new idea, namely:

- Abduction principles are data!

Example 12.17 (Abduction as explanation) Let Δ_1 be the following database:[11]

$$\Delta_1 =: \boxed{\begin{array}{l} \mathit{Holiday}(x,y)\ \mathit{if}\ \mathit{Available}(y) \wedge \mathit{Afford}(x,y)\\ \mathit{Afford}(x,y)\ \mathit{if}\ \mathit{Cost}(y,z) \wedge \mathit{Credit}(x,z)\\ \mathit{Credit}(x,z)\ \mathit{if}\ \mathit{In\_limit}(x,r)\\ \mathit{Sunnyplace}(a_i), i = 1\ldots 3\\ \mathit{Cost}(a_1, 200)\\ \mathit{Cost}(a_2, 250)\\ \mathit{Cost}(a_3, 700)\\ \mathit{Sunnyplace}(b)\\ \mathit{Cost}(b, 500) \end{array}}$$

A customer accesses the database by inserting his visa card into the slot and typing the query (*holiday*, a_1, a_2, a_3). The database can automatically get from the visa card his name x_0, and other credit details, and asks itself the queries:

$$?\mathit{Holiday}(x_0, a_i), i = 1, 2, 3,$$

An ordinary non-intelligent database will simply check if each query succeeds and output the details if it does and output fail if not. A more intelligent database might output an explanation in case of failure, such as:

You can't have a_1 because it is not available

or:

You can't have a_3 because you cannot get credit

An even more intelligent database may recognise, for example, that none of the a_i are available but they are all sunny places and b is a sunny place within the nearest price range. So using a metaprogram of 'interest', 'view' etc, might ask the user (in whatever way it 'communicates'):

[11] Note that our language does not contain '$\wedge$'. However, X if $Y \wedge Z$ is logically equivalent, in all our examples, to $Z \Rightarrow (Y \Rightarrow X)$. The use of '$\wedge$' makes our examples more readable.

Would you like (holiday, b)

and output the details. We are sure many papers in this volume address many of the issues related to the above capabilities.

Ultimately, though, we would like the human to ask in natural language rather than type the 'agreed' form of input (e.g. (holiday, a)) and furthermore reply naturally to a computer query. This might transcend simple 'yes' and 'no' answers: for example, a 'natural' form of response to the computer-generated query above might be:

'b is expensive'.

The computer should understand the English and deduce (as a human would) that the answer to the query is 'no', *and* that any holiday costing more than \$500 is not a candidate. Ideally, we would like to imbue the intelligent database with the same faculty, so that it could now reply:

There are no cheaper holidays

to which a possible form of response might be:

OK, I will take it.

which includes anaphora, *it* referring to b, and withdrawal of the previous denial of b as a suitable holiday.

There are a number of hidden problems here, but essentially we need to spell out two processes: one the reasoning underlying the parsing process, including the interpretation of anaphoric expressions, the access of related information, etc.; and the other the reasoning underlying question-answer dialogues, which may involve indirect answers, negotiation requiring changes in beliefs, plans and goals, and so on.

Example 12.18 Abduction can be data: Consider the following database and query:

a_1: A

a_2: $A \Rightarrow (B \Rightarrow S)$

a_3: B

....

a_4: X, abduce on structure to find X.

a_5: $B \Rightarrow D$

The goal is $?S \otimes D$.

By writing $S \otimes D$ for the goal we are saying we want to partition the database, which is a list of assumptions, into two parts, the first part must prove S and the second part must prove D. This is done in resource logics, where one pays attention to the question of which part of the database proves what.

Such considerations arise in many areas for example in what is known as *parsing as logic.*

Consider the text:

Mary hit John. He cried.

The way this can be analysed is by assuming that each word is assigned a wff of some resource logic (actually concatenation logic, see [Gabbay and Kempson, 1991]) with a label. This assignment is done at the lexical level. Thus a noun n is assigned $n' : NP$. An intransitive verb v_1 is assigned $v_1' : NP \Rightarrow S$. A transitive verb v_2 is assigned $v_2' : NP \Rightarrow (NP \Rightarrow S)$. The pronoun 'he' is assigned an abduction principle. Our problem becomes:

Data

1. Mary$'$: NP
2. hit$'$: $NP \Rightarrow (NP \Rightarrow S)$
3. John$'$: NP
4. he: Abduce on structure. Take the first literal up the list.
5. cried$'$: $NP \Rightarrow S$.

Query

Prove $?S$ or $S \otimes S$ or $?S \otimes S \otimes S \ldots$ etc, in order to show we have a text of sentences.

We are thus saying that Anaphora resolution makes use of structural abduction.

The reader should note that anaphora resolution is a complex area and we are not making any linguistic claims here beyond the intuitive example that abduction principles can be treated as data. We do admit however that logical principles underlying database management do seem to be operative in natural language understanding and we are working full steam ahead in making our case.

Coming back to our view of abduction as data, we are really saying:

- A database can either display data items or give us pointers to where to get them.

Thus a labelled database can appear as below:

n_1: data m

$\vdots \quad \vdots$

n_k: get datum from ...

$\vdots \quad \vdots$

Abductive Labelled Database

12.4.2 Abduction in Planning and Natural Language

We would like to give next a combined example of planning and parsing, based on ideas of [Gabbay and Kempson, 1991].

Example 12.19 (Planning) Consider the situation described by the diagram below.

There are three languages involved

1. The database language containing the predicates On(x, y) and Free(x)
2. The imperative (Input) command language with the predicates Move(x, y).
3. The mixed metalanguage with the connectives '$\wedge$' for 'and' and '$\Rightarrow g$' for 'precondition and action imply postcondition'.

a			t_1: On(a, b)	
b	c		t_1: On(b, tab)	$\leftarrow$ Move(a, c)
			t_1: On(c, tab)	
	a		t_2: On(a, c)	$\leftarrow$ Move(a, tab)
b	c			
b	c	a	t_3: On (a, tab).	

$$\text{On}(x, y) \wedge \text{ Free}(x) \wedge \text{ Free}(z) \wedge \text{Move}(x, z) \Rightarrow g \text{ On}(x, z) \wedge \text{Free}(y) \wedge \text{ Free}(x)$$

The diagram describes the initial layout of the blocks ready to respond to command. t_1 labels all data true at the initial situation and t_2 and t_3 the additional data after each of the actions. We have

$$t_1 < t_2 < t_3.$$

If we query the system with

$$?\ \text{On}(a, x)$$

we get three answers, with different labels, indicating where the answer was obtained in the database, namely:

$$\begin{aligned} &\vdash t_1 : \mathrm{On}(a, b) \\ &\vdash t_2 : \mathrm{On}(a, c) \\ &\vdash t_3 : \mathrm{On}(a, tab) \end{aligned}$$

The reply to the user is determined by the system as the answer proved with the stronger label, namely:

$$\mathrm{On}(a, tab)$$

Call the deductive system governing the planning consideration LDS_1. We remark in passing that this approach offers a *conceptual* (not computational) solution to the frame problem. Conceptually, given an initial labelled database and a sequence of actions to be performed, we model the sequence by another labelled database; the database obtained by adding the results of the actions to the initial database. We label the additions appropriately. This idea will be pursued elsewhere. There are several such 'non-monotonic' solutions in the literature. This is probably the most general. The present involves proving where the blocks are after which action. This system accepts commands in logical form Move(x, y). It does not accept commands in English. If the command comes in English, which we can represent as *move x onto y*, it needs to be parsed into the LDS_1 language. This is done in a parsing logic LDS_0. The following diagram explains the scheme, see Figure 12.1:

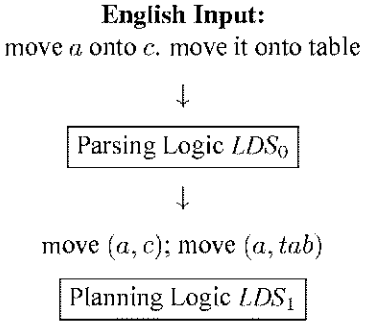

English Input:
move a onto c. move it onto table
Parsing Logic LDS_0
move (a, c); move (a, tab)
Planning Logic LDS_1

Figure 12.1

The following diagram describes the database-query problem of LDS_0:

move′: $NP \Rightarrow (NP \Rightarrow S)$ $?S \otimes S$
a': NP
c': NP
move′: $NP \Rrightarrow (NP \Rrightarrow S)$
it: use abduction. First use structural abduction to
get the first NP higher in the list, then use
inferential abduction to try and get maximal
inferential effects in LDS_1.
tab′: NP

Notice that the abduction principle in LDS_0 also uses inferential effect in LDS_1. Intuitively we are trying to abduce on who 'it' refers to. If we choose 'it' to be a block which is already on the table, it makes no sense to move it onto the table. Thus the command when applied to the database will produce no change. The abduction principle gives preference to abduced formulas which give some effect.

From the logical point of view we are using the following principle, (see also Example 12.18):

- Abduction principles can serve as items of data in one database Δ_0, being a pointer to another database Δ_1 where some computation is carried out jointly involving both databases and the result is the abduced data item in Δ_0.

12.4.3 Abduction in Logic Programming

Example 12.20 (Logic programming abduction) The abductive system in logic programming can be schematically put into our form by making use of the way the Prolog pointer scans the database. An abductive system has the form (Δ, I, A), where $\Delta = (C_1, \ldots, C_n)$ is a sequence of clauses and literals, where I are the integrity constraints and where A is the set of abducible atoms. This system can be translated into the following Horn clause database:

(0) C_0 = Abduce on the goal by checking whether the goal is in A and whether when added it satisfies the integrity constraints.

(1) C_1

$\vdots$ $\vdots$

(n) C_n

We are now ready for our next conceptual step. If an abductive principle is to be considered as a declarative item of data, say Q_{Abduce}, then what would be the meaning of

$$\Delta \; ?! \; Q_{Abduce}$$

For the imperative interaction $!Q_{Abduce}$, the meaning is obvious, we simply apply the abductive principle to the database and get a new database. However the query meaning of the abductive principle is a bit more difficult to explain in general, and we may need to provide an explanation for each specific case. For example, if the abduction principle abduces a formula B, then $\Delta ?Q_{Abduce}$ would mean $\Delta ?B$. This seems all right at first sight, however, the problem here can be that the abduction process by its nature tries to find B's which are not available in the database and so the answer to the query $\Delta ?Q_{Abduce}$ will always be no. This is clearly unacceptable. We must seek another meaning for the query.

Let us for the time being, accept only the imperative reading of $\Delta !Q_{Abduce}$. We can immediately allow ourselves to write databases with clauses containing Q_{Abduce} in them. Let us see through a few examples what this means.

Example 12.21 Let $\Delta = \{Q_{Abduce} \wedge B \Rightarrow D\}$. Think of the above as a database, and assume the computation procedure to be Prolog-like. We consider the following query

$$\Delta ?D$$

which reduces to

$$\Delta ?(Q_{Abduce}, B)$$

which reduces to

$$\Delta, B ?B$$

which succeeds.

Here we assumed that Q_{Abduce} yields B. Our database is similar in this case to the Prolog database:

$$\mathit{Assert}\ (B) \wedge B \Rightarrow D$$

Indeed, asserting is a form of unconditional abduction.

Example 12.22 (Abduction and negation by failure) From our point of view, negation by failure is abduction. This point has also been made in [Eshghi and Kowalski, 1990]. However, we want to make our position crystal clear to avoid confusion. We believe that abduction is a principle of reasoning of equal standing to deduction and that every logical system is comprised of both proof rules and various mechanisms including abductive rules. This view has developed through our interaction with the logics of common sense reasoning and related work in natural language understanding [Gabbay and Kempson, 1991]. Negation by failure is not central to the general abduction scheme, though it is an interesting example from our point of view.

We begin with a precisely specified proof system. The query $\Delta ?Q$ can be algorithmically checked. If the algorithm succeeds, the answer is yes. The algorithm may loop or it may fail. We may be able to prove that for the particular $\Delta ?Q$, the algorithm must fail (e.g. in a case where none of the rules can even be applied). In this case we can say $\Delta?Q$ *finitely fails* (relative to the algorithm). Thus the notion of *finite failure* can be defined for any proof theoretic system.

Given a system, we can consider the following abduction principles which we call $\mathit{Fail}(Q, B)$:

If $\Delta?Q$ *finitely fails then abduce (or assert)* B.

To make our example specific, let us choose a language and computation procedures. By an atomic literal let us understand either an atom $\{p, q, r, \ldots\}$ or an abduction principle $\mathit{Fail}(a, b)$, where a, b are atoms. By clauses let us understand Horn clauses of literals. Goals are conjunctions of literals. Thus we can write the following clauses:

1. $q \wedge \ \mathit{Fail}(a, b) \wedge c$
2. $\mathit{Fail}(q, r)$
3. $a \Rightarrow \ \mathit{Fail}(b, b)$.

To explain the computational meaning, we will translate into Prolog. Ordinary Prolog is not expressive enough for our purpose, so we use N-Prolog [Olivetti and Terracini, 1991; Gabbay and Reyle, 1984; Gabbay, 1985a] with negation by failure, mainly because it allows hypothetical reasoning, i.e. embedded implications.

We translate:

$$\mathit{Fail}(a, b) \mapsto (\neg a \Rightarrow b).$$

After translation, the database becomes:

1. $q \wedge (\neg a \Rightarrow b) \wedge c \Rightarrow p$
2. $\neg q \Rightarrow r$
3. $a \wedge \neg b \Rightarrow b$

which is meaningful computationally in goal directed N-Prolog.

A Horn clause with negation by failure of the form:

$$a \wedge \neg b \Rightarrow c$$

can be translated back into our abductive language as

$$a \Rightarrow \ \mathit{Fail}(b, c)$$

A Prolog goal of the form $\neg a$ an be translated as $\mathit{Fail}(a, \varnothing)$, $\varnothing$ is truth.

N-Prolog is not as expressive as our abductive language. In our abductive language we also have the imperative meaning of

$$\Delta ?\ \mathit{Fail}(a, b)$$

which means *apply* the abduction to Δ.

This would correspond to

$$\mathit{Assert}\ (\neg a \Rightarrow b)$$

in N-Prolog. N-Prolog does not allow for that. The syntax is defined in such a way that we do not get goals of the form $\Delta?(\neg a \Rightarrow b)$. The N-Prolog computation rule would require in this case to add $\neg a$ to Δ, which is not meaningful.

We note that the connection between abduction and negation by failure was observed in [Eshghi and Kowalski, 1990]. Their abductive systems have a special form. They need to rewrite the Horn clause program into a more convenient form, translating the Prolog $\neg a$ as a^* and adding the integrity constraint:

$$a \wedge a^* \Rightarrow .$$

Let us now do in detail the abduction for logic programming, without any rewriting of $\neg a$ into a^*.

Definition 12.23 (Logic programs)

1. *Our language contains the connectives* $\wedge, \neg, \Rightarrow$ *the constant* $\bot$ *and atomic letters* $p, q, \ldots$. *A clause has the form*

$$\bigwedge_i a_i \wedge \bigwedge_j \neg b_j \Rightarrow q$$

where a_i, b_j *and* q *are atoms.* q *is the Head of the clause and* $\bigwedge_i a_i \wedge \bigwedge_j \neg b_j$ *is the Body of the clause. The body may not appear in a clause. An integrity constraint has the same form except that* $q = \bot$. *An integrity constraint must have a body. A program is a set of clauses and integrity constraints.*

2. *Success or failure of an atom* q *from a program* Δ *is defined via the metapredicate* $S(\Delta, q, x)$, *for* $x = 1$ *(success) or* $x = 0$ *(failure) as follows:*

   - $S(\Delta, q, 1)$ *if* $q \in \Delta$
   - $S(\Delta, q, 0)$ *if* q *is not the head of any clause in* Δ.
   - $S(\Delta, q, 1)$ *if for some clause* $\bigwedge_i a_i \wedge \bigwedge_j \neg b_j \Rightarrow q$ *(or integrity constraint if* $q = \bot$*) we have* $\bigwedge_i S(\Delta, a_i, 1) \wedge \bigwedge_j S(\Delta, b_j, 0)$ *holds.*

- *$S(\Delta, q, 0)$ if for every clause (every integrity constraint if $q = \bot$) $\bigwedge_i a_i \wedge \bigwedge_j \neg b_j \Rightarrow q$ there exists an i such that $S(\Delta, a_i, 0)$ or a j such that $S(\Delta, b_j, 1)$ holds.*
- *We write $\Delta \vdash a$ if $S(\Delta, a, 1)$ holds and $\Delta \vdash \neg a$ if $S(\Delta, a, 0)$ holds.*
- *We also write $\Delta?a = x$ for $S(\Delta, a, x)$.*

3. *A program is consistent if $\bot$ is not successful.*

Let us consider an example. Consider the propositional logic program Δ below

$$\begin{array}{l} \neg a \Rightarrow b \\ \neg b \Rightarrow d \\ c \Rightarrow b \\ c \end{array}$$

clearly $\Delta \vdash c \wedge b$ and $\Delta \not\vdash a$.

If we add a to the database and delete c we get $\Delta + a - c \not\vdash b$, and therefore $\Delta \vdash d$. Δ is clearly non-monotonic. By adding a and deleting c we took out b and further added d.

Suppose we want to update Δ to Δ' by ensuring that the query $a \wedge \neg b$ succeeds from Δ'. Thus $\Delta' = \Delta + (a \wedge \neg b)$, where $+$ symbolises our revision process. For this we need an abduction algorithm that will follow the computation of the goal $\Delta?a \wedge \neg b$ and suggest what changes to make to Δ to ensure the success of this query. Δ' is the result of making these changes to Δ. There may be more than on way of doing the change so the suggestions will be disjunctive.

Let us see how it works for our particular example.

Try $\Delta?a = 1$. This fails, since a is not the head of any clause in Δ. So the suggestion here is to add a to Δ.

We write this as

$$\mathit{Ab}(\Delta, a, 1) = \{\{+a\}\}$$

$+a$ means add a to Δ.

We also want b to fail from Δ, so let us ask $\neg b$:
$\Delta?\neg b = 1$ iff $\Delta?b = 0$ iff $\Delta?\neg a = 0$ and $\Delta?c = 0$ iff both $\Delta?a = 1$ and $\Delta?c = 0$.

We know what to suggest here, namely, add a and take out c. Thus $\mathit{Ab}(\Delta, b, 0) = \{\{+a, -c\}\}$. Thus $\mathit{Ab}(\Delta, a \wedge \neg b, 1) = \mathit{Ab}(\Delta, a, 1) \wedge\!\!\!\wedge A(\Delta, b, 0) = \{\{+a, -c\}\}$ where $\wedge\!\!\!\wedge$ means that we choose the cases that are compatible. We need a formal definition.

Definition 12.24 (**of** $\wedge\!\!\!\wedge$)

1. *Let $S_1, \dots, S_k$ be sets of signed literals of the form $\pm q_i$. We say that they are compatible if for no atom q do we have $+q$ in one of the S_i and $-q$ in another.*
2. *Let $\mathsf{A}^i = \{\mathsf{S}^i_1, \dots, \mathsf{S}^i_{m_i}\}$. Define $\bigwedge\!\!\!\bigwedge \mathsf{A}_i$ as the set containing all elements of the form $\mathsf{S} = \bigcup_i \mathsf{S}^i_{k_i}$, where $\mathsf{S}^i_{k_i} \in A_i$ are all compatible. If none exists then $\bigwedge\!\!\!\bigwedge A_i = \varnothing$.*

Definition 12.25 (**Abduction algorithm**)

1. *A propositional logic program Δ is a list, $C_1, \dots, C_k$ of clauses of the form:*

$$C_i = \bigwedge_{j=1}^{m_i} a^i_j \wedge \bigwedge_{j=1}^{n_i} \neg b^i_j \Rightarrow q_i \qquad i = \mathbf{1}, \dots, k,$$

where a^i_j, b^i_j, q_i are atoms, with q_i possibly $\bot$.

2. *Let q be a literal. Define $Ab(\Delta, q, 1)$ and $Ab(\Delta, q, 0)$ by simultaneous induction as follows*

- *$Ab(\Delta, q, 1) = \{\{q\}\}$ if q is not the head of any clause.*
- *$Ab(\Delta, q, 0) = \varnothing$ if q is not the head of any clause.*

Suppose q is the head of the first r clauses

- *$Ab(\Delta, q, 1) = \bigcup_{i=1}^{r} (\bigwedge\!\!\!\bigwedge_{j=1}^{m_i} Ab(\Delta, a^i_j, 1) \wedge\!\!\!\wedge \bigwedge\!\!\!\bigwedge_{j=1}^{n_i} Ab(\Delta, b^i_j, 0))$*
- *$Ab(\Delta, q, 0) = \bigcup_{\alpha} A_\alpha$, where α ranges over all choice vectors such that for each $i, \alpha(i)$ is either some $a^i_{j(i)}, j \leq m_i$ or some $b^i_{j(i)}, j \leq n_i\}$ and*

$$\begin{aligned} \mathsf{A}_\alpha \quad = \quad & \bigwedge\!\!\!\bigwedge_{\{i|\alpha(i)=a^i_{j(i)}\}} Ab(\Delta, \alpha(i), 0) \wedge\!\!\!\wedge \\ & \bigwedge\!\!\!\bigwedge_{\{i|\alpha(i)=b^i_{j(i)}\}} Ab(\Delta, \alpha(i), 1) \end{aligned}$$

- *$Ab(\Delta, \neg q, x) = Ab(\Delta, q, \mathbf{1} - x)$, for $0 \leq x \leq 1$.*
- *$Ab(\Delta, \bigwedge_i A_i, 1) = \bigwedge\!\!\!\bigwedge_i Ab(\Delta, A_i, 1)$.*
- *$Ab(\Delta, \bigwedge_i A_i, 0) = \bigcup_i Ab(\Delta, A_i, 0)$.*

Given a Δ and a goal G, $Ab(\Delta, G, 1)$ (resp. $Ab(\Delta, G, 0)$) gives all the possible options for adding/deleting literals that would make the goal succeed resp. fail). It is up to us to make a choice of what to add/delete. This may involve other considerations, such as ranking or whatever measure we choose to put on the options. Given an inconsistent theory Δ, (i.e. such that $\Delta \vdash \bot$) we can activate

$Ab(\Delta, \bot, 0)$ and find out what to add/delete to restore consistency. $Ab(\Delta, A, 1)$ may be empty.

For example, let $A = a \wedge \neg b$ and $\Delta = a \Rightarrow b$. There is no way that $a \wedge \neg b$ can succeed. This happens because we are trying to leave the clauses of Δ untouched and revise Δ just by adding/subtracting literals. If $Ab(\Delta, A, 1)$ is empty we must be prepared to delete clauses from Δ. How can we do that?

Let us consider the example again

$$\Delta = a \Rightarrow b$$

move to Δ_{name}

$$\Delta_{name} = \{\neg name \wedge a \Rightarrow b\}$$

The clause $a \Rightarrow b$ is active as long as $name$ is not in the database. The minute *name* is added to the data, the clause is practically out.

To allow deleting clauses from Δ we move to Δ_{name} and apply Ab.

Δ_{name} is obtained from Δ by replacing each clause i, of the form $B_i \Rightarrow q_i$ by the clause $name_i \wedge B_i \Rightarrow q_i$ for $i = 1, \ldots, r$.

Let us see what happens with our program.

$$\Delta_n \text{ is } : \{\neg n \wedge a \Rightarrow b\}.$$

The update is: $a \wedge \neg b$.

$$\begin{aligned}
&Ab(\Delta_n, a, 1) = \{\{+a\}\}\\
&Ab(\Delta_n, \neg b, 1) = \{\{-a\}, \{+n\}\}\\
&Ab(\Delta_n, a \wedge \neg b, 1) = Ab(\Delta_n, a, 1) \mathbin{\wedge\!\!\!\wedge} Ab(\Delta_n, \neg b, 1) = \{\{+a, +n\}\}.
\end{aligned}$$

The new program $\Delta_n + (a \wedge \neg b)$ is $\{\neg n \wedge a \Rightarrow b, a, n\}$.

Obviously we have several options for abduction. The first one is to assume that all lines are continuous. This means we are working in a model of the plane of all points (x, y) where x, y are real numbers. It is plausible for us to assume that Euclid has these intuitions at the time. The second option is to assume, based on our knowledge of the extensive use and attention the Greeks gave to ruler and compass constructions, that Euclid had in mind that all ruler and compass construction points are available. In modern terms this means that we are looking at all points (x, y) such that $x, y \in \mathbb{E}$, where $\mathbb{E}$ is the extension field of the rationals closed under the additional operation of getting square roots, i.e. if $x \in \mathbb{E}$ and $x > 0 \Rightarrow \sqrt{x} \in \mathbb{E}$ (i.e. the set of all numbers we can get from rationals by the operations of addition $+$, multiplication $\times$, subtraction $-$, division $/$ and squareroot $\sqrt{\cdot}$)

Which abduction hypothesis do we adopt? Do we add the axiom that the plane is continuously full of points or do we add the axiom that the plane includes all rational as well as rule and compass points?

No geometrical abductive process can decide for us. Only our second logic, our reasoning based on ancient Greek culture can decide.

This example shows our second logic at play!

By the way, Plato's work is full of arguments in need of abductive 'help' and the Cohen and Keyt paper discusses the methodology of 'helping' Plato. In our terminology, the paper discusses the options for a second logic for supporting abduction for enthymemes in Plato.

12.4.4 A Conversation Between Two Intelligent Databases

We have the logical means to allow for two *LDS* databases to negotiate and reach an understanding. Imagine two databases S and H exchanging formulas continuously. At time n, the databases have evolved through the sequences

$$\Delta_1^S, \ldots, \Delta_n^S$$

and

$$\Delta_1^H, \ldots, \Delta_n^H.$$

At time n, database S sends a logical input I_n^S to database H and conversely, database H sends an input I_n^H to database S. The two databases use abduction to update themselves. Thus

$$\Delta_{n+1}^H = \mathit{Abduce}(\Delta_n^H, I_n^S)$$

and

$$\Delta_{n+1}^S = \mathit{Abduce}(\Delta_n^S, I_n^H).$$

To continue and communicate we need action principles for sending the next input. This is also part of our abduction scheme as hinted at in Example 12.19

Let us now consider an extended example:

Given a set of wffs Δ, taken as assumptions, and a formula or query Q, we can pose the logical question $\Delta ?Q$. The meaning of this question is: Does Q follow from Δ in the logic we are dealing with, *i.e.* is $\Delta \vdash Q$ true? There are three possibilities:

1. $\Delta \vdash Q$ holds

2. $\Delta \vdash \sim Q$ holds

3. Neither $\Delta \vdash Q$ nor $\Delta \vdash \sim Q$ hold.

We do not adopt any additional conventions such as $\Delta \not\vdash Q$ implies $\Delta \vdash \sim Q$, where Q is atomic (i.e. the closed-world assumption). We thus want a definite proof of the negation $\sim Q$ in all cases. We denote by $\Delta ?!Q = 1$ the situation where logically the database is consistent and is able to prove either Q or $\sim Q$. In other words $\Delta ?!Q = 1$ means that the database Δ decides the truth of Q in a definite way.

The situation we are going to study arises when there is an exchange of information between an asker, A and a responder R. At the outset of the exchange, A has a set of assumptions Δ_A: so too does R, i.e. Δ_R. Note that Δ_A may include A's beliefs about the content of Δ_R, and vice versa.

Suppose now that A asks R the query Q. If Δ_A were sufficient to prove or disprove Q then we can assume that A could have figured out Q for himself (unless, of course, he wanted to find out if R knew that Q). However, we will begin by assuming that $\Delta_A ?!Q \neq 1$.

R should now respond by supplying some new information B to be added to Δ_A to enable $\Delta_A \cup \{B\}?!Q = 1$. The answer could be:

- $B = Q$ (i.e. 'yes'), in which case $\Delta_A \cup \{Q\}?!Q = 1$;
- $B = {\sim}Q$, (i.e. 'no'), in which case $\Delta_A \cup \{{\sim}Q\}?!Q = 1$;
- some B such that $\Delta_A \cup \{B\}?!Q = 1$;
- some B such that $\Delta_A \cup \{B\}?!Q \neq 1$.

In the first three cases things are fine for A. The situation becomes more interesting when the new data B is still insufficient for A to (logically) determine an answer to Q.

R may have her own reasons for not answering directly and saying Q or $\sim Q$. Perhaps she wants to forestall any future questions. Perhaps she wants to give the reason for Q (or for $\sim Q$), namely B. For whatever reason, and assuming that R is trying to be co-operative rather than evasive, A expects the answer B to logically decide the query Q, i.e. for $\Delta_A \cup \{B\}?!Q = 1$ but nevertheless finds this not to be the case. At this point we invoke what we call *the Rule of Relevant Assumptions*; and require that further data B' should be abduced such that $\Delta_A \cup \{B, B'\}?!Q = 1$. We assume by relevance that R gives A the minimal input B which allows him to decide the query Q. To find B' we need to know and use the computation procedure we have for $\vdash$.

We call the relevance-theoretic principle which allows us to find the enrichment B' *The Rule of Relevant Assumptions (RRA): the principle of consistent least enrichment relative to a proof procedure for* $\vdash$, whereby:

1. Given a proof procedure for finding answers to queries of the form $\Delta ?Q$, then if Δ is a partial database and we ask a query $\Delta ?Q$ and we get the input B, then we try and prove Q from $\Delta \cup \{B\}$ and assume that any information $B_1, B_2, \ldots$ required for the success of the proof was intended to be added

to Δ along with the input B. $\Delta \cup \{B, B_1, B_2, \ldots\}$ must emerge as consistent. When we are talking about proof or deduction, we do not necessarily restrict ourselves to monotonic deduction. The deduction rules used in trying to prove $\Delta \cup \{B\} \vdash Q$ may be defeasible or non-monotonic. These rules correspond better to common sense reasoning. The nature of the logic involved is as yet unspecified by us.

2. If there are several ways of adding $\{B_i\}$ leading to success, we choose the one which involves the least deductive effort for some inferential effect. (These notions will be made precise in later work: they require specification of links between databases so that we can describe for example some concept of closeness between a pair of databases).

We note that the process of finding the necessary B_i requires a given proof procedure for the logic licensing the addition of such assumptions to the database. Thus for different proof procedures we may get different B_i. We also note that the asker A may get his enriched database $\Delta_A \cup \{B, B_i\}$ wrong, in which case further interaction would be required to clarify the matter.

Example 12.26 *Suppose we have a database Δ for an asker A and his query Q to a responder R. Δ could contain the fact that A has apples, and Q could be 'would you like [to eat] an apple?' Here we assume success in the parsing process is achieved and we have some outcome to be evaluated against the database, i.e.:*

$$\begin{aligned} \Delta &\ni \exists x[\mathit{Apple}(x)] \\ Q &= \exists x[\mathit{Apple}(x) \wedge \mathit{Eat}(x)] \end{aligned}$$

The answer A gets from the responder R may be 'yes' (Q) or 'no' ($\sim Q$), or A may get another answer B which allows A to conclude Q or $\sim Q$ (e.g. 'I'm not hungry'). Consider instead the answer: B = 'I don't eat South African apples'. Our proposed formal mechanism for parsing must bring the content of the sentence out, i.e.:

$$B = \forall x[\mathit{SA}(x) \wedge \mathit{Apple}(x) \Rightarrow \sim \mathit{Eat}(x)]$$

A has asked on the basis of Δ whether Q and got the answer B. We can therefore assume by the principle of relevance that $\Delta \cup \{B\}?!Q = 1$, i.e. the extra input information B is expected to (logically) decide the question Q. However, adding B to Δ does not decide Q in this case. Now A can invoke the principle of consistent least enrichment and try to abduce some further data B' such that $\Delta \cup \{B, B'\}?!Q = 1$.

We will now examine one way to compute the enrichment of Δ with some B' to give the desired effect. In this case, we will use as the RRA for this $\vdash$ the (propositional) abduction algorithm in Definition 13.1, below, whereby:

1. $\text{Abduce}(\Delta, Q) = \top$ *if* $\Delta?Q = 1$*;*
2. $\text{Abduce}(\Delta, q) = q$*, for* q *atomic such that* q *is not the head of any clause in* Δ*;*
3. $\text{Abduce}(\Delta, Q_1 \wedge Q_2) = \text{Abduce}(\Delta, Q_1) \wedge \text{Abduce}(\Delta, Q_2)$*;*
4. $\text{Abduce}(\Delta, A_1 \Rightarrow (A_2 \Rightarrow \ldots (A_n \Rightarrow q)\ldots)) = A_1 \Rightarrow (A_2 \Rightarrow \ldots (A_n \Rightarrow \text{Abduce}(\Delta \cup \{A_1, A_2, \ldots A_n\}, q))\ldots)$*;*
5. *Let* q *be atomic and let* $B^j = B^j_1 \Rightarrow (B^j_2 \Rightarrow \ldots (B^j_{n_j} \Rightarrow q)\ldots)$ *for* $j = 1, \ldots, m$ *be all clauses in* Δ *with head* q*. Then* $\text{Abduce}(\Delta, q) = \bigvee_{j=1}^{m} \bigwedge_{i=1}^{n_j} \text{Abduce}(\Delta, B^j_i)$.

If we introduce Skolem constants and Herbrand constants so that we are dealing only with propositional variables, put Δ *in clausal form, and rewrite* $\sim p$ *as* $p \Rightarrow \bot$*, we can apply this abduction algorithm to the example. We also focus on the relevant parts of* Δ*, i.e.* $\exists x[Apple(x)]$.

We want to know whether the answer to $\Delta \cup \{B\}?Q$ *is yes or no. We therefore try and prove both* Q *and* $\sim Q$*, abducing any further assumptions along the way. We have:*

1.	a	*(i.e.* Δ *Skolemised)*
2.	$a \Rightarrow (sa \Rightarrow (e \Rightarrow \bot))$	*(i.e.* B *Herbrandised in clausal form)*

We want to prove $Q = a \wedge e$ *and* $\sim Q = (a \wedge e) \Rightarrow \bot$*. We will examine each in turn.*

For Q *we have:*

$$\begin{aligned}\text{Abduce}(\Delta, a \wedge e) &= \text{Abduce}(\Delta, a) \wedge \text{Abduce}(\Delta, e)\\ &= \top \wedge \text{Abduce}(\Delta, e)\\ &= \text{Abduce}(\Delta, e)\end{aligned}$$

To compute $\text{Abduce}(\Delta, e)$ *we have two policies. Firstly, if we don't unify* e *with* $\bot$*, then the result of* $\text{Abduce}(\Delta, e) = e$*. Since we are assuming that the RRA yields some new information* B' *and* $Q \vdash e$ *anyway, we ignore this result.*

Secondly, we do let e *unify with* $\bot$*. Then the algorithm gives:*

$$\begin{aligned}\text{Abduce}(\Delta, e) &= \text{Abduce}(\Delta, a) \wedge \text{Abduce}(\Delta, sa) \wedge \text{Abduce}(\Delta, e)\\ &= sa \wedge \text{Abduce}(\Delta, e)\end{aligned}$$

We therefore need to solve the following 'equations' for the unknown $x = Abduce(\Delta, e)$*:*

$\Delta \cup \{x\} \vdash e$

$x = sa \wedge x$

x *is minimal in satisfying the two previous properties*

The only solution is $x = sa \wedge e$. This, however, will yield an inconsistent database when added to Δ. Therefore the RRA 'fails' for Q.

For $\sim Q$ we have:

$$\begin{aligned}\text{Abduce}(\Delta, (a \wedge e) \Rightarrow \bot) &= (a \wedge e) \Rightarrow \text{Abduce}(\Delta \cup \{a, e\}, \bot)\\ &= (a \wedge e) \Rightarrow sa\end{aligned}$$

Since, in effect, this came from the Herbrandised query $Apple(c) \wedge Eat(c) \Rightarrow \bot$ (from $\forall y[Apple(y) \wedge Eat(y) \Rightarrow \bot]$, itself rewritten from $\sim \exists y[Apple(y) \wedge Eat(y)]$, i.e. $\sim Q$), we unHerbrandise to give:

$$B' = \forall x[Apple(x) \wedge Eat(x) \Rightarrow SA(x)]$$

which is added to A's database Δ and so yields an answer to $\sim Q$. We now have $\Delta \cup \{B, B'\}?!Q = 1$, as required.

It should be noted that although adding B' is the minimal information enrichment for success (using an implicit ordering relation such that $p \leq (\Rightarrow gp) \Rightarrow q$), it does cost in deductive effort. This additional cost must be offset by some additional inferential effect for A, the details of which are not necessarily accessible to R. In this case, one natural such additional conclusion is that R is telling A that she believes that all the apples he has to eat are South African (and is inviting A to refute that belief before she will eat one of his apples). There may also be other conclusions in respect of other fruit R wouldn't eat for the same reason or, more vaguely, of R's political beliefs. A is encouraged to derive inferences in either of these directions given R's indirect mode of expression, but the only inference he is forced to make is that which induces an answer to his question in one step (this in itself is a reflection of the cognitive cost intrinsic to relevance).

Note that the need to add B' arose directly from the proof procedure involved. We 'hit' on the need to succeed with B' and we decided to add. Different proof procedures may give slightly different results especially when considerations like cost of deduction are involved. But given the question as input, no further enrichment (such as 'If the apples are Polish, R might not eat them either') is even attempted until it is established whether the indicated enrichment itself yields a definitive answer.

Chapter 13

An Abductive Mechanism for the Base Logic

> It is mere rubbish thinking of the origin of life; one might as well think of the origin of matter.
>
> Charles Darwin

13.1 Introduction

The abduction process is in general a recursive metafunction *Abduce* which follows the computation steps of the metapredicate *Success* (from the proof theory of the logic) and yields modification to the database at every step where the *Success* metapredicate seems to fail at its task. The suggested modifications are designed to help *Success* achieve its aim.

We begin with a discussion of the ideas involved in the definition of *Abduce*, for the *LDS* goal directed system for $\Rightarrow$. We then define the abduction algorithm for this system. Later sections proceed to give a simplified version of it for the case of intuitionistic $\Rightarrow$ (where no labels are used), and other implicational logics such as linear logic, relevance logic, the Lambek calcuus and a variety of strict implications.

Let our starting point be Definition 12.12. This definition gives the recursive rules for the metapredicate

$$\mathit{Success}\ (\Delta, \delta : A, \mathit{constraints}, \theta) = x.$$

for $x = 1$ we want the computation to succeed and for $x = 0$ we want it to fail. We now examine what kind of difficulties we encounter in defining the abduction algorithm using Definition 12.12 as a basis:

Case of immediate success or failure
We need to examine what can happen in the case of clauses 1 and 3 of Definition 12.12.

Case $x = 1$
Clause 1 says that *Success* $(\Delta, \delta : q, \mathit{constraints}, \theta) = 1$ (i.e. $\Delta\theta \vdash_{constraints} \delta\theta : q$), if two conditions hold:

(a) The constraints mentioned in *constraints* are all provable in $\mathcal{A}$ for the substitution θ.

(b) $\Psi_1(\Delta\theta, \delta\theta : q)$ holds.

If either (a) or (b) cannot be shown to hold then *success* cannot do its job and our abduction process can be activated to help.

If (a) cannot be shown to hold, we may choose to add axioms to our algebra $\mathcal{A}$ so that the new algebra $\mathcal{A}'$ can prove the constraints. This amounts to a change of the logic (compare with the last paragraph of Section 12.3). Our abduction policy is not to change the logic (at least not in this manner). Thus, in this case our abduction will not help. So let us consider the case where the constraints can be proved but where (b) fails, that is $\Psi_1(\Delta\theta, \delta\theta : q)$ does not hold. We can look for a Δ' such that $\Psi_1(\Delta'\theta, \delta\theta : q)$ does hold.[1]

Whether we can find such a reasonable Δ' depends on Ψ_1. For example, for the case of resource logics and the Ψ_1 suggested in Definition 12.12, namely the Ψ_1 saying that $\delta\theta : q \in \Delta\theta$, we can always find a Δ'; we simply add (input) $\delta : q$

[1] We are already making a serious assumption here in that we want to succeed immediately through the Ψ_1 predicate. Another option is to continue the computation in some manner. The following example illustrates our options.

$$(c \Rightarrow a) \Rightarrow c \vdash ?a$$

In this example a does not unify with any head of clause, so our policy would be to add a to the database. We can adopt a weaker rule however:

If we are stuck with q then we continue with any other head x and add $x \Rightarrow q$ to the database.
In this case we add $c \Rightarrow a$ to the database and carry on with

$$(c \Rightarrow a) \Rightarrow c, c \Rightarrow a \vdash ?c$$

and succeed.

This policy does not guarantee success. Take for example

$$p \Rightarrow q \vdash ?q$$

This reduces to

$$p \Rightarrow q \vdash ?q$$

and adding $q \Rightarrow p$ does not give success.

We can, of course, modify the rule and say add $x \Rightarrow q$ and continue with x whenever x was not asked as a goal in the path leading to q otherwise modify the data to make Ψ_1 succeed (i.e. add q). We shall study different options in the chapter on meatalevel. Meanwhile, let us continue and give one good example of how the abduction machinery can work.

into Δ, i.e. let $\Delta' = \Delta + (\delta : q)$. In other words, using the predicate Ψ_2, we find the Δ' such that $\Psi_2(\Delta, \Delta', \delta : q)$ holds.[2]

For a general Ψ_1, we do not know whether a Δ' can be found. We are going to have to stipulate some propertes of Ψ_1:

- Ψ_1 *Abduction axiom*:
  For every finite Δ and any $\delta : q$ and any $x \in \{0, 1\}$ such that $\Psi_1(\Delta, \delta : q) = x$ ($\Psi_1 = 1$ means Ψ_1 holds, $\Psi_1 = 0$ means Ψ_1 does not hold), there is an effective procedure to produce a finite set (which may be empty) of databases $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_k\}$ such that $\Psi_1(\Gamma_i, \delta : q) = 1 - x$. We denote this set by $\mathbf{Ab}_1(\Delta, \delta : q, 1 - x)$.

Let us see what this algorithm can do for some of the examples already examined. Note that the nature of the abduction depends on the application area it is being used. In the application area all of these algorithms have a meaning, and the choice of algorithm for Ψ_1 and indeed for any other Ψ will be dictated by the needs of the application.

- For $\Psi_1(\Delta, \delta : q) = (\delta : q \in \Delta)$ the algorithm is to let Γ be $\Delta + (\delta : q)$, as already mentioned.

- For the condition $\Psi_1(\Delta, \delta : q) = (\{\delta : q\} = \Delta)$, let the algorithm be to delete all other data from Δ except $\delta : q$ if $\delta : q \in \Delta$ and if $\delta : q \not\in \Delta$ let the algorithm give us nothing. (We do not want to add $\delta : q$ to Δ for reasons having to do with the application area.)

Case $x = 0$
This case is the opposite of the case $x = 1$. We want the computation to fail. So if the computation succeeds, this means that the *constraints* can be proved as well as that Ψ_1 holds. We will not fiddle with the logic and make the constraints unprovable. We will concentrate on Ψ_1. We have $\Psi_1(\Delta\theta, \delta\theta : q)$ holds and we want to make it not hold. Again we need to stipulate an algorithm that yields for each Δ and $\delta : q$ such that $\Psi_1(\Delta, \delta : q)$ holds a set (possibly empty) of $\{\Gamma_1, \ldots, \Gamma_k\}$ options such that $\Psi_1(\Gamma_i, \delta : q)$ fails to hold. We use $\mathbf{Ab}_1(\Delta, \delta : q, 0)$ to produce the set, and say that $\Psi_3(\Delta, \Gamma_i, \delta : q)$ holds.

For example, we can adopt the algorithm that takes $\delta : q$ out of Δ (i.e. let $\Gamma = \Delta - \{\delta : q\}$), i.e. we solve $\Psi_3(\Delta, \Gamma, \delta : q)$.

See Definition 12.25 for an example.

[2] If our language contains $\perp$ or in general our theories have a notion of inconsistency, we may not wish just to insert $\delta : q$ into Δ to get $\Delta' = \Delta + (\delta : q)$. Δ' may be unacceptable to us. In such cases a revision process is needed and we may understand '+' as a revision functor. Thus Ψ_1 involves revision.

Case of $\Rightarrow$:
Let us now consider the case of $\Rightarrow$.

In order to deal with abduction for a goal of the form $\delta : B \Rightarrow C$ we are going to need some additional machinery. We know that *Success*$(\Delta, \delta : B \Rightarrow C,$ *constraints*$,\theta) = x$ iff *success*$(\Delta + (a : B), f(\delta, a) : C,$ *constraints*$,\theta) = x$, where a is a new atomic label and $+$ is the insertion operation. The abduction process should replace Δ by a new Γ which will facilitate the success/failure of the computation. Suppose now that using abduction Γ_a replaces $\Delta + (a : B)$ in giving the desired success/failure value x to the goal $?f(\delta, a) : C$. We now ask ourselves, what is the appropriate replacement Γ of Δ? We want a theory Γ such that upon adding $a : B$ to it (i.e. forming $\Gamma + (a : B)$) we get Γ_a. So we want (using the metaoperator *The* x *such that* $F(x)$ *holds*):

The Γ *such that* $(\Gamma + (a : b) = \Gamma_a)$

let $\mathbf{Ab}_{\Rightarrow}(\Gamma_1, a : B)$ be a metafunction giving *the* Γ *such that* $(\Gamma + (a : B) = \Gamma_1)$, when it exists and $\varnothing$ otherwise.

Let us get a rough idea what this Γ can be.

If we ignore the labels for a moment, we really want the database $\Gamma = (B \Rightarrow \Gamma_1)$, because when we add B to Γ we can get Γ_1.

Thus (ignoring labels) we get

$$\mathbf{Ab}_{\Rightarrow}(\Gamma_1, B) = \{B \Rightarrow X \mid X \in \Gamma_1\}.$$

Since Γ_1 is supposed to be *Abduce*$(\Delta + B, C)$ we get the 'rough' equation:

- *Abduce*$(\Delta, \delta : B \Rightarrow C) = (a : B) \Rightarrow$ *Abduce* $(\Delta + (a, B), f(\delta, a) : C)$

where a is a new constant and the operation $(a : B) \Rightarrow \Gamma_a$ has some meaning for theories Γ_a and units $a : B$. We are not defining this operation but are using it intuitively to explain the kind of operation we need to do abduction in the case of $\Rightarrow$.

If we don't have labels, say in intuitionistic logic, let us see what we get.

Assume the abduction problem is *Abduce*$(\{A \Rightarrow (B \Rightarrow q)\}, B \Rightarrow q)$.

This reduces to

$$B \Rightarrow \textit{Abduce}(\{B, A \Rightarrow (B \Rightarrow q)\}, q)$$

Abduce$(\{B, A \Rightarrow (B \Rightarrow q\}, q)$ reduces to *Abduce*$(\{B, A \Rightarrow (B \Rightarrow q)\}, A)$ and *Abduce*$\{B, A \Rightarrow (B \Rightarrow q)\}, B\}$.

The second computation, *Success*$(\{B, A \Rightarrow (B \Rightarrow q)\}, B) = 1$ is successful and so *Abduce* does not change the database but the first one recommends that we add A to the database. So it is intuitively clear that

$$\textit{Abduce}(\{B, A \Rightarrow (B \Rightarrow q)\}, q) = \{\{B, A \Rightarrow (B \Rightarrow q), A\}\}$$

Therefore our new theory is:

$$\begin{aligned}&\mathit{Abduce}(\{A \Rightarrow (B \Rightarrow q)\}, B \Rightarrow q) = \\ &\quad \{\{B \Rightarrow B, B \Rightarrow (A \Rightarrow (B \Rightarrow q)), B \Rightarrow A\} = \\ &\quad \{\{B \Rightarrow A, B \Rightarrow q\}\} = \{\Delta'\}.\end{aligned}$$

Let us take another point of view of abduction. Since intuitionistic logic is monotonic, let us look at $\mathit{Abduce}(\Delta, A)$ as telling us *what to add* to Δ to obtain a Δ' which makes A succeed. In this case $\mathit{Abduce}(\{A \Rightarrow (B \Rightarrow q\}, B \Rightarrow q) = B \Rightarrow \mathit{Abduce}(\{A \Rightarrow B \Rightarrow q), B\}, q) = B \Rightarrow A$.

So $B \Rightarrow A$ needs to be added to Δ to get Δ'. Thus

$$\begin{aligned}\Delta' &= \{A \Rightarrow (B \Rightarrow q), B \Rightarrow A\} \\ &= \{B = q, B \Rightarrow A\}.\end{aligned}$$

We are getting the same result!

Let us see if we can be more specific about the operation $(a : B) \Rightarrow \Gamma$, of the labels.

Consider the database $\Gamma_1 = \{\alpha : A\}$. We want to turn it into $\Gamma = \{\beta : B \Rightarrow A\}$ such that $\Gamma + (a : B)$ is 'equivalent' to $\alpha : A$.

Let us try the modus ponens:

$$\frac{\beta : B \Rightarrow A; a : B}{f(\beta, a) : A}.$$

thus we have to solve *the* β *such that* $(f(\beta, a) = \alpha)$, and let

$$(a : B) \Rightarrow \{\alpha : A\} = \{\beta : B \Rightarrow A\}.$$

Thus $\mathbf{Ab}_{\Rightarrow}(\{\alpha : A\}, a : B) = \{\{\beta : B \Rightarrow A\}\}$.

Of course in a general *LDS* database the $\mathbf{Ab}_{\Rightarrow}$ function can be more complex to define.

Case of database decomposition:
We now consider the case where an atomic query $\delta : q$ unifies with a head of a clause in the database. This is rule 4 of Definition 12.12.

This case decomposes the database into several databases using the Ψ_4 predicate. This decomposition forces us to consider the question of what to do with the replacement databases proposed by *Abduce* after the decomposition. We will get output for the abduction metafunction at the points of difficulty of the *Success* predicate but these replace databases deep into some iterated decomposition. How do we propagate the replacements upwards?

Let us illustrate the problem by being more specific. The abductive algorithm for the Ψ_1 case, both for $x = 0$ and $x = 1$, for example, replaces Δ

with a new Γ. This Γ is proposed in the middle of an overall computation. We need to examine the 'replacement' operation and how it interacts with decomposition. A simple example will illustrate our problem (we will ignore the constraints). Suppose we have a clause $E = (q_1 \Rightarrow (q_2 \Rightarrow a)) \in \Delta$. We ask $\mathit{Success}(\Delta, \delta : a, \mathit{constraints}, \theta) = 1$. Rule 4 of Definition 12.12 tells us to look at partitions databases Δ'_1 and Δ'_2 such that $\Psi_4(\Delta, \{E\}, \Delta'_1, \Delta'_2)$ holds and labels δ_1, δ_2 such that

$$\mathit{success}(\Delta'_i, \delta_i : q_i, \mathit{constraints}', \theta_i) = 1 \text{ for } i = 1, 2 \text{ hold.}$$

E appears in Ψ_4 as a parameter. It may be that to make these succeed we abduce Γ_1, Γ_2 for the two subcomputations.

Our question is how does the replacement of Δ'_i by Γ_i yield a replacement of Δ by Γ? What is Γ? How do we construct it?

Recall that the idea of abduction is to ask for the initial database Δ and a goal $\delta : A$ the predicate $\mathit{Success}(\Delta, \delta : A, \mathit{constraints}, \theta) = x$ and we expect our abduction metafunction to answer: $\mathit{Abduce}(\Delta, \delta : A, \mathit{constraints}, \theta, x) = \{\Gamma\}$. Thus Γ is the result of abduction and then we are supposed to be assured that $\mathit{Success}(\Gamma, \delta : A, \mathit{constraints}, \theta) = x$ succeeds!

Let us present our problem clearly. We have $\Delta'_1, \Delta'_2, \Gamma_1, \Gamma_2$ and Δ and we are looking for a Γ. What we know is the following:

- $\Psi_4(\Delta, \{E\}, \Delta'_1, \Delta'_2)$
- Γ_1 is abduced to replace Δ'_1
- Γ_2 is abduced to replace Δ'_2
- Ψ_2 holds for Γ_i and $\delta_i : q_i$.

We need to know a corresponding Γ to replace Δ such that $\Psi_4(\Gamma, \{E\}, \Gamma_1, \Gamma_2)$ holds.

To find the Γ we need to know more about the Ψs.

We need some basic axioms relating the properties of the Ψs. We perhaps need to stipulate inverse functions that allow one to retrieve Δ from any $\Delta_1, \ldots, \Delta_n$, i.e. we need

- $\forall n \forall \Delta_1, \ldots, \Delta_n \exists ! \Delta \Psi_4(\Delta, \Delta_1, \ldots, \Delta_n)$.

  Let us introduce a new function $\mathbf{Ab}_+$ with $\mathbf{Ab}_+(\Delta_1, \ldots, \Delta_n) = \Delta$.
- $\bigwedge_{j=1}^{n} \Psi_4(\Delta_n, \Delta_n^1, \ldots, \Delta_n^{k(n)}) \wedge \Psi_4(\Delta, \Delta_1, \ldots, \Delta_n) \Rightarrow$ $\Psi_4(\Delta, \Delta_1^1, \ldots, \Delta_1^{k(1)}, \ldots, \Delta_n^1, \ldots, \Delta_n^{k(n)})$.
- $\forall \Delta \forall n \exists \Delta_1, \ldots, \Delta_n \Psi_4(\Delta, \Delta_1, \ldots, \Delta_n)$. (Note that the decomposition need not be disjoint and Δ_i can be empty.)

- $\Psi_2(\Delta, \Delta', a : B) \Rightarrow \Psi_3(\Delta', \Delta, a : B)$

We also want some compatibility of Ψ_4 with Ψ_2 and Ψ_3. If we regard Ψ_4 as a decomposition of Δ into $\Delta_1, \ldots, \Delta_n$ and regard $\Psi_2(\Delta, \Delta', a : B)$ as insertion of $a : B$ into Δ to form Δ' then do we expect $\Psi_4(\Delta', \Delta, \{a : B\})$ to hold?

We can have an even stronger condition, namely that Δ can be generated by successive inputs of its own elements, namely:

- $\forall \Delta \exists n \exists (\alpha_1 : A_1, \ldots, \alpha_n : A_n) \exists \Delta_1, \ldots, \Delta_{n-1} \bigwedge_{i=1}^{n-1} \Psi_2(\Delta_{i+1}, \Delta_i, \alpha_{i+1} : A_{i+1}) \wedge \Psi_2(\varnothing, \Delta_1, \alpha_1 : A_1)$.

The problem of soundness

This is not all that is needed. We started from a decomposition Δ'_1, Δ'_2 of Δ. We abduced and got Γ_1, Γ_2 and formed Γ. Now we must assume the original clause $E = (q_1 \Rightarrow (q_2 \Rightarrow a))$ is in Γ. In such a case, it is obvious that we can use the clause to decompose Γ back into Γ_1, Γ_2, and ensure that the computation from Γ succeeds.[3]

What happens if we want to fail? In this case for every clause, say $q_1 \Rightarrow (q_2 \Rightarrow a)$ and every decomposition, say Δ'_1, Δ'_2 the *abduce* predicate will choose some replacement, say Γ_1 for Δ'_1 for which the computation $\Gamma_1?q_1$ fails. Thus when we form Γ to replace Δ, we do not care if $q_1 \Rightarrow (q_2 \Rightarrow a) \in \Gamma$. If it is in Γ, however, we must worry whether it gives rise to *new* decompositions Γ'_1, Γ'_2 such that $\Gamma'_i?q_i$ succeed. In fact we must worry about new clauses in Γ with heads q for which there are successful decompositions. This is again the problem of ensuring soundness

There are two ways for going around this difficulty. The first is to make some strong assumptions on the decomposition and $\mathbf{Ab}_+$ predicates. One such possible assumption is that Γ has the same clauses with head q as Δ and the changes are only in addition or deletion of atoms. Such an assumption may not be enough, however, without knowing the specific properties of Ψ_4 we may not be able to formulate additional assumptions.

The second method may be in principle more general. We compose $\Gamma = \mathbf{Ab}_+(\Gamma_1, \Gamma_2)$ to replace Δ but we are not assured (by any strong assumptions) that $\Gamma?q$ fails. We compute *Success*(Γ, q). If it fails, then fine, we take this Γ as the result of the abduction. If it succeeds, then we either say that the abduction process produces nothing or we apply abduction again to $\Gamma?q$. This will produce a Γ_1, and now we iterate the process. To ensure that this process does not loop forever, we could introduce some complexity assumptions either on the number of iterations or on Ψ_4 and $\mathbf{Ab}_+$ which will tell us that abduction involves adding or deleting wffs from a finite set of wffs constructed from the initial database and query. This assumption is enough to allow for a historical loop checker to work.

[3] This means that the function $\mathbf{Ab}_+(\Delta_0, \Delta_1, \ldots, \Delta_n) = \Delta$ must satsify $\Delta_0 \subseteq \Delta$. Δ_0 is the parameter $\{E\}$ in our case.

13.2 Abduction Algorithm for $\Rightarrow$

Following our discussion, we are now ready to write the abduction algorithm.

Definition 13.1 (Abduce metafuction)

1. $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 1) = \{\Delta\}$ *if* $\text{Success}(\Delta, \delta : q, \text{constraints}, \theta) = 1$.
2. $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 1) = \varnothing$ *if* $\mathcal{A} \nvdash \text{constraints}\theta$
3. $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 1) = \{\Gamma_1, \ldots, \Gamma_k\}$ *if* $\mathcal{A} \vdash \text{constraints}\,\theta$ *and* $\Psi_1(\Delta\theta, \delta\theta : q)$ *does not hold and* $\mathbf{Ab}_1(\Delta\theta, \delta\theta : q, 1) = \{\Gamma_1, \ldots, \Gamma_k\}$. *Note that each* $\Gamma \in \mathbf{Ab}_1$ *satisfies* $\Psi_1(\Gamma\theta, \delta\theta : q)$. $\mathbf{Ab}_1$ *may yield* $\varnothing$ *if the abduction is not possible.*
4. $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 0) = \{\Delta\}$ *if* $\text{Success}(\Delta, \delta : q, \text{constraints}, \theta) = 0$.
5. $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 0) = \{\Gamma_1, \ldots, \Gamma_k\}$ *if* $\mathcal{A} \vdash \text{constraints}\,\theta$ *and* $\Psi_1(\Delta\theta, \delta\theta : q)$ *does hold and* $\mathbf{Ab}_1(\Delta\theta, \delta\theta : q, 0) = \{\Gamma_1, \ldots, \Gamma_k\}$.

   Note that for each $\Gamma \in \mathbf{Ab}_1$ *we have that* $\Psi_1(\Gamma\theta, \delta\theta : q)$ *does not hold.* $\mathbf{Ab}_1$ *may give* $\varnothing$ *if the abduction is not possible.*
6. $\text{Abduce}(\Delta, \delta : B \Rightarrow C, \text{constraints}, \theta, x) = \{\Gamma_1, \ldots, \Gamma_k\}$, *where for each* Γ_i *there exists a* Γ'_i *such that* $\Gamma'_i \in \text{Abduce}(\Delta + (a : B), f(\delta, a) : C, \theta, x)$ *and* $\Gamma_i = \mathbf{Ab}_{\Rightarrow}(\Gamma'_i, (a : B))$.

We now consider the case of decomposition. We need some notation first. Let Δ *be a database and* $\delta : q$ *a goal. By a decomposition* $\mathbb{D}(\Delta, \delta : q)$ *we mean a choice of a clause* E *of the form* $\alpha : A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q) \ldots)$ *in* Δ *called the clause of* $\mathbb{D}$ *and of databases* $\Delta'_1, \ldots, \Delta'_n$ *such that* $\Psi_4(\Delta, \{E\}, \Delta'_1, \ldots, \Delta'_n)$ *holds.* n *is referred to as the length of the decomposition and depends on* E. *To introduce the rules for* Abduce *recall that by rule 4 of Definition 12.12,* $\text{Success}(\Delta, \delta : q, \text{constraints}, \theta) = 1$ *(resp.* $= 0$*) if for some* $\mathbb{D}$ *(resp. all* $\mathbb{D}$*) and some* $\delta_1, \ldots, \delta_n$ *we have for all* i *(resp. some* i*)* $\text{Success}(\Delta'_i, \delta_i : A_i, \text{constraints}', \theta') = 1$ *(resp.* $= 0$*).*

7. $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 1)$ *wants to make success out of failure so it will give all options for success. So for each* $\mathbb{D}$*, it will give a set of* $\Gamma^{\mathbb{D}}_{\mathbf{e}_1}, \Gamma^{\mathbb{D}}_{\mathbf{e}_2}, \ldots$ *that will make this choice of* $\mathbb{D}$ *succeed. For the choice of* $\mathbb{D}$ *to succeed we need all* $\text{Success}(\Delta'_i, \delta_i : A_i, \text{constraints}', \theta') = 1$.

   Let $\text{Abduce}(\Delta'_i, \delta_i : A_i, \text{constraints}', \theta', 1) = \{\Gamma^{\mathbb{D}}_{i,1}, \Gamma^{\mathbb{D}}_{i,2} \ldots \Gamma^{\mathbb{D}}_{i,k(i)}\}$, *i.e. the set of options for replacing* Δ'_i *and succeeding.*

The replacement for Δ *will be all possible choices* $\Gamma^{\mathbb{D}}_{\mathbf{e}} = \{\mathbf{Ab}_{+}(\{E\}, \Gamma^{\mathbb{D}}_{1,\mathbf{e}(1)}, \ldots, \Gamma^{\mathbb{D}}_{n,\mathbf{e}(n)}) \mid \mathbf{e}$ *a function giving each* $1 \leq i \leq n$ *a value* $1 \leq \mathbf{e}(i) \leq k(i)\}$. *We therefore define:* $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 1) = \{\Gamma^{\mathbb{D}}_{\mathbf{e}} \mid \mathbb{D}, \mathbf{e}$ *as above and such that the clause of* $\mathbb{D}$ *is in* $\Gamma_{\mathbf{e}}\mathbb{D}\}$.

8. *Consider now* $\text{Success}(\Delta, \delta : q, \text{constraints}, \theta) = 0$. *This fails if for some* $\mathbb{D}$ *we have for all* $i = 1, \ldots, n, \text{Success}(\Delta'_i, \delta_i : A_i, \text{constraints}'\theta') = 1$.

   So for every $\mathbb{D}$ *we choose an index* $1 \leq \mathbf{c}(\mathbb{D}) \leq n$ *and replace* ${\Delta'}_{\mathbf{c}(\mathbb{D})}$ *by* ${\Gamma'}_{\mathbf{c}(\mathbb{D})}$ *from the set* $\text{Abduce}({\Delta'}_{\mathbf{c}(\mathbb{D})}, \delta_{\mathbf{c}(\mathbb{D})} : A_{\mathbf{c}(\mathbb{D})}, \text{constraints}', \theta', 0)$.

   Let $\Gamma^{\mathbb{D}}_{\mathbf{c}} = \mathbf{Ab}_{+}(\{E\}, \Delta'_1, \ldots, {\Delta'}_{\mathbf{c}(\mathbb{D})-1}, {\Gamma'}_{\mathbf{c}(\mathbb{D})}, {\Delta'}_{\mathbf{c}(\mathbb{D})+1}, \ldots, \Delta'_n)$.

   Let $\text{Abduce}(\Delta, \delta : q, \text{constraints}, \theta, 0) = \{\Gamma^{\mathbb{D}}_{\mathbf{c}} \mid \mathbb{D}$ *and* $\mathbf{c}$ *as above and* $\text{Success}(\Gamma^{\mathbb{D}}_{\mathbf{c}}, \delta : q, \text{constraints}, \theta) = 1\}$.

9. *In case there are no strong assumptions on* $\mathbf{Ab}_{+}$ *and* Ψ_4 *ensuring that the* $\Gamma^{\mathbb{D}}_{\mathbf{c}}$ *produced in the preceding item 8 above are sound, then we iterate the* Abduce *computation for each candidate* $\Gamma^{\mathbb{D}}_{\mathbf{c}}$ *until we either find such candidates or give up and produce* $\varnothing$.

Theorem 13.2 (Soundness of abduction) *If* $\Gamma \in \text{Abduce}(\Delta, \delta : A, \text{constraints}, \theta, x)$, *then* $\text{Success}(\Gamma, \delta : A, \text{constraints}, \theta) = x$.

Of course, $\text{Abduce}(\Delta, \delta : A, \text{constraints}, \theta, x)$ *may be empty.*

Proof. By induction on the definition of *Abduce*.

We follow the clauses of Definition 13.1

1. clear

2. nothing to prove

3. By definition $\mathbf{Ab}_1(\Delta\theta, \delta\theta : q, 1)$ produces Γ_i which satisfy Ψ_2 and so $\mathit{Success} = 1$ holds.

4. Clear

5. By definition $\mathbf{Ab}_1(\Delta\theta, \delta\theta : q, 0)$ produces Γ_i for which there is failure of Ψ_1 and so $\mathit{Success} = 0$ holds.

6. Clear from the definitions.

7. The abduction replaces Δ by Γ. To check the success predicate for Γ we had to assume that $\mathit{Success}(\Gamma, \delta : q, \mathit{constraints}, \theta) = 1$.

8–9. Again measures were introduced in this clause to ensure that the choices of Γ produced by the *Abduce* function do the job. ■

13.3 Case Study: Abduction for Intuitionistic Implications

Intuitionistic logic is monotonic and requires no labels and therefore no constraints. Doing abduction for it becomes simple. In Definition 12.12, $\Psi_1(\Delta, q)$ becomes $q \in \Delta$. The decomposition $\Psi_4(\Delta, \{E\}, \Delta_1, \ldots, \Delta_n)$ does not need the parameter Δ_0 and can be taken as $\bigwedge_{i=1}^{n} \Delta_i = \Delta$. $\Psi_2(\Delta, \Delta', A$ is $\Delta' = \Delta \cup \{A\}$ and $\Psi_3(\Delta, \Delta', A)$ is $\Delta' = \Delta - \{A\}$.

The computation rules for *Success*$(\Delta, A) = x, x \in \{0, 1\}$ become the following:
Assume our initial goal is G

Definition 13.3 (Success for intuitionistic logic)

1. Immediate success case
Success$(\Delta, q) = 1$, *for q atomic if $q \in \Delta$.*

2. Implication case
Success$(\Delta, B \Rightarrow C) = x$ *if* Success$(\Delta \cup \{B\}, C) = x$.

3. Immediate failure case
Success$(\Delta, q) = 0$, *q atomic, if q is not the head of any clause in Δ.*

4. Cut reduction case
Success$(\Delta, q) = 1$ *(resp.* Success$(\Delta, q) = 0$*) for q atomic, if for some (resp. all) clauses of the form $A_1 \Rightarrow (A_2 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q)\ldots) \in \Delta$ we have that for all i (resp. for some i) we have* Success$(\Delta, A_i) = 1$ *(resp.* Success$(\Delta, A_i) = 0$*).*

5. Consequence
We have $\Delta \vdash A$ iff Success$(\Delta, A) = 1$.

The above rules (1) to (5), identify $\Rightarrow$ as intuitionistic implication. To obtain $\Rightarrow$ as classical implication, we add the

6. Restart Rule.
Success$(\Delta, q) = 1$ *if* Success$(\Delta, G) = 1$, *where G is the initial goal.*

We can therefore see *Abduce*(Δ, A, x) as yielding a set of wffs to add (for $x = 1$) or delete (for $x = 0$) from Δ to make A succeed or respectively fail.

We can therefore expect the following:

- $\Delta + \mathit{Abduce}(\Delta, A, 1) \vdash A$
- $\Delta - \mathit{Abduce}(\Delta, A, 0) \nvdash A$

We can simplify our notation and write $\Delta?Q = x$ for $\mathit{Success}(\Delta, Q, \mathit{constraints}, \theta) = x$ and $\mathit{Abduce}^{\pm}(\Delta, Q) = \{\Gamma_1, \ldots, \Gamma_k\}$, where Γ_i are all alternative sets of wffs to be added or taken out of Δ to yield the desired results.

We therefore have the following connection between the old and new notations:

$$\mathit{Abduce}(\Delta, Q, \mathit{constraints}, \theta, 1/0) = \{\Delta \pm \Gamma \mid \Gamma \in \mathit{Abduce}^{\pm}(\Delta, Q)\}$$

We are not going to define Abduce^{-} in this section. It involves problems of deletion an dneeds special attention.[4]

Definition 13.4 (Abduce$^+$ for intuitionistic logic)

1. $\text{Abduce}^+(\Delta, Q) = \{\varnothing\}$ *if* $\Delta?Q = 1$

2. $\text{Abduce}^+(\Delta, q) = \{q\}$ *if q is atomic and is not the head of any clause in Δ*

3. $\text{Abduce}^+(\Delta, A_1 \Rightarrow (A_2 \Rightarrow \ldots (A_n \Rightarrow q) \ldots) = \{A_1 \Rightarrow (A_2 \Rightarrow \ldots (A_n \Rightarrow X) \ldots) \mid X \in \text{Abduce}^+(\Delta \cup \{A_1, \ldots, A_n\}, q)\}$.

For clause (4) below, we need to assume that

$$B^j = (B_1^1 \Rightarrow \ldots \Rightarrow (B^j_{n(j)} \Rightarrow q) \ldots), j = 1, \ldots, m,$$

lists all clauses of heads q in Δ.

4. $\text{Abduce}^+(\Delta, q) = \{\bigcup\{X_i \in \text{Abduce}^+(\Delta, B_i^j) \mid i = 1, \ldots, n(j)\} \mid j = 1, \ldots, m\}$.

Examples 13.5

1. Consider $\Delta = \{a\}?q$.
The Abduce$^+$ *wff is q.*
The new abduced theory for success is $\{a, q\}$.

[4] We have to be careful with the definition of Abduce^-. Consider the abduction problem of wanting q to fail from $\{p, p \Rightarrow q\}$. Obviously the answer is to delete p. Thus $\mathit{Abduce}^-(\{p, p \Rightarrow q\}, q) = \{p\}$. However, if we consider the equivalent problem of wanting $p \Rightarrow q$ to fail from $p \Rightarrow q$, i.e. $\mathit{Abduce}^-(\{p \Rightarrow q\}, p \Rightarrow q)$, we want to reduce it to $\mathit{Abduce}^-(\{p \Rightarrow q, p\}, q)$. The obvious thing to write is:

- Add to the database the condition $p \Rightarrow \mathit{Abduce}^-(\{p \Rightarrow q, p\}, q)$, which should mean if p is added then delete what Abduce^- tells you to delete. In metalevel notation we get $p \Rightarrow$ Delete $\{p\}$, as the Abduced$^-$ wff.

For this we need metalevel. This is addressed in [Gabbay *et al.*, 2002; Gabbay *et al.*, 2004].

2. *Consider* $\Delta = \{a \Rightarrow (b \Rightarrow q), a, c \Rightarrow (d \Rightarrow q)\}?q$.
The Abduce$^+$ *candidates are* $\{\{b\}, \{c \Rightarrow d\}\}$. *There are two new abduced theories for success* $\Delta \cup \{b\}$ *and* $\Delta \cup \{c, d\}$.

3. *Consider* $\Delta = \{q, a \Rightarrow q\}?q$. *The* Abduce$^-$ *candidate is* $\{q\}$. *The new abduced theory for failure is* $\Delta - \{q\} = \{a \Rightarrow q\}$.

Theorem 13.6 $\Delta \cup \Gamma \vdash Q$ *for any* $\Gamma \in$ Abduce$^+(\Delta, Q)$.

Proof. The proof is by induction on the definition of the *Abduce* predicate.

1. If the set *Abduce*(Δ, Q) contains $\varnothing$ then this means $\Delta \vdash Q$.

2. Assume that $Q = q$ is atomic and that it is not the head of any clause. Then the abduced formula is q and clearly $\Delta, q \vdash q$.

3. Assume Q has the form

$$Q = (A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q)\ldots).$$

Then the abduced set is

$$A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow \mathit{Abduce}^+(\Delta \cup (A_1, \ldots, A_n), q), \ldots).$$

where $A_1 \Rightarrow \ldots (A_n \Rightarrow \{\Gamma_1, \ldots, \Gamma_k\})\ldots)$ denotes the set of all theories of the form $\Gamma_i^{\Rightarrow}$

$$\Gamma_i^{\Rightarrow} = \{A_1 \Rightarrow (A_2 \ldots \Rightarrow (A_n \Rightarrow X)\ldots) \mid X \in \Gamma_i\}.$$

We need to show $\Delta \cup \Gamma_i^{\Rightarrow} \vdash Q$ for any $\Gamma_i^{\Rightarrow} \in A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow \mathit{Abduce}^+(\Delta \cup \{A_1 \ldots, A_n\}, q)\ldots)$.

By the induction hypothesis

$$\Delta \cup \{A_1, \ldots, A_n\} \cup \Gamma_i \vdash q \text{ for any } \Gamma_i \in \mathit{Abduce}^+(\Delta \cup (A_1, \ldots, A_n), q)$$

hence

$$\Delta \cup \{A_1, \ldots, A_n\} \cup \Gamma_i^{\Rightarrow} \vdash q$$

hence

$$\Delta \cup \Gamma_i \vdash Q$$

4. Assume $Q = q$ is atomic and let $B^j = (B_1^j \Rightarrow \ldots \Rightarrow (B_{n(j)}^j \Rightarrow q)\ldots), j = 1, \ldots, m$, be all the clauses in Δ with head q.

Then we need to show for

$$\Delta \cup \Gamma \vdash q \text{ for } \Gamma \in \mathit{Abduce}^{+}(\Delta, q)$$

where $\mathit{Abduce}^{+}(\Delta, q) = \{\Gamma_1^j \cup \ldots \cup \Gamma_{n(j)}^j \mid \Gamma_i^j \in \mathit{Abduce}^{+}(\Delta, B_i^j), j = 1, \ldots, m\}$.

By the induction hypothesis for j fixed, we have

$$\Delta \cup \Gamma_i^j \vdash B_i^j$$

for $i = 1, \ldots, n(j)$.

Hence for each j, since B^j is in Δ, we have:

$$\Delta \cup \bigcup_{i=1}^{n(j)} \Gamma_i^j \vdash q.$$

■

To appreciate the dependence of abduction on the computation procedures let us give another algorithm for intuitionistic logic. The computation metapredicate is $\mathit{DSuccess}(\Delta, A, H) = x$, where Δ is a set of wffs (the database), A is the current goal and H is a list of previous atomic goals. $x \in \{0, 1\}$. The 'D' in *DSuccess* stands for diminishing (resource), when we use a clause we throw it out. We need some definitions

Definition 13.7 *Let* $(x_1, \ldots, x_n)$ *be a list. Let* y *be an element and* x *be in the list. We define the relation* x is accessible to y *as follows.*

1. x *is accessible to* y *in one step if for some* x_i, y_j *in the list we have* $i \geq j$ *and* $x_i = y$ *and* $x_j = x$.

2. x *is accessible to* y *in* $m + 1$ *steps if for some* y', y' *is accessible to* y *in one step and* x *is accessible to* y *in* m *steps.*

3. x *is intuitionistic accessible to* y *if for some* m, x *is accessible to* y *in* m-*steps.*

4. x *is classically accessible to* y *if* x *is in* H.

Definition 13.8 Diminishing resource DSuccess *for inuitionistic logic.*

1. Immediate DSuccess case:
DSuccess$(\Delta, q, H) = 1$ *if* $q \in \Delta$

2. Immediate failure case*:*
DSuccess$(\Delta, q, H) = 0$ *if neither* q *nor any accessible* $p \in H$ *to* q *is head of any clause in* Δ.

3. Implication case*:*
   $\mathrm{DSuccess}(\Delta, B \Rightarrow C, H) = x$ *if* $\mathrm{DSuccess}(\Delta \cup \{B\}, C, H) = x$

4. Cut reduction case*:*
   $\mathrm{DSuccess}(\Delta \cup \{A_1 \Rightarrow (A_1 \Rightarrow (\ldots(A_n \Rightarrow q)\ldots)\}, q, H) = 1$ *(respectively* $= 0$*) iff for all* $i = 1, \ldots, n$ *(respectively for some* i*) we have* $\mathrm{DSuccess}(\Delta, A_i, H * (q)) = 1$ *(respectively* $= 0$*), where* $H * (q)$ *is the list obtained by appending* q *to* H *at the end.*

5. Bounded Restart Case
   $\mathrm{DSuccess}(\Delta, q, H) = 1$ *iff for some* y *intuitionistically accessible to* q *in* H *we have* $\mathrm{DSuccess}(\Delta, y, H * (q)) = 1$.

6. Restart Rule
   $\mathrm{DSuccess}(\Delta, q, H) = 1$ *iff for some* y *in* H $\mathrm{DSuccess}(\Delta, y, H) = 1$.

7. Consequence[5]
   We have $\Delta \vdash q$ *in intuitionistic logic (respectively in classical logic) iff* $\mathrm{DSuccess}(\Delta, q, \varnothing) = 1$*, using Bounded Restart, (respectively using Restart), where* $\varnothing$ *is the empty list.*

Example 13.9

$$c \Rightarrow a, (c \Rightarrow a) \Rightarrow c \vdash ?a, \varnothing$$

reduces to

$$(c \Rightarrow a) \Rightarrow c \vdash ?c, (a)$$

rreduces to

$$\varnothing \vdash ?c \Rightarrow a, (a, c)$$

reduces to

$$c \vdash ?a, (a, c)$$

Since c is accessible to a we continue to reduce to

$$c \vdash ?c, (a, c, a)$$

DSuccess

Example 13.10 (Of abduction) *Try now*

$$(c \Rightarrow a) \Rightarrow c \vdash ?c, \varnothing$$

[5] See [Gabbay and Olivetti, 2000].

reduces to

$$\varnothing \vdash ?c \Rightarrow a, (c)$$

reduces to

$$c \vdash ?a, (c)$$

c is not accessible to a and therefore we fail unless we abduce. We can abduce by adding a.

Note that in classical logic $(c \Rightarrow a) \Rightarrow c \vdash c$ and hence we should be able to succeed. The restart rule in our context says that any x in H is accessible and hence we can continue in this case and ask for c and succeed.

Definition 13.11 (DAbduce for intuitionistic logic) *The abduction on the* DSuccess *computation follows the same lines as the abduction for* Success *in Definition 13.3*

1. $\text{DAbduce}^+(\Delta, Q, H) = \varnothing$ *if* $\text{DSuccess}(\Delta, q, H) = 1$
2. $\text{DAbduce}^+(\Delta, q, H) = \{q\}$ *if q is atomic and neither q nor any accessible y from H is the head of any clause in Δ.*
3. $\text{DAbduce}^+(\Delta, A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q), H) = \{A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow X)\ldots) \mid X \in \text{DAbduce}^+(\Delta \cup \{A_1, \ldots, A_n\}, q, H)\}$ *where* $A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow X)\ldots) = \{A_1 \Rightarrow \ldots \Rightarrow `(A_n \Rightarrow y) \mid y \in X\}$.
4. $\text{DAbduce}^+(\Delta, q, H) = \{\Gamma_B \mid B \in \Delta$ *has the form* $B_1 \Rightarrow \ldots \Rightarrow (B_k \Rightarrow y)$, *$y$ is q or is accessible to q from H*$\}$ *where* Γ_B *is*

$$\Gamma_B = \bigcup_{i=1}^{k} X_i, X_i \in \text{DAbduce}^+(\Delta - \{B\}, B_i, H * (y)).$$

13.4 Case Study: Abduction for Relevance Logic

The formal abductive machinery for relevance logic can be easily obtained from the intuitionistic case of the previous section. Intuitionistic implication becomes relevance implication when we ask that all assumptions be used. However, relevance logic is of particular interest for abduction because it is intimately involved in the second background logic (which help decide what to abduce and may involve further abductions). We shall therefore define an abduction process for relevance ⇒ directly and independently of intuitionistic logic, and add some additional frills such as cost function **C** and judgement function **J**.

A database in relevance logic will have the form for a labelled set of wffs. The labels are atomic labels a_i for different assumptions but are annotated as $+a_i$ or $-a_i$, the sign signifying whether the assumption has been used in the proof or not.

Let $\mathbf{C}(x)$ be a Canadian dollar cost function and $\mathbf{J}(x)$ be a judgement annotation function.

Definition 13.12

1. *A database has the form $\Delta = (\Delta_0, \mathbf{C}, \mathbf{J})$, where Δ_0 is a set of signed labelled formulas of the form $+a_i : A_i$, a_i are all pairwise disjoint atoms and $\mathbf{C}(a_i) = c_i$ is a cost function and $\mathbf{J}(a_i) = \alpha_i$ is a wff annotation in some other language.*

2. *We define the predicate* Success*(*$(\Delta, A, c) = 0$ *or 1 where Δ is a database, A a wff and C an amount of Canadian dollars. The predicate is also dependent on a cost policy and judgement policy which is implicit. We also keep implicit the complexity of the computation.*

   (a) *Immediate success (one step success)*
   success$(\Delta, A, c) = 1$ *immediately if $A = q$ is atomic $\pm a : q \in \Delta, \mathbf{C}(a) \leq c$ and all other wffs in Δ are signed positive, and $\mathbf{J}(a)$ is such that we are not contradicting our policy.*

   (b) *Immediate failure*
   Success$(\Delta, a, c) = 0$ *immediately if either one or more of the following holds:*

      i. *$\pm a : A_1 \Rightarrow (A_2 \Rightarrow \cdots \Rightarrow (A_n \Rightarrow q) \cdots)$ is not in Δ for any A_i. This means q is not the head of any clause in Δ.*

      ii. *Although some clauses in Δ have head q, the cost $\mathbf{C}(a)$ of the use of the clause is greater than c (the money we have got).*

      iii. *Although q is not the head of any complex clause in Δ (i.e. $A_1 \Rightarrow \cdots \Rightarrow (A_n \Rightarrow q) \cdots)$, with $n \geq 1$), q is in Δ in the form of $\pm a : q$ and is affordable $\mathbf{C}(a) \leq c$ but there are other unused clauses in Δ of the form $-b : B$, and so we cannot declare success nor can we continue with other clauses in the hope that the other unused data will be lost.*

      iv. *All the clauses with head q which are affordable have judgement value which contradict our policy for success.*

   Note that our abduction mechanism will have to address each of the above cases!

   (c) *n-stage success/failure for $\Rightarrow$*
   Success$(\Delta, A_1 \Rightarrow \cdots \Rightarrow (A_n \Rightarrow q) \cdots), c) = x$ *iff* Success$(\Delta + A_1 + \ldots + A_n, q, c') = x$ *where $x = 0$ or $x = 1$ and $\Delta + A_1 + \ldots + A_n$ is the database obtained from Δ by inserting into it the items $-a_1 : A_1, \ldots, -a_n : A_n$, where a_i are* completely new *atomic names.* $\mathbf{C}(a_i)$

and $\mathbf{J}(a_i)$ *are defined at this state at runtime according to some policy.* c' *is a new cost function again defined according to some policy (say* $c' = c + \sum_{i=1}^{n} \mathbf{C}(a_i)$*)).*

(d) Unification case for success
Success$(\Delta, q, c) = 1$, *using at most* $n+1$ *steps, if for some* $\pm a : A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q) \ldots)$ *(called the* deduction clause*) in* Δ *we have the following holding:*

   i. $\mathbf{C}(a) \leq c$ *(the clause is affordable).*

   ii. The database Δ *can be split into* n *databases (not necessarily disjoint)*[6] *such that* $\Delta = \bigcup_i \Delta_i$ *and for each* i Success$(\Delta_i', A_i, c_i) = 1$ *using at most* n *steps, where* Δ_i' *is obtained from* Δ_i *by switching the label* $\pm a$ *of the deduction clause into* $+a$ *(should it appear in* Δ_i*).*[7] *We also have* $c_1 + \ldots + c_n = c - \mathbf{C}(a)$.

(e) Unification case for failure
Success$(\Delta, q, c) = 0$*, using at most* $n+1$ *steps if for* each *candidate deduction clause* $\pm a : A_1 \Rightarrow \ldots \Rightarrow ((A_n \Rightarrow a) \ldots)$ *either of the following holds:*

   i. $\mathbf{C}(a) > c$ *(not affordable).*

   ii. for each *decomposition* $\Delta = \bigcup \Delta_i$*, as described in the preceding item (d) there exists an* $1 \leq i \leq n$ *such that* Success$(\Delta_i', A_i, c_i) = 0$ *using at most* n *steps.*

   iii. $\mathbf{J}(A)$ *does not allow for success.*

Note that to do abduction we need to address each of these cases.

The next example will show how the computation works, as well as prepare the ground for the definition of the abductive process.

Example 13.13 *Let* $\Delta = \{-a_1 : A_1, -a_2 : A_1, -a_3 : A_1 \Rightarrow A_2 \Rightarrow B)\}$. *Let the cost of each clause be \$ 10, and let us ignore* $\mathbf{J}$.
Success$(\Delta, B, \$1000)$ *if the following holds (we ignore the cost since we have lots of money!).*
$\bigvee_{i=1}^{2} \bigwedge_{j=1}^{2}$ Success$(\Delta_j^i, A_j, C_{ij})$ *where*

$$\Delta_i^1 = \{-a_i : A_1, +a_3 : A_1 \Rightarrow (A_2 \Rightarrow B\}$$
$$\Delta_1^2 = \varnothing, \Delta_2^2 = \{-a_1 : A_1, -a_2 : A_1 + a_3 : A_1 \Rightarrow (A_2 \Rightarrow B)\}$$

Both options $i = 1, i = 2$ *of division fail to yield success. The reason being that there are two copies of* A_1 *in the database and no copy of* A_2.

[6] For the case of linear logic we can give up the signed labels and have at this stage the requirement that Δ_i are disjoint and $\Delta - \{a : A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q) \ldots)\} = \bigcup_{i=1}^{n} \Delta_i$.

[7] At this point we can allow for other adjustments.

Our abduction options are to delete the redundant copy of A_1 and to add a copy of A_2.

The abduced database will be

$$\{-a_1 : A_1, -a_2 : A_2, -a_3 : A_1 \Rightarrow (A_2 \Rightarrow B)\}.$$

Let us see how we get that formally.

Option 1 for success

We need $\text{Success}(\Delta^1_1, A_1, c^1_1 = 1$ *and* $\text{Success}(\Delta^1_2, A_2, c^1_2) = 1$.

The former works. The second has $\{-a_2 : A_1, +a_3 : A_1 \Rightarrow (A_2 \Rightarrow)\}$ *and the goal is* A_2. *This is a case of immediate failure. So we take out what is not used so that we do not fail because of it and put in what we need, i.e.* $-a_4 : A_2$.

The database is therefore $\{-a_1 : A_1, -a_4 : A_2, -a_3 : A_1 \Rightarrow (A_2 \Rightarrow B)\}$.

Option 2 for success:

$\text{Success}(\Delta^2_1, A_1, c^2_1) = 1$ *and* $\text{Success}(\Delta^2_2, A_2, c^2_2) = 1$. *The first predicate requires adding* $-a_5 : A_1$. *The second predicate requires taking out the unused* $-a_1 : A_1$ *and* $-a_2 : A_1$ *and adding* $-a_6 : A_2$.

So the abduced database is

$$\{-a_5 : A_1, -a_6 : A_2, -a_3 : A_1 \Rightarrow (A_2 \Rightarrow B)\}.$$

Obviously the two options are the same but for the different use of atomic names.

Definition 13.14 (The abduce function) *The function* $\mathbf{Ab}^+(\Delta, A, c)$ *tells us what to add to the database to make* A *succeed from* Δ.

We will ignore financial consideration since the obvious solution to that is to add more money. Other restrictions may be computational time but then this is so important that it needs special attention. We will deal with time separately.

$\mathbf{Ab}^+(\Delta, A)$ *is a set of alternative update actions, each of which changes* Δ *to a new database* Δ' *which proves* A. *We write it as*

$$\mathbf{Ab}^+(\Delta, A) = \{\Delta'_i\}.$$

The definition of $\mathbf{Ab}^+$ *is recursive on the computation stages of* $\text{Success}(\Delta, A) = 0$.

For notational convenience we let $\mathbf{Ab}^+(\Delta, A) = \{\Delta\}$, *when* $\Delta \vdash A$, *i.e. when* $\text{Success}(\Delta, A) = 1$.

1. $\mathbf{Ab}^+(\Delta, q)$ *for* q *atomic and immediate failure contains* Δ', *where* Δ' *is obtained from* Δ *by deleting all unused items and adding* $-a : q$, *for a new atomic name* a.

2. *Consider* $\mathrm{Success}(\Delta_1(A_1 \Rightarrow \ldots (A_n \Rightarrow B)\ldots))$. *This fails iff Success*$(\Delta + A_1 + \cdots + A_n), B)$ *fails.*

   $\mathbf{Ab}^+(\Delta + A_1, + \cdots + A_n, B)$ *gives us options* $\Delta'_1, \ldots, \Delta'_m$ *of what to add to* $\Delta + A_1 + \cdots + A_n$ *to make* B *provable. Let* $\Delta_i = \{-a^i_j : X_{i,j}\}$.

   Then let $A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow \Delta_i)\ldots)$ *be the theory*

$$\Delta^*_i = \{-a^i_j : A_1 \Rightarrow \ldots (A_n \Rightarrow X_{i,j})\ldots)\}.$$

   Then we define

$$\begin{aligned}\mathbf{Ab}^+(\Delta, A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow B)\ldots) = \{&A_1 \Rightarrow \ldots (A_n \Rightarrow \Gamma)\ldots) \\ &\mid \Gamma \in \mathbf{Ab}^+(\Delta + A_1 \ldots + A_n, B)\}\end{aligned} \tag{13.1}$$

3. *Assume* $\mathrm{Success}(\Delta, q) = 0$ *because for every candidate deduction clause* $Z = \pm a : A_1 \Rightarrow \ldots \Rightarrow (A_n \Rightarrow q)\ldots)$ *in* Δ *and any division of* $\Delta = \bigcup_i + \Delta^Z_i$ *there exists some* i *s.t.* $\mathrm{Success}(\Delta_i, A_i) = 0$. *Let* $\mathbf{Ab}^+(\Delta_i, A_i)$ *be the abduced family for this case and recall that we can let* $\mathbf{Ab}^+(\Delta_i, A_i) = \{\Delta_i\}$ *if* $\Delta_i \vdash A$.

   We let $\mathbf{Ab}^+(\Delta, q)$ *be* $\bigcup_{\textit{all clauses } Z \textit{ and all divisions}} \bigcup_i \mathbf{Ab}^+(\Delta_i, A_i)$.

Remark 13.15 *The role of the background logic is to decide which option* $\Delta' \in \mathbf{Ab}^+(\Delta, A)$ *to adopt as the abduced theory* Δ.

Theorem 13.16 (Soundness of abduction) *Let* $\Delta' \in \mathbf{Ab}^+(\Delta, A)$ *then* Success $(\Delta', A) = 1$.

Proof. By induction on the recursive definition of $\mathbf{Ab}^+$. ■

13.5 Conclusion

This section concludes the formal part of the book. We hope the reader gained an idea about how the formal models would look. We repeat our observation that proper formal modelling is better done at the end of this series of volumes since all the concepts and mechanisms are interdependent. In fact they will be recursively interdependent in the formal model we hope to present in the last volume.

This Page is Intentionally Left Blank

Bibliography

[Achinstein, 1977a] Peter Achinstein. Discussion. In Frederick Suppe, editor, *The Structure of Scientific Theories*, page 364. Urbana: University of Illinois Press, 2nd edition, 1977.

[Achinstein, 1977b] Peter Achinstein. History and the philosophy of science: A reply to Cohen. In Frederick Suppe, editor, *The Structure of Scientific Theories*, pages 350–360. Urbana: University of Illinois Press, 2nd edition, 1977.

[Achinstein, 1991] P. Achinstein. *Particles and Waves*. Oxford: Oxford University Press, 1991.

[Aizawa, 1994] K. Aizawa. Representations without rules, connectionism and the syntactic argument. *Synthese*, 101:465–492, 1994.

[Aizawa, 2000] K. Aizawa. Connectionist rules: A rejoinder to Horgan and Tienson's *Connectionism and the Philosophy of Psychology*. *Acta Analytica*, 22:59–85, 2000.

[Aliseda-LLera, 1997] Atocha Aliseda-LLera. *Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence*. Amsterdam: Institute for Logic, Language and Computation, 1997. PhD dissertation (ILLC Dissertation Series 1997-4).

[Aliseda, forthcoming] Atocha Aliseda. *Abductive Reasoning: Logical Investigation into the Processes of Discovery and Evaluation*. Dordrecht: Kluwer, forthcoming.

[Allen, 1994] R.J. Allen. Factual ambiguity and a theory of evidence. *Northwestern University Law Review*, 88:604–660, 1994.

[Amundson and Lauder, 1994] R. Amundson and G.V. Lauder. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy*, 9:443–469, 1994.

[Anderson and Belnap Jr., 1959] A.R. Anderson and Nuel Belnap Jr. A simple treatment of truth functions. *The Journal of Symbolic Logic*, 24:301–302, 1959.

[Anderson and Belnap, 1975] Alan Ross Anderson and Nuel D. Belnap, Jr. *Entailment: The Logic of Relevance and Necessity*, volume 1. Princeton, NJ: Princeton University Press, 1975.

[Aravurdan and Dung, 1994] C. Aravurdan and P.M. Dung. Belief dynamics, abduction and databases. In C. MacNish, D. Pearch, and L.M. Pereira, editors, *Logics in Artificial Intelligence*, pages 66–85. Berlin: Springer-Verlag, 1994.

[Armour-Garb, 2004] Brad Armour-Garb. Wrestling with (and without) dialetheism. *Australasian Journal of Philosophy*, 2004. To appear.

[Arnauld and Nicole, 1996] Antoine Arnauld and Pierre Nicole. *Logic or the Art of Thinking*. Cambridge: Cambridge University Press, 1996. Originally published in 1662. This volume is edited by Jill Vance Buroker.

[Arruda, 1989] A.I. Arruda. Aspects of the historical development of paraconsistent logic. In Graham Priest, Richard Routley, and Jean Norman, editors, *Paraconsistent Logic: Essays on the Inconsistent*, pages 99–130. Munich: Philosophia Verlag, 1989.

[Axsom *et al.*, 1987] D.S. Axsom, S. Yates, and S. Chaiken. Audience response as a heuristic case in persuasion. *Journal of Personality and Social Psychology*, 53:30–40, 1987.

[Bacon, 1905] Francis Bacon. Novum organum. In R.L. Ellis and J. Spedding, editors, *The Philosophical Works of Francis Bacon*, pages 212–387. London: Routledge, 1905.

[Bartha, forthcoming] Paul Bartha. Analogical reasoning and plausibility in the sciences. Preprint, forthcoming.

[Barwise and Seligman, 1997] John Barwise and J. Seligman. *Information Flow: The Logic of Distributed Systems*. Cambridge and New York: Cambridge University Press, 1997.

[Batens, 1980] D. Batens. Paraconsistent extensional propositional logic. *Logique et Analyse*, 23:195–234, 1980.

[Baum, 1975] Robert Baum. *Logic*. New York: Holt Rinehart, 1975.

[Beth, 1969] E.W. Beth. Semantic entailment and formal derivability. In J. Hintikka, editor, *The Philosophy of Mathematics*, pages 9–41. Oxford: Oxford University Press, 1969.

[Bigelow and Pargetter, 1987] J. Bigelow and R. Pargetter. Functions. *The Journal of Philosophy*, 84:181–196, 1987.

[Bikhchandani *et al.*, 1992] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as information cascades. *Journal of Political Economy*, 100:992–1026, 1992.

[Bikhchandani *et al.*, 1998] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. Learning from the behavior of others: Confirmity, fads, and informational cascades. *Journal of Economic Perspectives*, 12:151–170, 1998.

[Blois, 1984] M.S. Blois. *Information and Medicine: The Nature of Medical Description*. New York: Basic Books, 1984.

[Boutilier and Becher, 1995] C. Boutilier and V. Becher. Abduction as belief revision. *Artificial Intelligence*, 77:43–94, 1995.

[Boyd, 1979] R. Boyd. Metaphor and thought. In A. Ortony, editor, *Metaphor and Theory change*, pages 356–408. Cambridge University Press, Cambridge, 1979.

[Brandon, 1990] R.N. Brandon. *Adaptation and Environment.* Princeton, NJ: Princeton University Press, 1990.

[Brockhaus, 1991] Richard R. Brockhaus. Realism and psychologism in 19th Century logic. *Philosophy and Phenomenological Research*, LI:493–524, 1991.

[Bruner, 1957] J.S. Bruner. On perceptual readiness. *Psychological Review*, 64:123–152, 1957.

[Bruza and Song, 2002] P.D. Bruza and D. Song. Inferring query models by computing information flow. In *Proceedings of ACM/CIKM 2002*, pages 260–269, 2002.

[Bruza *et al.*, 2004] P.D. Bruza, D. Song, and R.M. McArthur. Abduction in semantic space: Towards a logic of discovery. *Logic Journal of the IGPL*, 12:97–110, 2004.

[Buchler, 1955] Justus Buchler. *Philosophical Writings of Peirce.* New York: Dover, 1955.

[Burgess *et al.*, 1998] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences, discoveries. *Discourse Processes*, 25:211–257, 1998.

[Burks, 1946] Arthur W. Burks. Empiricism and vagueness. *Journal of Philosophy*, 43:447–486, 1946.

[Burton, 1999] Robert G. Burton. A neurocomputational approach to abduction. *Mind*, 9:257–265, 1999.

[Byrne, 1968] Edmund F. Byrne. *Probability and Opinion: A study in the medieval presuppositions of post-medieval theories of probability.* The Hague: Martinus Nijhoff, 1968.

[Canfield, 1964] John Canfield. Teleological explanation in biology. *British Journal for the Philosophy of Science*, 14:285–295, 1964.

[Carlson and Pelletier, 1995] Gregory N. Carlson and Francis Jeffry Pelletier. *The Generic Book.* Chicago: Chicago University Press, 1995.

[Carnap, 1950] Rudolf Carnap. *The Logical Foundations of Probability.* Chicago, IL: University of Chicago Press, 1950.

[Cartwright, 1983] Nancy Cartwright. How approximations take us away from theory and towards the truth. *Pacific Philosophical Quarterly*, 64:273–280, 1983.

[Catanzarite and Greenburg, 1979] V. Catanzarite and A. Greenburg. Neurologist — a computer program for diagnosis in neurology. *Proceedings of the 3rd Symposium of Computer Applied Medical Care, IEEE*, pages 64–72, 1979.

[Cherry, 1966] Colin Cherry. *On Human Communication.* Cambridge, MA: MIT Press, 1966.

[Chomsky, 1972] Noam Chomsky. *Language and Mind.* New York: Harcourt Brace Jovanovich, enlarged edition, 1972.

[Churchland, 1989] Paul Churchland. *A Neurocomputational Perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT Press, 1989.

[Churchland, 1995] Paul M. Churchland. *The Engine of Reason, The Seat of the Soul.* Cambridge: The MIT Press, 1995.

[Cohen, 1977] L. Jonathan Cohen. *The Probably and the Provable.* Oxford: Clarendon Press, 1977.

[Cohen, 1979] L. Jonathan Cohen. Rescher's theory of plausible reasoning. In Ernest Sosa, editor, *The Philosophy of Nicholas Rescher: Discussion and Replies*, pages 49–60. Princeton, NJ: Princeton University Press, 1979.

[Cohen, 1992] L. Jonathan Cohen. *An Essay on Belief and Acceptance.* Oxford: Clarendon Press, 1992.

[Collingwood, 1946] R.G. Collingwood. *The Idea of History.* Oxford: Oxford University Press, 1946.

[Coombs and Hartley, 1987] M. Coombs and R. Hartley. The mgr algorithm and its application to the generation of explanations for novel events. Technical Report MCCS-87-97, Computer Research Laboratory, New Mexico University, 1987.

[Cooper, 1990] G. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

[Cornman, 1980] James W. Cornman. *Skepticism, Justification, and Explanation.* Dordrecht: Reidel, 1980.

[Crombie, 1997] E. James Crombie. What is deduction? In Nathan Houser, Don D. Roberts, and James Van Evra, editors, *Studies in the Logic of Charles Sanders Peirce*, pages 460–476. Bloomington, IN: Indianna University Press, 1997.

[Cross and Harris, 1991] R. Cross and J. W. Harris. *Precedent in English Law.* Oxford University Press, Oxford, 1991.

[Cross and Wilkins, 1964] Rupert Cross and Nancy Wilkins. *An Outline of the Law of Evidence.* London: Butterworths, 1964.

[Cummins, 1975] R. Cummins. Functional analysis. *The Journal of Philosophy*, 72:741–765, 1975.

[Cuppens and Demolombe, 1988] F. Cuppens and R. Demolombe. Cooperative answering: a methodology to provide intellgient access to databases. In *Proceedings of 2nd International Conference on Expert Database Systems.* Tysons Corner, VA, 1988.

[Cuppens and Demolombe, 1989] F. Cuppens and R. Demolombe. How to recognize interesting topics to provide cooperative answering. *Information Systems*, 14:163–173, 1989.

[DaCosta and Bueno, 1996] N. DaCosta and O. Bueno. Consistency, paraconsistency and truth. *Ideas y Valores*, 100:48–60, 1996.

[Darden and Cain, 1989] L. Darden and J. A. Cain. Selection type theories. *Philosophy of Science*, 56:106–129, 1989.

[Darden, 1974] Lindley Darden. *Reasoning in Scientific Change: The Field of Genetics at its Beginnings*. PhD thesis, University of Chicago, 1974.

[Darden, 1976] Lindley Darden. Reasoning in scientific change: Charles Darwin, Hugo de Vries, and the discovery of segregation. *Studies in the History and Philosophy of Science*, 7:127–169, 1976.

[Darden, 1991] Lindley Darden. *Theory Change in Science Strategies from Mendelian Genetics*. New York, Oxford: Oxford University Press, 1991.

[Daston, 1988] Lorraine Daston. *Classical Probability in the Enlightenment*. Princeton, NJ: Princeton University Press, 1988.

[Davidson and Hintikka, 1969] Donald Davidson and Jaakko Hintikka. *Words and Objections: Essays on the Work of W.V. Quine*. Amsterdam: Kluwer, 1969.

[Davidson, 1984] Donald Davidson. *Inquiries into Truth and Interpretation*. New York: Oxford University Press, 1984.

[d'Avila Garcez and Lamb, 2004] A.S. d'Avila Garcez and L.C. Lamb. Reasoning about time and knowledge in neural-symbolic learning systems. In S. Thrum and B. Schoelkopk, editors, *Advances in Neural Information Processing Systems 16: Proceedings of the NIPS 2003 Conference*, Vancouver, BC, 2004. Cambridge, MA: MIT Press.

[d'Avila Garcez *et al.*, 2002] A.S. d'Avila Garcez, K. Broda, and Dov M. Gabbay. *Neural-Symbolic Learning Systems: Foundations and Applications*. Berlin: Springer-Verlag, 2002.

[de Kleer, 1986] Johan de Kleer. An assumption-based tms. *Artificial Intelligence*, 28:127–162, 1986.

[Demolombe and Jones, 1999] R. Demolombe and A.J.I. Jones. Sentences of the kind 'sentence *p* is about topic *t*'. In H. J. Ohlbach and U. Reyle, editors, *Logic, Language and Reasoning*, pages 115–133. Dordrecht and Boston: Kluwer, 1999.

[Dennis, 1999] I.H. Dennis. *The Law of Evidence*. London: Sweet & Maxwell, 1999.

[Detlefsen, 1986] M. Detlefsen. *Hilbert's Program: An Essay on Mathematical Instrumentalism*. Dordrecht and Boston: Reidel, 1986.

[Doyle, 1979] Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 7:231–272, 1979.

[Dray, 1957] William Dray. *Laws and Explanation in History*. Oxford: Oxford University Press, 1957.

[Duhem, 1904–1905] Pierre Duhem. *The Aim and Structure of Physical Theory*. London: Atheneum 1962, 1904–1905. Philip P. Wiener, translator.

[Dunn, 1994] J.M. Dunn. Relevant logic and entailment. In Dov M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume III, pages 117–224. Dordrecht: Reidel, 1994. Originally published in 1986.

[Dvorak and Kuipers, 1989] D. Dvorak and B. Kuipers. Model-based monitoring of dynamic systems. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1238–1243. Detroit, MI: Morgan Kaufmann, 1989.

[Eagly and Chaiken, 1993] A.H. Eagly and S. Chaiken. *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich, 1993.

[Eisele, 1985] Carolyn Eisele. *Historical Perspectives on Peirce's Logic of Science*. The Hague: Mouton, 1985. In 2 volumes.

[Eliasmith and Thagard, 1997] C. Eliasmith and Paul Thagard. Waves, particles, and explanatory coherence. *British Journal for the Philosophy of Science*, 48:1–19, 1997.

[Eshghi and Kowalski, 1990] K. Eshghi and R. A. Kowalski. Abduction compared with negation by failure. In *Proceedings of Logic Colloquium 1990*. Springer, 1990.

[Fann, 1970] K.T. Fann. *Peirce's Theory of Abduction*. The Hague: Nijhoff, 1970.

[Finger and Wasserman, to appeara] Finger and Wasserman. Approximate and limited reasoning: Semantics, proof theory, expressivity and control. *Journal of Logic and Computation*, to appear.

[Finger and Wasserman, to appearb] Marcelo Finger and Renata Wasserman. Logics for approximate reasoning: Approximating classical logic 'from above'. *Journal of Logic and Computation*, to appear.

[Flach and Kakas, 2000] Peter A. Flach and Antonis C. Kakas, editors. *Abduction and Induction: Essays on Their Relation and Integration*. Dordrecht and Boston: Kluwer, 2000.

[Fleck, 1996] J. Fleck. Informal information flow and the nature of expertise in financial service. *International Journal of Technology Management*, 11:104–128, 1996.

[Fodor, 1981] Jerry A. Fodor. *Representations: Philosophical Essays on Foundations of Cognitive Science*. Cambridge, MA: MIT Press, 1981.

[Fodor, 1983] Jerry A. Fodor. *The Modularity of Mind*. Cambridge, MA: MIT Press, 1983.

[Franklin, 2001] James Franklin. *The Science of Conjecture: Evidence and Probability before Pascal*. Baltimore, MD: The Johns Hopkins University Press, 2001.

[Frege, 1914] Gottlob Frege. Logic in mathematics. In Hans Hermes, Friederich Kambartel, and Friederich Kaulbach, editors, *Posthumous Writings*, pages 203–250. Chicago, IL: University of Chicago Press, 1914.

[French and Ladyman, 1997] S. French and J. Ladyman. Superconductivity and structures: revisiting the London Approach. *Studies in History and Philosophy of Modern Physics*, 28:363–393, 1997.

[Frey, 1998] B.J. Frey. *Graphical Models for Machine Learning and Digital Communication.* Cambridge, MA: MIT Press, 1998.

[Friedman and Simpson, 2000] Harvey Friedman and S. Simpson. Issues and problems in reverse mathematics. *Computability Theory and its Applications: Contemporary Mathematics*, 257:127–144, 2000.

[Gabbay and Hunter, 1991] Dov M. Gabbay and Anthony Hunter. Making inconsistency respectable, part i. In P. Jourand and J. Kelemen, editors, *Fundamentals of Artificial Intelligence Research*, volume 535, pages 19–32. Berlin: Springer-Verlag, 1991.

[Gabbay and Hunter, 1993] Dov M. Gabbay and A. Hunter. Making inconsistency respectable part 2: Meta-level handling of inconsistency. In *LNCS 747*, pages 129–136. Berlin: Springer-Verlag, 1993.

[Gabbay and Kempson, 1991] D. M. Gabbay and R. Kempson. Labelled abduction and relevance, 1991. unpublished.

[Gabbay and Olivetti, 2000] Dov M. Gabbay and N. Olivetti. *Goal Directed Proof Theory.* Kluwer, 2000.

[Gabbay and Reyle, 1984] D. M. Gabbay and U. Reyle. N-prolog, an extension of prolog with hypothetical implication. *Journal of Logic Programming*, 4:319–355, 1984.

[Gabbay and Woods, 1999] Dov M. Gabbay and John Woods. Cooperate with your logic ancestors. *Journal of Logic, Language and Information*, 8, 1999.

[Gabbay and Woods, 2000] Dov M. Gabbay and John Woods. Editorial. *Journal of Logic and Computation*, 10(1):1–2, 2000.

[Gabbay and Woods, 2001a] Dov M. Gabbay and John Woods. More on noncooperation in dialogue logic. *Logic Journal of the IGPL*, 9:321–339, 2001.

[Gabbay and Woods, 2001b] Dov M. Gabbay and John Woods. The new logic. *Logic Journal of the IGPL*, 9:157–190, 2001.

[Gabbay and Woods, 2001c] Dov M. Gabbay and John Woods. Non-cooperation in dialogue logic. *Synthese*, 127:161–186, 2001.

[Gabbay and Woods, 2002] Dov M. Gabbay and John Woods. Formal approaches to practical reasoning: A survey. In Dov M. Gabbay, Ralph H. Johnson, Hans Jürgen Ohlbach, and John Woods, editors, *Handbook of the Logic of Argument and Inference: The Turn Towards the Practical*, volume 1 of *Studies in Logic and Practical Reasoning*, pages 445–478. Amsterdam: North-Holland, 2002.

[Gabbay and Woods, 2003a] Dov M. Gabbay and John Woods. *Agenda Relevance: A Study in Formal Pragmatics*, volume 1 of *A Practical Logic of Cognitive Systems*. Amsterdam: North-Holland, 2003.

[Gabbay and Woods, 2003b] Dov M. Gabbay and John Woods. Normative models of rationality. *Logic Journal of IGPL*, 11:597–613, 2003.

[Gabbay and Woods, 2004] Dov M. Gabbay and John Woods. The practical turn in logic, volume 13. In Dov M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, pages 15–130. Dordrecht and Boston: Kluwer, 2nd revised edition, 2004.

[Gabbay and Woods, 2005] Dov M. Gabbay and John Woods. *Formal Models of Practical Reasoning*. 2005. In preparation.

[Gabbay *et al.*, 2002] Dov M. Gabbay, Odinaldo Rodriguez, and John Woods. Belief contraction, anti-formulae and resource overdraft part 1 deletion in resource bounded logics. *Logic Journal of the IGPL*, 10:601–652, 2002.

[Gabbay *et al.*, 2003] Dov M. Gabbay, Gabriella Pigozzi, and John Woods. Controlled revision. *Journal of Logic and Computation*, 13:15–35, 2003.

[Gabbay *et al.*, 2004] Dov M. Gabbay, Odinaldo Rodriguez, and John Woods. Belief contraction, anti-formulae and resource overdraft part 2 deletion in resource unbounded logics. In Shahid Rahman, John Symosn, Dov M. Gabbay, and Jean paul van Bendegem, editors, *Logic Epistemology and the Unity of Science*, pages 291–326. Kluwer, 2004.

[Gabbay, 1985a] Dov M. Gabbay. N-prolog, part ii. *Journal of Logic Programming*, 5:251–283, 1985.

[Gabbay, 1985b] Dov M. Gabbay. Theoretical foundations for non-monotonic reasoning. In Karel Apt, editor, *Logics and Models of Concurrent Systems*, pages 439–457. Berlin: Springer-Verlog, 1985.

[Gabbay, 1996] Dov M. Gabbay. *Labelled Deductive Systems*. Oxford: Oxford University Press, 1996.

[Gabbay, 1998a] Dov M. Gabbay. *Elementary Logic: A Procedural Perspective*. Upper Saddle River, NJ: Prentice Hall, 1998.

[Gabbay, 1998b] Dov M. Gabbay. *Fibring Logics*. Oxford: Oxford University Press, 1998. Vol. 38 of *Oxford Logic Guides*.

[Gabbay, 2000] Dov M. Gabbay. Abduction in labelled deductive systems. In *Handbook of Defeasible Reasoning and Uncertainty Management, Volume 4*, pages 99–154. Dordrecht: Kluwer Academic Publishers, 2000. Volume edited by D.M. Gabbay and R. Kruse.

[Gabbay, 2001a] Dov M. Gabbay. Dynamics of practical reasoning: A position paper. In M. Zakharyaschev, K. Segerberg, M. de Rijke, and H. Wansing, editors, *Advances in Modal Logic*, volume 2, pages 179–224. CSLI Publications, 2001.

[Gabbay, 2001b] Dov M. Gabbay. What is a logical system 2. In *Logical Consequence: Rival Approaches*, pages 81–104. Oxford: Hermes Science Publications, 2001. Proceedings of the 1999 Conference of the SEP. Vol 1.

[Gadamer, 1975] Hans Georg Gadamer. *Truth and Method.* New York: Continuum, 1975.

[Gallie, 1964] W.B. Gallie. *Philosophy and the Historical Understanding.* London: Chatto and Windus, 1964.

[Gärdenfors, 1988] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* Cambridge Massachusetts:Bradford Books, The MIT Press, 1988.

[Gärdenfors, 2000] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought.* Cambridge, MA: MIT Press, 2000.

[Geffner, 1992] Hector Geffner. *Default Reasoning: Causal and Conditional Theories.* Cambridge, MA: MIT Press, 1992.

[Genesereth and Nilsson, 1987] Michael R. Genesereth and Nils J. Nilsson. *Logical Foundations of Artificial Intelligence.* Palo Alto, CA: Morgan Kaufmann, 1987.

[Gentner, 1981] D. Gentner. Studies of inerence from lack of knowledge. *Memory and Cognition*, 9:434–443, 1981.

[Gentner, 1983] D. Gentner. Structure mapping: a theoretical framework for analogy. *Cognitive Science*, 7:155–170, 1983.

[Gentzen, 1935] Gerhard Gentzen. Untersuchungen uberdas logische schliessen. *Mathematisches, Zeitschrift*, 39:176–210, 149–167, 1935.

[Gervás, 1995] P. Gervás. *Logical Considerations in the Interpretation of Presuppositional Sentences.* PhD thesis, Department of Computing, Imperial College, London, 1995.

[Gigerenzer and Selten, 2001] G. Gigerenzer and R. Selten, editors. *Bounded Rationality: The Adaptive Toolbox.* Cambridge, MA: MIT Press, 2001.

[Ginsberg, 1988] A. Ginsberg. Theory of revision via prior operationalization. In Jeff Shrager and Pat Langley, editors, *Computational Models of Scientific Discovery and Theory Formation.* San Mateo, CA: Morgan Kaufmann Publishers, 1988.

[Gochet, 2002] P. Gochet. The dynamic turn in Twentieth Century logic. *Synthese*, 130:175–184, 2002.

[Goddu, 2002] G.C. Goddu. The most important and fundamental distinction in logic. *Informal Logic*, 22:1–17, 2002.

[Gödel, 1944] Kurt Gödel. Russell's mathematical logic. In *The Philosophy of Bertrand Russell*, pages 123–153. New York: Tudor, 3rd edition, 1944.

[Gödel, 1990a] Kurt Gödel. Remarks before the Princeton bicentennial conference on problems in mathematics. In Solomon Feferman, John W. Dawson Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay, and Jean van Heijenoort, editors, *Kurt Gödel Collected Works, Volume II, Publications 1938–1974*, pages 150–153. New York and Oxford: Oxford University Press, 1990. Originally published in 1946.

[Gödel, 1990b] Kurt Gödel. What is Cantor's continuum problem? In Solomon Feferman, John W. Dawson Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay, and Jean van Heijenoort, editors, *Kurt Gödel Collected Works, Volume II, Publications 1938–1974*, pages 176–187. New York and Oxford: Oxford University Press, 1990. Originally published in 1947.

[Goldman, 1986] Alvin I. Goldman. *Epistemology and Cognition.* Cambridge, MA: Harvard University Press, 1986.

[Good, 1983] I.J. Good. *Good Thinking: The Foundations of Probability and its Application.* Minneapolis, MN: University of Minnesota Press, 1983.

[Govier, 1988] Trudy Govier. *A Practical Study of Argument.* Belmont, CA: Wadsworth, 1988.

[Gray, 2002] John Gray. *Straw Dogs: Thoughts on Humans and Other Animals.* London: Granta, 2002.

[Greene, 2004] Brian Greene. *The Fabric of the Cosmos: Space, Time, and the Texture of Reality.* New York: Knopf, 2004.

[Gregory, 1970] R. Gregory. *The Intelligent Eye.* New York: McGraw Hill, 1970.

[Gregory, 1980] R.L. Gregory. Perception as hypotheses. In R.L. Gregory, editor, *Proceedings of the Royal Socity of London, Volume B290*, pages 181–197. New York: Oxford University Press, 1980.

[Grice, 1989] H.P. Grice. *Studies in the way of the world.* Harvard University Press, Cambridge, MA, 1989.

[Guarini, 2001] Marcello Guarini. A defence of connectionism against the "SYNTACTIC" argument. *Synthese*, 128:287–317, 2001.

[Hacking, 1975] Ian Hacking. *The Emergence of Probability.* London: Cambridge University Press, 1975.

[Hacking, 1983] Ian Hacking. *Representing and Intervening: Introductory topics in the Philosophy of Natural Science.* Cambridge: Cambridge University Press, 1983.

[Hahn and Schilpp, 1998] L.E. Hahn and P.A. Schilpp, editors. *The Philosophy of W.V. Quine.* Chicago and La Salle, IL: Open Court, expanded edition, 1998.

[Hailpern, 2003] Joseph Y. Hailpern. *Reasoning About Uncertainty.* Cambridge, MA: MIT Press, 2003.

[Hamlyn, 1990] D.W. Hamlyn. *In and Out of the Black Box.* Oxford: Basil Blackwell, 1990.

[Hannibal, 2002] Martin Hannibal. *The Law of Criminal and Civil Evidence.* London: Longman, 2002.

[Hansen and Kauffeld, 2005] H.V. Hansen and F. Kauffeld, editors. *Presumptions and Burdens of Proof: An Anthology.* Tuscaloosa, AB: University of Alabama, 2005.

[Hanson, 1958] N.R. Hanson. *Patterns of Discovery.* Cambridge: Cambridge University Press, 1958.

[Hanson, 1961] N.R. Hanson. Is there a logic of scientific discovery? In Herbert Feigl and Grover Maxwell, editors, *Current Issues in the Philosophy of Science*, pages 20–35. New York: Holt, Rinehart and Winston, 1961.

[Harman, 1965] Gilbert H. Harman. The inference to the best explanation. *Philosophical Review*, 74:88–95, 1965.

[Harman, 1986] Gilbert Harman. *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press, 1986.

[Heim, 1991] I. Heim. On the projection problem for presuppositions. In S. Davis, editor, *Pragmatics*, pages 397–405. Oxford: Oxford University Press, 1991.

[Henkin, 1950] Leon Henkin. Completeness in the theory of types. *Journal of Symbolic Logic*, 15:81–91, 1950.

[Hesse, 1966] Mary Hesse. *Models and Analogies in Science*. Notre Dame, IN: University of Notre Dame Press, 1966.

[Hesse, 1974] M. Hesse. *The Structure of Scientific Inference*. Berkeley: University of California, 1974.

[Hilpinen, 2004] Risto Hilpinen. Peirce's logic. In Dov M. Gabbay and John Woods, editors, *The Rise of Modern Logic: From Leibniz to Frege*, volume 3 of *Handbook of the History of Logic*, pages 611–658. Dordrecht: North-Holland, 2004.

[Hintikka and Bachman, 1991] Jaakko Hintikka and James Bachman. *What if ...? Toward Excellence in Reasoning*. Mountain View, CA: Mayfield, 1991.

[Hintikka and Halonen, 1995] Jaakko Hintikka and Ilpo Halonen. Semantics and pragmatics for why-questions. *Journal of Philosophy*, 92:636–657, 1995.

[Hintikka and Remes, 1974] Jaakko Hintikka and U. Remes. *The Method of Analysis: Its Geometrical Origin and its General Ssignificance*. Dordrecht: Reidel, 1974.

[Hintikka *et al.*, 2002] Jaakko Hintikka, Ilpo Halonen, and Arto Mutanen. Interrogative logic as a general theory of reasoning. In Dov M. Gabbay, Ralph H. Johnson, Hans Jürgen Ohlbach, and John Woods, editors, *Handbook of the Logic of Argument and Inference: The Turn Towards the Practical*, volume 1 of *Studies in Logic and Practical Reasoning*, pages 295–337. Amsterdam: North-Holland, 2002.

[Hintikka, 1955] Jaakko Hintikka. *Two Papers on Symbolic Logic*. Helsinki: Akateeminen Kirjak, 1955.

[Hintikka, 1999a] Jaakko Hintikka. *Selected Papers*, volume 5. Dordrecht and Boston: Kluwer, 1999.

[Hintikka, 1999b] Jaakko Hintikka. What is abduction? In Jaakko Hintikka, editor, *Selected Papers*, volume 5, pages 91–113. Dordrecht and Boston: Kluwer, 1999. Originally published in the *Transactions of the Charles S. Peirce Society*, 34 (1998); 503–533.

[Hitchcock, 1983] David Hitchcock. *Critical Thinking*. New York: Methuen, 1983.

[Hitchcock, 2002] David Hitchcock. A note on implicit premisses. *Informal Logic*, 22:158–159, 2002.

[Hobbs *et al.*, 1990] J.R. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation on abduction. Technical Report 499, SRI International, Artificial Intelligence Center, Computing and Engineering Sciences Division, Menlo Park, CA, 1990.

[Holyoak and Thagard, 1995] Keith J. Holyoak and Paul Thagard. *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press, 1995.

[Horan, 1989] B.L. Horan. Functional explanations in sociobiology. *Biology and Philosophy*, 4:131–158 and 205–228, 1989.

[Horgan and Tienson, 1988] T. Horgan and J. Tienson. Settling into a new paradigm. *Southern Journal of Philosophy*, 26:97–113, 1988. *Connectionism and the Philosophy of Mind: Proceedings of the 1987 Spindel Conference*, special supplement.

[Horgan and Tienson, 1989] T. Horgan and J. Tienson. Representations without rules. *Philosophical Topics*, 17:147–174, 1989.

[Horgan and Tienson, 1990] T. Horgan and J. Tienson. Soft laws. In Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein, editors, *The Philosophy of the Human Sciences*, volume 15 of *Midwest Studies in Philosophy*, pages 256–279. Notre Dame, IN: University of Notre Dame Press, 1990.

[Horgan and Tienson, 1992] T. Horgan and J. Tienson. Cognitive systems as dynamical systems. *Topoi*, 11:27–43, 1992.

[Horgan and Tienson, 1996] T. Horgan and J. Tienson. *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press, A Bradford Book, 1996.

[Horgan and Tienson, 1999a] T. Horgan and J. Tienson. Authors' replies. *Acta Analytica*, 22:275–287, 1999.

[Horgan and Tienson, 1999b] T. Horgan and J. Tienson. Short précis of *Connectionism and the Philosoph of Psychology*. *Acta Analytica*, 22:9–21, 1999.

[Horwich, 1982] P. Horwich. *Probability and Evidence*. Cambridge: Cambridge University Press, 1982.

[Houser *et al.*, 1997] Nathan Houser, Don D. Roberts, and James van Evra. *Studies in the Logic of Charles Sanders Peirce*. Bloomington and Indianapolis, IN: Indiana University Press, 1997.

[Howells, 1996] J. Howells. Tacit knowledge, innovation and technology transfer. *Technology, Analysis and Strategic Management*, 8:91–105, 1996.

[Howson and Urbach, 1993] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Tradition*. Lasalle, IL: Open Court, second edition, 1993.

[Howson, 2000] C. Howson. The logic of bayesian probability. In D. Corfield and J. Williamson, editors, *Foundations of Bayesianism*. Dordrecht and Boston: Kluwer, 2000.

[Hrycej, 1990] T. Hrycej. Gibbs sampling in Bayesian networks. *Artificial Intelligence*, 46:351–363, 1990.

[Hutchins, 1995] E. Hutchins. *Cognition in the Wild.* Cambridge, MA: MIT Press, 1995.

[Irvine, 1989] A.D. Irvine. Epistemic Logicism and Russell's Regressive Method. *Philosophical Studies*, 55:303–327, 1989.

[Jackson, 1996] Sally Jackson. Fallacies and heuristics. In Johan van Benthem, Frans H. van Eemeren, Rob Grootendorst, and Frank Veltman, editors, *Logic and Argumentation*, pages 101–114. Amsterdam: North-Holland, 1996.

[Jacobs and Jackson, 1983] Scott Jacobs and Sally Jackson. Speech act structure in conversation: Rational aspects of pragmatic coherence. In Rober T. Craig and Karen Tracy, editors, *Conversational Coherence: Form, Structure, and Strategy*, pages 47–66. Newbury Park, CA: Sage, 1983.

[Jacobs *et al.*, 1985] Scott Jacobs, M. Allen, Sally Jackson, and D. Petrel. Can ordinary arguers recognize a valid conclusion if it walks up and bites them in the butt? In J.R. Cox, M.O. Sillars, and G.B. Walker, editors, *Argument and Social Practice: Proceedings of the Fourth SCA/FA Conference on Argumentation*, pages 665–674. Annandale, VA: Speech Communication Association, 1985.

[Jacquette, 1986] Dale Jacquette. Intentionality and intensionality: Quotation contexts and the modal wedge. *The Monist*, 69:598–608, 1986.

[Jacquette, 2003a] Dale Jacquette, editor. *Philosophy, Psychology and Psychologism: Critical and Historical Readings on the Psychological Turn in Philosophy.* Dordrecht and Boston: Kluwer, 2003.

[Jacquette, 2003b] Dale Jacquette. Psychologism revisited in logic, metaphysics and epistemology. In Dale Jacquette, editor, *Philosophy, Psychology and Psychologism: Critical and Historical Readings on the Psychological Turn in Philosophy*, pages 245–262. Dordrecht and Boston: Kluwer, 2003.

[Jacquette, 2003c] Dale Jacquette. Psychologism, the philosophical shibboleth. In Dale Jacquette, editor, *Philosophy, Psychology and Psychologism: Critical and Historical Readings on the Psychological Turn in Philosophy*, pages 1–12. Dordrecht and Boston: Kluwer, 2003.

[Jacquette, 2004] Dale Jacquette. Assumption and mechanical simulation of hypothetical reasoning. In Arkadiusz Chrudzimski and Wolfgang Heumer, editors, *Phenomenology and Mind.* Verlag, 2004.

[Jáskowski, 1948] S. Jáskowski. Rachunek zdak dla systemów dedukcyzjych sprzecznych. *Studia Societatis Torunesis*, 1:55–77, 1948. Sectio A.

[Jeffrey, 1983] R. Jeffrey. *The Logic of Decision.* Chicago, IL: University of Chicago Press, 2nd edition, 1983. First published in 1965.

[Jennings and Schotch, 1981] R.E. Jennings and P.K. Schotch. Some remarks on (weakly) weak modal logics. *Notre Dame Journal of Formal Logic*, 22:309–314, 1981.

[Johnson and Blair, 1983] Ralph H. Johnson and J. Anthony Blair. *Logical Self Defence.* Toronto, ON: McGraw Hill, 2nd edition, 1983.

[Johnson, 1981] Ralph H. Johnson. Charity begins at home. *Informal Logic Newsletter*, 3:4–9, June 1981.

[Johnson, 1999] George Johnson. *Strange Beauty: Murray Gell-Mann and the Revolution in Twentieth-Century Physics.* New York: Alfred A. Knopf, 1999.

[Josephson and Josephson, 1994] J.R. Josephson and S.G. Josephson, editors. *Abductive Inference: Computation, Philosophy, Technology.* Cambridge: Cambridge University Press, 1994.

[Kahneman *et al.*, 1982] D. Kahneman, P. Slovic, and A. Tversky. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press, 1982.

[Kakas and Mancarella, 1990] A. Kakas and P. Mancarella. Knowledge assimilation and abduction. In João P. Martins and Michael Reinfrank, editors, *Truth Maintenance Systems: Proceedings of the European Conference on Artificial Intelligence, ECAI.* Berlin: Springer-Verlag, 1990.

[Kakas and Mancarella, 1994] A. Kakas and P. Mancarella. Abduction and abductive logic programming. *ICLP*, 23:18–19, 1994.

[Kakas *et al.*, 1995] A. Kakas, R.A. Kowalski, and F. Toni. Abductive logic programming. *Journal of Logic and Computation*, 2:719–770, 1995.

[Kant, 1929] Immanuel Kant. *Critique of Pure Reason.* London: MacMillan, 1929. Norman Kemp Smith, translator.

[Kant, 1933] Immanuel Kant. *Critique of Pure Reason.* London: Macmillan, 1933. Norman Kemp Smith, translator. Originally published 1781–1782.

[Kant, 1974] Immanuel Kant. *Logic.* Mineola, NY: Cover Publications, 1974. Robert S. Hartman and Wolfgang Schwarz translators. First published in 1800.

[Kapitan, 1992] Tomis Kapitan. Peirce and the autonomy of abductive reasoning. *Erkenntnis*, 37:1–26, 1992.

[Kapitan, 1997] Tomis Kapitan. Peirce and the structure of abductive inference. In Nathan Houser, Don D. Roberts, and James Van Evra, editors, *Studies in the Logic of Charles Sanders Peirce*, pages 477–496. Bloomington, IN: Indianna University Press, 1997.

[Kaplan, 1996] M. Kaplan. *Decision Theory as Philosophy.* Cambridge: Cambridge University Press, 1996.

[Kelly, 1989] G.M. Kelly. Elementary observations on 2-categorical limits. *Bull. Aust. Math. Soc.*, 39(2):301–317, Apr 1989.

[Kempson, 1975] Ruth Kempson. *Presupposition and the delimination of semantics.* Cambridge: Cambridge University Press, 1975.

[Keynes, 1921] John Maynard Keynes. *A Treatise in Probability.* New York and London: Harper and Row and Macmillan, 1921.

[Kleer and Williams, 1987] J. De Kleer and B.C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[Klotter and Ingram, 2003] John T. Klotter and Jefferson Ingram. *Criminal Evidence.* Cincinatti, OH:Anderson Publishing, 8th edition, 2003.

[Klotter, 1992] John C. Klotter. *Criminal Evidence.* Cincinnati, OH: Anderson Publishing, 5th edition, 1992.

[Kolb and Whishaw, 2001] Bryan Kolb and Ian Q. Whishaw. *An Introduction to Brain and Behavior*. New York: Worth, 2001.

[Konolege, 1996] K. Konolege. A general theory of abduction. In G. Brewska, editor, *Principles of Knowledge Representation*, pages 129–152. Stanford, CA: CLSI, 1996.

[Kowalski, 1979] R.A. Kowalski. *Logic for Problem Solving*. New York: Elsevier, 1979.

[Kraus *et al.*, 1990] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotronic reasoning: Preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[Kraus, 2003] Manfred Kraus. Charles Peirce's theory of abduction and the Aristotelian enthymeme from signs. In Frans H. van Eemeren, J. Anthony Blair, Charles A. Willard, and A. Francisca Snoeck Henkemans, editors, *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation.* Amsterdam: Sic Sat, 2003.

[Krifka *et al.*, 1995] Manfred Krifka, Francis Jeffry Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link, and Germano Chierchia. Genericity: An introduction. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, pages 1–124. Chicago, IL: The Univeristy of Chicago Press, 1995.

[Kruijff, 1998] Geert-Jan M. Kruijff. Peirce's late theory of abduction. In *Logica.* Liblice: Czech Republic, 1998.

[Kuhn, 1962] Thomas Kuhn. *The Structure of Scientific Revolutions*. Cambridge, MA: Harvard University Press, 1962.

[Kuhn, 1967] Thomas Kuhn. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, 1967.

[Kuhn, 1970] Thomas Kuhn. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, enlarged edition, 1970.

[Kuhn, 1977] Thomas Kuhn. *The Essential Tension*. Chicago, IL: University of Chicago Press, 1977.

[Kuipers and Wisniewski, 1994] Theo A.F. Kuipers and A. Wisniewski. An erotetic approach to explanation by specification. *Erkenntnis*, 40.3:377–402, 1994.

[Kuipers, 1999] Theo A.F. Kuipers. Abduction aiming at empirical progress of even truth approximation leading to a challenge for computational modelling. *Foundations of Science*, 4:307–323, 1999.

[Kuipers, 2000] Theo A.F. Kuipers. *From Instrumentalism to Constructive Realism: On Some Relations Between Confirmation, Empirical Progress and Truth Approximation.* Dordrecht and Boston: Kluwer, 2000.

[Kuipers, 2001] Theo A.F. Kuipers. *Structures in Science: Heuristic Patterns Based on Cognitive Structures.* Amsterdam: Kluwer, 2001. Part of the Synthese Library.

[Kyburg, 1983] Henry Kyburg. *Epistemology and Inference.* Minneapolis, MN: University of Minnesota Press, 1983.

[Lakatos, 1968] Imre Lakatos. Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, 69:149–186, 1968.

[Lakatos, 1970] Imre Lakatos. Falsification and methodology of scientific research programmes. In Imre Lakatos and A. Musgrave, editors, *Criticism and the Growth of Knowledge*, pages 91–196. Cambridge: Cambridge University Press, 1970.

[Langley *et al.*, 1987] P. Langley, H.A. Simon, G.L. Bradshaw, and J.M. Zytkow. *Scientific Discovery.* Cambridge, MA: MIT Press, 1987.

[Laplace, 1904] Pierre Laplace. Mémoires. In *Oeuvres completes de Laplace*, volume 13/14. Gauthier-Villars 1878–1912, 1904.

[Laplace, 1951] Pierre Laplace. *Philosophical Essay on Probabilities.* New York: Dover, 1951. Frederick Wilson Truscott and Frederick Lincoln Emory, translators. Originally published in 1795.

[Laudan, 1980] Larry Laudan. Why was the logic of discovery abandoned? In Thomas Nickles, editor, *Scientific Discovery, Logic and Rationality*, pages 173–183. Dordrecht: Reidel, 1980.

[Lauritzen and Spiegelharter, 1988] S. Lauritzen and D. Spiegelharter. Local computation with probabilities in graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society B*, 50:157–224, 1988.

[Lehrer, 1974] Keith Lehrer. *Knowledge.* Oxford: Oxford University Press, 1974.

[Lempert, 1986] R. Lempert. The new evidence scholarship: Analyzing the process of proof. *Boston University Law Review*, 66:439–477, 1986.

[Levi, 1949] Edward H. Levi. *An introduction to legal reasoning.* University of Chicago Press, Chicago, 1949.

[Levi, 1980] I. Levi. *The Enterprise of Knowledge.* Cambridge, MA: MIT Press, 1980.

[Levinson, 2001] Stephen C. Levinson. *Presumptive Meanings: The Theory of Generalised Conversational Implicature.* Cambridge, MA: MIT Press, 2001.

[Levy and Bullinaria, 1999] J.F. Levy and J.A. Bullinaria. Learning lexical properties from word usage patterns: Which context words should be used? In

R.F. French and J.P. Sounge, editors, *Connectionist Models of Learning Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop*, pages 273–282. Berlin: Springer-Verlag, 1999.

[Lewis, 1973] David Lewis. *Counterfactuals*. Cambridge, MA: Harvard University Press, 1973.

[Lewontin, 2003] Richard Lewontin. Science and simplicity. *New York Review of Books*, 50(7), May 2003.

[Lichter, 1995] Tilman Lichter. Bill Clinton is the First Lady of the USA: Making and unmaking analogies. *Synthese*, 104:285–297, 1995.

[Lilien *et al.*, 1992] Gary L. Lilien, Philip Kotler, and K. Sridhar Moorthy. *Marketing Models*. Englewood Cliffs, NJ: Prentice Hall, 1992.

[Lipton, 1991] Peter Lipton. *Inference to the Best Explanation*. London: Routledge, 1991. Second edition, 2004.

[Llewellyn, 1930] K. N. Llewellyn. *The Bramble Bush*. Oceana, New York, 1930.

[Lloyd, 1987] J.W. Lloyd. *Foundations of Logic Programming*. Berlin: Springer-Verlag, 2nd edition, 1987.

[Lorenzen and Lorenz, 1978] Paul Lorenzen and Kuno Lorenz. *Dialogische Logik*. Darmstadt: Wissenschaft-liche Buchgesellschaft, 1978.

[Lowe, 2000] W. Lowe. What is the dimensionality of human semantic space? In R.F. French and J.P. Sounge, editors, *Connectionist Models of Learning Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop*, pages 303–311. Berlin: Springer-Verlag, 2000.

[Lowe, 2001] W. Lowe. Towards a theory of semantic space. In J.D. Moore and K. Stenning, editors, *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 576–581. Mahwah NJ: Erlbaum, 2001.

[Lund and Burgess, 1996] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments and Computers*, 28:203–208, 1996.

[Lycan, 1988] William G. Lycan. *Judgement and Justification*. Cambridge: Cambridge University Press, 1988.

[MacCormick, 1994] Neil MacCormick. *Legal Reasoning and Legal Theory*. Oxford: Oxford University Press, 1994.

[Magnani, 2001a] Lorenzo Magnani. *Abduction, Reason and Science: Processes of Discovery and Explanation*. New York: Kluwer, Plenum, 2001.

[Magnani, 2001b] Lorenzo Magnani. Philosophy and geometry: Theoretical and historical issues. Dordrecht: Kluwer, 2001.

[Maher, 1993] P. Maher. *Betting on Theories*. Cambridge: Cambridge University Press, 1993.

[Maloney and Mulherin, forthcoming] Michael T. Maloney and J. Harold Mulherin. The complexity of price discovery in an efficient market: The stock market reaction to the *challenger* crash. *Journal of Corporate Finance*, forthcoming.

[McAllister, 1996] J. McAllister. *Beauty and Revolution in Science*. Ithaca, NY: Cornell University Press, 1996.

[Meheus *et al.*, forthcoming] Joke Meheus, Liza Verhoeven, Maarten Van Dyck, and Dagmar Provijn. Ampliative adaptive logics and the foundation of logic-based approaches to abduction. In Lorenzo Magnani, Nancy J. Nersessian, and Claudio Puzzi, editors, *Logical and Computational Aspects of Model-Based Reasoning*. Dordrecht and Boston: Kluwer, forthcoming.

[Menschel, 2002] Robert Menschel. *Markets, Mobs and Mayhem*. New York: Wiley, 2002.

[Mill, 1959] J.S. Mill. *A System of Logic*. London: Longman's Green, 1959.

[Mill, 1974] John Stuart Mill. A system of logic. In J.M. Robson and J. Stillinger, editors, *The Collected Works of John Stuart Mill*, volume VII and VIII. Toronto, ON: University of Toronto Press, 1974. Originally published in 1843, London: Longman and Green. Volume VII was published in 1973 and Volume VIII was published in 1974.

[Miller *et al.*, 1982] R. Miller, H. Pople, and J Meyers. Internist-I: An experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307:468–476, 1982.

[Miller, 2000] Geoffrey Miller. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. New York: Doubleday, 2000.

[Millikan, 1984] Ruth Millikan. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press, 1984.

[Millikan, 1989] Ruth Garrett Millikan. An ambiguity in the notion of 'function'. *Biology and Philosophy*, 4:172–176, 1989.

[Mink, 1969] Louis O. Mink. *Mind, History and Dialectic: The Philosophy of R.G. Collingwood*. Bloomington, IN: Indiana University Press, 1969.

[Minsky, 1975] Marvin Minsky. Frame-system theory. In R.C. Schank and B.L. Nash-Webber, editors, *Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing*. Cambridge, MA: Yale University Press, 1975. Preprints of a conference at MIT, June 1975. Reprinted in P.N. Johnson-Laird and P.C. Wason, editors. *Thinking: Readings in Cognitive Science*, Cambridge: Cambridge University Press 1977; pp. 355–376.

[Mitchell, 1989] S.D. Mitchell. The causal background of functional explanation. *International Studies in the Philosophy of Science*, 3:213–230, 1989.

[Murphy, 2000] Peter Murphy. *Murphy on Evidence*. London: Blackstone, 7th edition, 2000. First published in 1980.

[Musgrave, 1989] Alan Musgrave. Deductive heuristics. In Kostas Gavroglu, Yorgos Goudaroulis, and Pantelis Nicolacopoulos, editors, *Imre Lakatos and Theories of Scientific Change*, pages 15–32. Dordrecht: Kluwer, 1989.

[Myers, 2002] David G. Myers. *Intuition: Its Powers and Perils*. New Haven, CT: Yale University Press, 2002.

[Nagel, 1974] Thomas Nagel. What is it like to be a bat? *Philosophical Review*, 83:435–450, 1974.

[Nagel, 1977] Ernest Nagel. Teleology revisited. *Journal of Philosophy*, 74:261–301, 1977.

[Neander, 1991] Karen Neander. Function as selected effects: The conceptual analyst's defense. *Philosophy of Science*, 58:168–184, 1991.

[Nersessian, 1994] N.J. Nersessian. Opening the black box: Cognitive science and history of science. Technical Report GIT–COG SCI 94/23, Georgia Institute of Technology, July 1994. Partially published in *Osiris*.

[Nersessian, 1995] N.J. Nersessian. Should physicists preach what they practice? constructive modelling in doing and learning physics. *Science and Education*, 4:203–220, 1995.

[Newton, 1713] Isaac Newton. *Philosopiae Naturalis Principia Mathematica*. Berkeley and Los Angeles: University of California Press, 1713. Andrew Motte, translator (1729); revised by Florian Cajori, including General Scholium of 1713. Reprinted 1960.

[Nickles, 1980] Thomas Nickles. Introductory essay: Scientific discovery and the future of philosophy of science. In Thomas Nickles, editor, *Scientific Discovery, Logic and Rationality*, pages 1–59. Dordrecht: Reidel, 1980.

[Nietsche, 1966] Friedrich Nietsche. *Beyone Good and Evil*. New York: Random House, 1966. Walter Kaufmann, translator.

[Norman, 1993] D.A. Norman. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Reading, MA: Addison-Wesley, 1993.

[Oakley and Johnson-Laird, 1987] P. Oakley and P.N. Johnson-Laird. Towards a cognitive of emotions. *Cognition and Emotions*, 1:29–50, 1987.

[Oatley, 1996] I. Oatley. Inference in narrative and science. In D.R. Olson and N. Torrance, editors, *Models of Thought: Exploration in Culture and Cognition*, pages 123–140. Cambridge: Cambridge University Press, 1996.

[O'Keefe, 1990] D.J. O'Keefe. *Persuasion: Theory and Research*. Thousand Oaks, CA: Sage, 1990.

[Olivetti and Terracini, 1991] N. Olivetti and L. Terracini. N-prolog and equivalence of logic programs part 1. Technical report, University of Turin, 1991.

[O'Rorke and Ortony, 1992] P.O. O'Rorke and A. Ortony. Abductive explanations of emotions. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pages 283–323. Hillsdale, NJ: Erlbaum, 1992.

[O'Rorke and Ortony, 1995] P.O. O'Rorke and A. Ortony. Explaining emotions. *Cognitive Science*, 18:283–323, 1995.

[Paavola and Hakkarainen, forthcoming] Sami Paavola and Kai Hakkarainen. Three abductive solutions to the Mano Paradox: with instinct and distributed cognition. In *Studies in Philosophy and Education*. forthcoming.

[Pap, 1962] Arthur Pap. *An Introduction to the Philosophy of Science*. New York: Free Press, 1962.

[Parret, 1978] Herman Parret. A note on pragmatic universals of language. In *Language Unviersals*, pages 125–140. Tubingen: Narr, 1978.

[Patel *et al.*, 1997] M. Patel, J.A. Bullinaria, and J.P. Levy. Extracting semantic representations from large text corpora. In R.F. French and J.P. Sounge, editors, *Connectionist Models of Learning, Development and Evolutions: Proceedings of the Fourth Neural Computation and Psychology Workshop*, pages 199–212. London: Springer-Verlag, 1997.

[Pauker *et al.*, 1976] S. Pauker, G. Gorry, J. Kassirer, and M. Schwarz. Towards the simulation of clinical cognition. *American Journal of Medicine*, 60:981–996, 1976.

[Paul, 1993] G. Paul. Approaches to abductive reasoning: An overview. *Artificial Intelligence Review*, 7:109–152, 1993.

[Peacock, 2001] Kent Peacock. Flandern is wrong. Draft, 2001.

[Pearl, 1987] Judea Pearl. Distributed revision of composite beliefs. *Artificial Intelligence*, 33:173–215, 1987.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[Pearl, 2000] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2000.

[Peirce, 1931–1958] C.S. Peirce. *Collected Works*. Cambridge, MA: Harvard University Press, 1931–1958. A series of volumes, the first appearing in 1931.

[Peirce, 1955] C.S. Peirce. Perceptual judgements. In Julius Buhler, editor, *Philosophical Writings of Peirce*, pages 302–305. New York: Dover, 1955.

[Peirce, 1992] Charles Sanders Peirce. *Reasoning and the logic of things: The Cambridge Conference lectures of 1898*. Cambridge, MA: Harvard University Press, 1992. Kenneth Laine Ketner, editor, with an introduction by Kenneth Laine Ketner and Hilary Putnam.

[Peng and Reggia, 1990] Yun Peng and James A. Reggia. *Abductive Inference Models for Diagnostic Problem-solving*. New York and Berlin: Springer-Verlag, 1990.

[Pennington and Hastie, 1986] N. Pennington and R. Hastie. Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, 51:242–258, 1986.

[Petty and Cacioppo, 1986] R.E. Petty and J.T. Cacioppo. *Communication and Persuasion*. New York: Springer-Verlag, 1986.

[Petty *et al.*, 1981] R.E. Petty, J.T. Cacioppo, and R. Goldman. Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41:847–855, 1981.

[Polanyi, 1966] Michael Polanyi. *The Tacit Dimension.* London: Routledge & Kegan Paul, 1966.

[Polya, 1945] G. Polya. *How to Solve it. A New Aspect of Mathematical Method.* Princeton, NJ: Princeton University Press, 1945.

[Polya, 1954] George Polya. *Induction and Analogy in Mathematics*, volume 1 of *Mathematics and Plausible Reasoning.* Princeton, NJ: Princeton University Press, 1954.

[Polya, 1962] G. Polya. *Mathematical Discovery. On Understanding, Learning, and Teaching Problem Solving*, volume I. New York: Wiley, 1962.

[Poole *et al.*, 1987] D. L. Poole, R. Goebel, and R. Aleliunas. Theorist: a logical reasoning system for defaults and diagnosis. In N. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352. Springer Varlag, New York, 1987.

[Pople, 1975] H. Pople. Dialog: A model of diagnostic logic for internal medicine. *Proceedings of International Joint Conference on Artificial Intelligence, IJCAI*, 1975.

[Popper and Eccles, 1983] Karl R. Popper and John C. Eccles. *The Self and Its Brain.* London: Routledge and Kegan Paul, 1983. Originally published in 1977.

[Popper, 1934] K.R. Popper. *The Logic of Scientific Discovery.* London: Hutchinson, 1934. In large part a translation of *Logik der Forschung*, Vienna: Springer, 1934.

[Popper, 1963] Karl Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge and Kegan, 1963.

[Port and van Gelder, 1995] R.F. Port and T. van Gelder, editors. *Mind as Motion: Explorations of the Dynamics of Cognition.* Cambridge, MA: MIT Press, 1995.

[Priest, 2002] Graham Priest. Paraconsistent logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 6, pages 287–393. Dordrecht and Boston: Kluwer, 2nd edition, 2002.

[Pro, 1973] Proceedings of the Third International Joint Conference on Artificial Intelligence, IJCAAI-73. *On the Mechanization of Abductive Logic.* Stanford, CA: Morgan Kauffmann, 1973.

[Punch *et al.*, 1986] W. Punch, M. Tanner, and J. Josephson. Design consideration for Peirce, a high-level language for hypothesis assembly. *Proceedings of Expert Systems in Government Symposium*, pages 279–281, 1986.

[Putnam, 1975a] Hilary Putnam. 'Degree of Confirmation' and inductive logic. In Hilary Putnam, editor, *Mathematics, Matter and Method*, pages 270–292. Cambridge: Cambridge University Press, 1975.

[Putnam, 1975b] Hilary Putnam. Probability and confirmation. In Hilary Putnam, editor, *Mathematics, Matter and Method*, pages 293–304. Cambridge: Cambridge University Press, 1975. Originally published in *The Voice of America, Forum Philosophy of Science* volume 10 in 1963.

[Putnam, 1992] Hilary Putnam. Comments on the lectures. In Kenneth Laine Ketner, editor, *Reasoning and the Logic of Things*, pages 55–102. Cambridge MA: Harvard University Press, 1992.

[Quine, 1951] W.V. Quine. Two dogmas of empiricism. In A.P. Martinich, editor, *The Philosophy of Language*, pages 39–52. New York: Oxford University Press, 1951.

[Quine, 1953] W.V. Quine. Two dogmas of empiricism. In W.V. Quine, editor, *From a Logical Point of View*, pages 20–46. Cambridge, MA: Harvard University Press, 1953. Originally published in [Quine, 1951].

[Quine, 1960] W.V. Quine. *Word and Object*. Cambridge, MA and New York: MIT Press and John Wiley, 1960.

[Quine, 1969a] W. V. Quine. Epistemology naturalized. In *Ontological Relativity and other Essays*, pages 69–90. New York: Columbia University Press, 1969.

[Quine, 1969b] W.V. Quine. Natural kinds. In Nicholas Rescher, editor, *Essays in Honor of Carl G. Hempel*, pages 5–23. Dordrecht: Reidel, 1969.

[Quine, 1976] W.V. Quine. *The Ways of Paradox, and Other Essays*. Cambridge, MA: Harvard University Press, 1976.

[Quine, 1992] W.V. Quine. *Pursuit of Truth*. Cambridge, MA and London: Harvard University Press, revised edition, 1992.

[Quine, 1995] W.V. Quine. *From Stimulus to Science*. Cambridge, MA: Harvard University Press, 1995.

[Ramoni *et al.*, 1992] M. Ramoni, M. Stefanelli, L. Magnani, and G. Barosi. An epistemological framework for medical knowledge-based systems. In *IEEE Transactions on Systems, Man and Cybernetics*, volume 22, pages 1361–1375. 1992.

[Ramsey, 1931] Frank Ramsey. Truth and probability. In Frank Ramsey, editor, *The Foundations of Mathematics and Other Essays*, pages 156–198. London: Routledge and Kegan Paul, 1931. Originally published in 1926.

[Raz, 1978] Joseph Raz. *Practical Reasoning*. Oxford: Oxford University Press, 1978.

[Read and Newhall, 1993] S.J. Read and A. Marcus Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65:429–447, 1993.

[Read, 1988] Stephen Read. *Relevant Logic: A Philosophical Examination of Inference*. Oxford: Blackwell, 1988.

[Reggia, 1981] J. Reggia. Knowledge-based decision support system development through kms. Technical Report TR-1121, Department of Computer Science, University of Maryland, Oct 1981.

[Reichenbach, 1938] Hans Reichenbach. *Experience and Prediction.* Chicago, IL: University of Chicago Press, 1938.

[Reiter, 1980] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 12:81–132, 1980.

[Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[Rensink, 2004] Ronald Rensink. Visual sensing without seeing. *Psychological Science*, 15:27–32, 2004.

[Rescher, 1964] Nicholas Rescher. *Hypothetical Reasoning*. Amsterdam: North-Holland, 1964.

[Rescher, 1976a] Nicholas Rescher. Peirce and the economy of research. *Philosophy of Science*, 43:71–98, 1976.

[Rescher, 1976b] Nicolas Rescher. *Plausible Reasoning: An Introduction to the Theory and Practice of Plausible Inference.* Assen and Amsterdam: Van Gorcum, 1976.

[Rescher, 1977] Nicholas Rescher. *Methodological Pragmatism: A systems-theoretic approach to the theory of knowledge*. Oxford: Blackwell, 1977.

[Rescher, 1995] Nicholas Rescher. Plausibility. In Ted Honderich, editor, *The Oxford Companion to Philosophy*. Oxford: Oxford University Press, 1995.

[Rescher, 1996] Nicholas Rescher. *Priceless Knowledge? Natural Science in Economic Perspective*. Lanham, MD: Rowman and Littlefield, 1996.

[Rescorla and Wagner, 1972] R.A. Rescorla and A.R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In *Classical Conditioning II: Current Research and Theory*, pages 64–99. New York: Appleton-Century-Crofts, 1972.

[Ricoeur, 1977] Paul Ricoeur. *The Rule of Metaphor: Multi-disciplinary Studies of the Creation of Meaning in Language*. Toronto, ON: University of Toronto Press, 1977. Robert Czerny, Kathleen McLaughlin and John Costello, translators. Originally published in 1975.

[Rock, 1983] I. Rock. *The Logic of Perception.* Cambridge, MA: MIT Press, 1983.

[Rodych, 2005] Victor Rodych. Are Platonism and pragmatism compatible? In Kent Peacock and Andrew Irvine, editors, *Mistakes of Reason*. Toronto, ON: University of Toronto Press, 2005. To appear.

[Rosch, 1978] Eleanor Rosch. Principles of categorization. In Eleanor Rosch and B.B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Hillsdale, NJ: Erlbaum, 1978.

[Rose and Rose, 2000] Steven Rose and Hilary Rose. *Alas, Poor Darwin: Arguments Against Evolutionary Psychology*. London: Jonathan Cape, 2000.

[Ruse, 1973] M. Ruse. *The Philosophy of Biology*. London: Hutchinson, 1973.

[Russell, 1907] Bertrand Russell. The regressive method of discovering the premises of mathematics. In Douglas Lackey, editor, *Essays in Analysis*. London: George Allen and Unwin, 1907. Published in 1973.

[Russell, 1973] Bertrand Russell. *Essays in Analysis*. London: George Allen and Unwin, 1973. Douglas Lackey (ed.).

[Sahlgren, 2002] M. Sahlgren. Towards a flexible model of word meaning. In *Proceedings of the 19th National Conference on Artificial Intelligence*. AAAI Spring Symposium, Stanford University, Palo Alto, CA, March 25–27 2002. Paper presented at the Symposium.

[Salmon, 1978] Wesley C. Salmon. Why ask 'why?'. *Proceedings and Addresses of the Americal Philosophical Association*, 51:683–705, 1978.

[Salmon, 1984] Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press, 1984.

[Savary, 1995] C. Savary. Discovery and its logic: Popper and the friends of discovery. *Philosophy of the Social Sciences*, 25:318–344, 1995.

[Schaerf and Cadoli, 1995] Marco Schaerf and Marco Cadoli. Tractable reasoning via approximation. *Artificial Intelligence*, 74:249–310, 1995.

[Schank and Abelson, 1977] Roger Schank and Robert Abelson. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

[Schank and Ranney, 1991] P. Schank and M. Ranney. Modeling an experimental study of explanatory coherence. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pages 892–897. Hillsdale, NJ: Erlbaum, 1991.

[Schank and Ranney, 1992] P. Schank and M. Ranney. Assessing explanatory coherence: A new method for integrating verbal data with models of on-line belief revision. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pages 599–604. Hillsdale, NJ: Erlbaum, 1992.

[Scott, 1972] D. Scott. Continuous lattices. In *Proceedings of the 1971 Dalhousie Conference: LNM 274*, pages 97–136. Amsterdam: Springer-Verlag, 1972.

[Scriven, 1976] Michael Scriven. *Reasoning*. New York: McGraw-Hill, 1976.

[Shagrin *et al.*, 1985] Morton L. Shagrin, William J. Rapaport, and Randall R. Dipert. *Logic: A Computer Approach*. New York: McGraw-Hill Book Company, 1985.

[Shanks and Dickinson, 1987] D.R. Shanks and A. Dickinson. Associative accounts of causality judgment. In G.G. Bower, editor, *The Psychology of Learning and Motivation*, volume 21, pages 229–261. San Diego, CA: Academic Press, 1987.

[Shapere, 1977] Dudley Shapere. Scientific theories and their domains. In *The Structure of Scientific Theories*, pages 518–565. Urbana: University of Illinois Press, 2nd edition, 1977.

[Shi, 2001] Yanfei Shi. *The Economics of Scientific Knowledge: A Rational Choice Neo-Instituitionalist Theory of Science.* Cheltenham, UK: Edward Elgar, 2001.

[Shiffrin, 1997] Richard M. Shiffrin. Attention, automatism and consciousness. In Jonathan D. Cohen and Jonathan W. Schooler, editors, *Scientific Approaches to Consciousness*, pages 49–64. Mahwah, NJ: Erlbaum, 1997.

[Shrager and Langley, 1990] J. Shrager and P. Langley. *Computational Models of Scientific Discovery and Theory.* San Mateo, CA:Morgan Kaufmann, 1990.

[Shubin and Ulrich, 1982] H. Shubin and J. Ulrich. Idt: An intelligent diagnostic tool. In *Proceedings of the National Conference on Artificial Intelligence, AAAI*, pages 290–295, 1982.

[Shugan, 1980] Steven M. Shugan. The cost of thinking. *Journal of Consumer Research*, 7:99–111, 1980.

[Simon *et al.*, 1981] H. Simon, P. Langley, and G. Bradshaw. Scientific reasoning as problem solving. *Synthese*, 47:1–27, 1981.

[Simon, 1957] H.A. Simon. *Models of Man.* New York: John Wiley, 1957.

[Simon, 1973] H. A. Simon. The structure of ill-structured problems. *Artificial Intelligence*, 4:181–202, 1973.

[Simon, 1977] Herbert Simon. Does scientific discovery have a logic. In Herbert Simon, editor, *Models of Discovery*, pages 326–337. Dordrecht: Reidel, 1977.

[Simon, 1982] Herbert A. Simon. *Models of Bounded Rationality: Behavioral Economics and Business Organization*, volume 2. Cambridge and London: The MIT Press, 1982.

[Sintonen, 1993] Matti Sintonen. In search of explanations: From why-questions to Shakespearian questions. *Philosophica*, 51:55–81, 1993.

[Smith and Medin, 1981] Edward E. Smith and Douglas L. Medin. *Categories and Concepts*. Cambridge, MA: Harvard University Press, 1981.

[Smullyan, 1968] R.M. Smullyan. *First Order Logic.* Amsterdam: Springer-Verlag, 1968.

[Song and Bruza, 2000] D. Song and P. Bruza. Fundamental properties o fthe core matching functions for information. In *Proceedings of the 13th International Florida Artificial Intelligence Society Conference*, volume D. Song and K. Wong and P. Bruza and C. Cheng, 2000.

[Song and Bruza, 2001] Dawei Song and Peter D. Bruza. Discovering information flow using a high dimensional conceptual space. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001 Proceedings of the 24th Annual International Conference on Research and De-*

velopment in Information Retrieval, pages 327–333, New Orleans, LA, 2001. ACM.

[Song and Bruza, 2003] D. Song and P. Bruza. Toward context sensitive information inference. *Journal of the American Society for Information Science and Technology*, 52:321–334, 2003.

[Sperber and Wilson, 1986] Dan Sperber and Deirdre Wilson. *Relevance*. Oxford: Basil Blackwell, 1st edition, 1986.

[Stalnaker, 1973] Robert Stalnaker. Presupposition. *Journal of Philosophical Logic 2*, pages 447–457, 1973.

[Stanovich, 1999] Keith A. Stanovich. *Who is Rational? Studies of Individual Differences in Reasoning*. Mahawah, NJ: Erlbaum, 1999.

[Starmans, 1996] Richard Starmans. *Logic, Argument and Commonsense*. Tilburg: Tilburg University Press, 1996.

[Stigler, 1961] George J. Stigler. The economics of information. *The Journal of Political Economy*, LXIX(3):213–224, 1961.

[Stough, 1969] Charlotte L. Stough. *Greek Skepiticism*. Berkeley and Los Angeles, CA: University of California Press, 1969.

[Strong, 1999] John W. Strong, editor. *MacCormick on Evidence*. St. Paul, MN: West Group, 5th edition, 1999.

[Struss and Dressler, 1989] P. Struss and O. Dressler. Physical negation — integrating fault models into the general diagnostic engine. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1318–1323. Detroit, MI: Morgan Kaufmann, 1989.

[Suppe, 1977] Frederick Suppe, editor. *The Structure of Scientific Theories*. Urbana, IL: University of Illinois Press, 2nd edition, 1977.

[Suppes, 1962] Patrick Suppes. Models of data. In Ernest Nagel, Patrick Suppes, and Alfred Tarski, editors, *Logic, Methodology and Philosophy of Science*, pages 252–261. Stanford, CA: Standford University Press, 1962. Originally published in 1960.

[Surowiecki, 2004] James Surowiecki. *The Wisdom of Crowds*. New York: Doubleday, 2004.

[Swanson and Smalheiser, 1997] D.R. Swanson and N.R. Smalheiser. An interactive system for finding complementary literature, a stimulus for scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.

[Swanson, 1986] D.R. Swanson. Fish oil, Raynaud's Syndrome and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30:7–18, 1986.

[Tallis, 1999] Raymond Tallis. *The Explicit Animal: A Defence of Human Consciousness*. London: Macmillan and New York: Martin's Press, 2nd edition, 1999.

[Tanaka, 2003] Koji Tanaka. Three schools of paraconsistency. *Australasian Journal of Logic*, 1:28–42, 2003.

[Tangen and Allen, in press] M. Tangen and Lorraine G. Allen. Cue-interaction and judgments of causality: Contributions of causal and associative processes. *Memory and Cognition*, in press.

[Tarski, 1956] Alfred Tarski. The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*, pages 152–278. Oxford: Clarendon Press, 1956.

[Thagard and Kunda, 1998] Paul Thagard and Z. Kunda. Making sense of people: Coherence mechanisms. In S.J. Read and L.C. Miller, editors, *Connectionist Models of Social Reasoning and Social Behavior*, pages 3–26. Hillsdale, NJ: Erlbaum, 1998.

[Thagard and Verbeurgt, 1998] Paul Thagard and K. Verbeurgt. Coherence as constraint satisfaction. *Cognitive Science*, 22:22–24, 1998.

[Thagard, 1988] Paul Thagard. *Computational Philosophy of Science*. Princeton, NJ: Princeton University Press, 1988.

[Thagard, 1989] Paul Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12:381–433, 1989.

[Thagard, 1992] Paul Thagard. *Conceptual Revolutions*. Princeton, NJ: Princeton University Press, 1992.

[Thagard, 1993] Paul Thagard. Computational tractability and conceptual coherence: Why do computer scientists believe that $p \neq np$? *Canadian Journal of Philosophy*, 23:349–364, 1993.

[Thagard, 1995] Paul Thagard. *Abductive Reasoning: Logic, Visual Thinking, and Coherence*. Waterloo, ON: University of Waterloo, 1995.

[Thagard, 2000] Paul Thagard. *Coherence in Thought and Action*. Cambridge, MA: MIT Press, 2000.

[Thalos, 2002] Miriam Thalos. Explanation is a genius: An essay on the varieties of scientific explanation. *Synthese*, 13:317–354, 2002.

[Thomas, 1977] Stephen N. Thomas. *Practical Reasoning in Natural Language*. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

[Thomson, 1971] Judith Jarvis Thomson. A defence of abortion. *Philosphy and Public Affairs*, 1:47–66, 1971.

[Tillers and Green, 1988] Peter Tillers and Eric D. Green. *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism*. Dordrecht and Boston: Kluwer, 1988.

[Timmerman *et al.*, 1998] W. Timmerman, F. Westerhof, T. van der Wall, and B.C. Westerink. Striatal dopamine-glutamate interactions reflected in substantia nigra reticulata firing. *Neuroreport*, 9:3829–3836, 1998.

[Toulmin, 1972] Stephen Toulmin. *Human Understanding*, volume I. Princeton, NJ: Princeton University Press, 1972.

[Turing, 1950] A. Turing. Checking a large routine. In *Rep. Conf. High Speed Automatic Calculating Machines*, Institute of Compupter Science, University of Toronto, ON, Jan 1950.

[Uglow, 1997] Steve Uglow. *Evidence*. London: Sweet & Maxwell, 1997.

[van Benthem and van Rooy, 2003] Johan van Benthem and Robert van Rooy. Connecting the different faces of information. *Journal of Logic, Language and Information*, 12:375–379, 2003.

[van Benthem, 1996] Johan van Benthem. *Exploring Logical Dynamics*. Stanford, CA: CSLI, 1996.

[van den Bosch, 1997] Alexander P.M. van den Bosch. Rational drug design as hypothesis formation. In P. Weingartner, G. Schurz, and G. Dorn, editors, *209th International Wittgenstein Symposium I*, pages 102–108. Kirchberg am Wechsel: The Austrian Ludwig Wittgenstein Society, 1997.

[van den Bosch, 1998] Alexander P.M. van den Bosch. Qualitative drug lead discover. In *Working Notes of the International Congress on Discovery and Creativity*, pages 163–165. Ghent, 1998.

[van den Bosch, 2001] Alexander P.M. van den Bosch. *Rationality in Discovery*. Amsterdam: Institute for Logic, Language and Computation, 2001.

[van Eemeren and Grootendorst, 1983] Frans H. van Eemeren and Rob Grootendorst. *Speech Acts in Argumentative Discourse*. Dordrecht: Foris Publication, 1983.

[van Flandern, 1999] Thomas van Flandern. Status of the NEAR challenge. *Meta Res. Bull*, 8:31–32, 1999.

[van Fraassen, 1980] Bas C. van Fraassen. *The Scientific Image*. Oxford: Clarendon Press, 1980.

[Van Fraassen, 1989] Bas Van Fraassen. *Laws of Symmetry*. Oxford: Oxford University Press, 1989.

[Velleman, 2003] J. D. Velleman. Narrative explanation. *Philosophical Review*, 112, 2003.

[Waldmann and Holyoak, 1992] M.R. Waldmann and K.J. Holyoak. Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121:222–236, 1992.

[Waldmann, 2000] M.R. Waldmann. Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26:53–76, 2000.

[Walton, 1992a] Douglas Walton. *Plausible Argument in Everyday Conversation*. Albany, NY: SUNY Press, 1992.

[Walton, 1992b] Douglas N. Walton. *Slippery Slope Arguments*. Oxford: Clarendon Press, 1992.

[Walton, 2002] Douglas Walton. The sunk costs fallacy or argument from waste. *Argumentation*, 16:473–503, 2002.

[Walton, 2004] Douglas Walton. *Abductive Reasoning*. Tuscaloosa: University of Alabama Press, 2004.

[Weber and Perkins, 1992] Robert J. Weber and David N. Perkins. *Inventive Minds: Creativity in Technology*. Oxford: Oxford University Press, 1992.

[Weeber *et al.*, 2001] M. Weeber, H. Klein, L. Jong van den Berg, and R. Vos. Using concepts in literature-based discovery simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52:548–557, 2001.

[Whewell, 1840 and 1967] W. Whewell. *The Philosophy of the Inductive Sciences*. New York: Johnson Reprint Corporation, 1840 and 1967.

[White, 1968] Hayden V. White. *The Uses of History: Essays in Intellectual and Social History*. Detroit: Wayne State University, 1968.

[Whitehead and Russell, 1910] Alfred North Whitehead and Bertrand Russell. *Principia Mathematica*, volume I of III. Cambridge: Cambridge University Press, 1910. Second edition printed in 1925.

[Wiles, 1995] Andrew J. Wiles. Modular elliptic curves and fermat's last theorem. *Annala of Mathematics*, 141:443–551, 1995.

[Wiles, 2000] Andrew Wiles. Solving fermat. An interview on NOVA Online, November 2000.

[Wilson, 1959] N.L. Wilson. Substances without substrata. *Review of Metaphysics*, 12:521–539, 1959.

[Wimsatt, 1972] W.C. Wimsatt. Teleology and the logical structure of function statements. *Studies in History and Philosophy of Science*, 3:1–80, 1972.

[Wisniewski, 1995] Andrzej Wisniewski. *The Posing of Questions: Logical Foundations of Erotetic-Inferences*. Dordrecht and Boston: Kluwer, 1995.

[Wolfram, 1984] Stephen Wolfram. Computer softwear in science and mathematics. *Scientific American*, 251:188, September 1984.

[Woods and Hudak, 1989] John Woods and Brent Hudak. By parity of reasoning. *Informal Logic*, XI:125–140, 1989. Reprinted in [Woods, 2004].

[Woods and Hudak, 1992] John Woods and Brent Hudak. Verdi is the Puccini of music. *Synthese*, 92:189–220, 1992. Reprinted in [Woods, 2004].

[Woods and Walton, 1972] John Woods and Douglas Walton. On fallacies. *Journal of Critical Analysis*, 5:103–111, 1972. reprinted in *Fallacies: Selected Papers 1972-1982* Dordrecht and Providence, RI: Foris Publications 1989; 1–10.

[Woods *et al.*, 2002] John Woods, Ralph H. Johnson, Dov M. Gabbay, and Hans Jürgen Ohlbach. Logic and the practical turn. In Dov M. Gabbay, Ralph H. Johnson, Hans Jürgen Ohlbach, and John Woods, editors, *Handbook of the Logic of Argument and Inference: The Turn Toward the Practical*, Studies in Logic and Practical Reasoning, pages 1–40. Amsterdam: North-Holland, 2002.

[Woods, 1989] John Woods. The relevance of relevant logic. In J. Norman and R. Sylvan, editors, *Directions in Relevant Logics*, pages 77–86. Dordrecht: Kluwer Academic Publisher, 1989.

[Woods, 1998] John Woods. A captious nicety of argument: The philosophy of w.v. quine. In Lewis Edwin Hahn and Paul Arthur Schilpp, editors, The Library of Living Philosophers Volume *XVIII*, pages 687–727. Chicago and LaSalle, IL: Open Court, expanded edition, 1998.

[Woods, 1999] John Woods. Aristotle. *Argumentation*, 13:203–220, 1999. [File of Fallacies] Reprinted in [Woods, 2004].

[Woods, 2001] John Woods. *Aristotle's Earlier Logic*. Oxford: Hermes Science Publications, 2001.

[Woods, 2003] John Woods. *Paradox and Paraconsistency: Conflict Resolution in the Abstract Sciences*. Cambridge and New York: Cambridge University Press, 2003.

[Woods, 2004] John Woods. *The Death of Argument: Fallacies in Agent-Based Reasoning*. Dordrecht and Boston: Kluwer, 2004.

[Woods, 2005a] John Woods. Cognitive yearning and the fugitivity of truth. In Kent Peacock and Andrew Irvine, editors, *Mistakes of Reason: Essays in Honour of John Woods*. Toronto, ON: University of Toronto Press, 2005. to appear in 2005.

[Woods, 2005b] John Woods. Dialectical considerations on the logic of contradiction. *Logic Journal of the IGPL*, 2005. To appear.

[Woodworth and Sells, 1935] R.S. Woodworth and S.B. Sells. An atmosphere effect in formal syllogistic-reasoning. *Journal of Experimental Psychology*, 18:451–460, 1935.

[Wouters, 1999] A.G. Wouters. *Explanation Without a Cause*. Utrecht: Zeno, 1999.

[Wright, 1973] L. Wright. Functions. *The Philosophical Review*, 82:139–168, 1973.

[Wright, 1976] L. Wright. *Teleological Explanations: An Etiological Analysis of Goals and Functions*. Berkeley, CA: University of California Press, 1976.

[Ziff, 1960] P. Ziff. *Semantic Analysis*. Ithaca, NY: Cornell University Press, 1960.

[Zimmermann, 1989] Manfred Zimmermann. The nervous system and the context of information theory. In R.F. Schmidt and G. Thews, editors, *Human Physiology*, pages 166–175. Berlin: Springer-Verlag, 2nd edition, 1989. Marguerite A. Biederman-Thorson, translator.

[Zipf, 1949] George K. Zipf. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison Wesley, 1949.

Index

ELSEVIER

A Practical Logic of Cognitive Systems

VOLUME 2

The Reach of Abduction

Insight and Trial

BY
DOV M. GABBAY AND JOHN WOODS