PROBABILITY THEORY
THE LOGIC OF SCIENCE

E. T. Jaynes

edited by G. Larry Bretthorst

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

© E. T. Jaynes 2003

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10/13 pt *System* LaTeX 2ε [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Jaynes, E. T. (Edwin T.)
Probability theory: the logic of science / by E.T. Jaynes; edited by G. Larry Bretthorst.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 59271 2
1. Probabilities. I. Bretthorst, G. Larry. II. Title.
QA273 .J36 2003
519.2 – dc21 2002071486

ISBN 0 521 59271 2 hardback

Contents

Editor's foreword *page* xvii
Preface xix

Part I Principles and elementary applications

1 Plausible reasoning 3
  1.1 Deductive and plausible reasoning 3
  1.2 Analogies with physical theories 6
  1.3 The thinking computer 7
  1.4 Introducing the robot 8
  1.5 Boolean algebra 9
  1.6 Adequate sets of operations 12
  1.7 The basic desiderata 17
  1.8 Comments 19
    1.8.1 Common language vs. formal logic 21
    1.8.2 Nitpicking 23
2 The quantitative rules 24
  2.1 The product rule 24
  2.2 The sum rule 30
  2.3 Qualitative properties 35
  2.4 Numerical values 37
  2.5 Notation and finite-sets policy 43
  2.6 Comments 44
    2.6.1 'Subjective' vs. 'objective' 44
    2.6.2 Gödel's theorem 45
    2.6.3 Venn diagrams 47
    2.6.4 The 'Kolmogorov axioms' 49
3 Elementary sampling theory 51
  3.1 Sampling without replacement 52
  3.2 Logic vs. propensity 60
  3.3 Reasoning from less precise information 64
  3.4 Expectations 66
  3.5 Other forms and extensions 68

3.6 Probability as a mathematical tool 68
3.7 The binomial distribution 69
3.8 Sampling with replacement 72
3.8.1 Digression: a sermon on reality vs. models 73
3.9 Correction for correlations 75
3.10 Simplification 81
3.11 Comments 82
3.11.1 A look ahead 84
4 Elementary hypothesis testing 86
4.1 Prior probabilities 87
4.2 Testing binary hypotheses with binary data 90
4.3 Nonextensibility beyond the binary case 97
4.4 Multiple hypothesis testing 98
4.4.1 Digression on another derivation 101
4.5 Continuous probability distribution functions 107
4.6 Testing an infinite number of hypotheses 109
4.6.1 Historical digression 112
4.7 Simple and compound (or composite) hypotheses 115
4.8 Comments 116
4.8.1 Etymology 116
4.8.2 What have we accomplished? 117
5 Queer uses for probability theory 119
5.1 Extrasensory perception 119
5.2 Mrs Stewart's telepathic powers 120
5.2.1 Digression on the normal approximation 122
5.2.2 Back to Mrs Stewart 122
5.3 Converging and diverging views 126
5.4 Visual perception – evolution into Bayesianity? 132
5.5 The discovery of Neptune 133
5.5.1 Digression on alternative hypotheses 135
5.5.2 Back to Newton 137
5.6 Horse racing and weather forecasting 140
5.6.1 Discussion 142
5.7 Paradoxes of intuition 143
5.8 Bayesian jurisprudence 144
5.9 Comments 146
5.9.1 What is queer? 148
6 Elementary parameter estimation 149
6.1 Inversion of the urn distributions 149
6.2 Both *N* and *R* unknown 150
6.3 Uniform prior 152
6.4 Predictive distributions 154

6.5 Truncated uniform priors 157
6.6 A concave prior 158
6.7 The binomial monkey prior 160
6.8 Metamorphosis into continuous parameter estimation 163
6.9 Estimation with a binomial sampling distribution 163
  6.9.1 Digression on optional stopping 166
6.10 Compound estimation problems 167
6.11 A simple Bayesian estimate: quantitative prior information 168
  6.11.1 From posterior distribution function to estimate 172
6.12 Effects of qualitative prior information 177
6.13 Choice of a prior 178
6.14 On with the calculation! 179
6.15 The Jeffreys prior 181
6.16 The point of it all 183
6.17 Interval estimation 186
6.18 Calculation of variance 186
6.19 Generalization and asymptotic forms 188
6.20 Rectangular sampling distribution 190
6.21 Small samples 192
6.22 Mathematical trickery 193
6.23 Comments 195
7 The central, Gaussian or normal distribution 198
7.1 The gravitating phenomenon 199
7.2 The Herschel–Maxwell derivation 200
7.3 The Gauss derivation 202
7.4 Historical importance of Gauss’s result 203
7.5 The Landon derivation 205
7.6 Why the ubiquitous use of Gaussian distributions? 207
7.7 Why the ubiquitous success? 210
7.8 What estimator should we use? 211
7.9 Error cancellation 213
7.10 The near irrelevance of sampling frequency distributions 215
7.11 The remarkable efficiency of information transfer 216
7.12 Other sampling distributions 218
7.13 Nuisance parameters as safety devices 219
7.14 More general properties 220
7.15 Convolution of Gaussians 221
7.16 The central limit theorem 222
7.17 Accuracy of computations 224
7.18 Galton’s discovery 227
7.19 Population dynamics and Darwinian evolution 229
7.20 Evolution of humming-birds and flowers 231

7.21 Application to economics 233
7.22 The great inequality of Jupiter and Saturn 234
7.23 Resolution of distributions into Gaussians 235
7.24 Hermite polynomial solutions 236
7.25 Fourier transform relations 238
7.26 There is hope after all 239
7.27 Comments 240
7.27.1 Terminology again 240
8 Sufficiency, ancillarity, and all that 243
8.1 Sufficiency 243
8.2 Fisher sufficiency 245
8.2.1 Examples 246
8.2.2 The Blackwell–Rao theorem 247
8.3 Generalized sufficiency 248
8.4 Sufficiency plus nuisance parameters 249
8.5 The likelihood principle 250
8.6 Ancillarity 253
8.7 Generalized ancillary information 254
8.8 Asymptotic likelihood: Fisher information 256
8.9 Combining evidence from different sources 257
8.10 Pooling the data 260
8.10.1 Fine-grained propositions 261
8.11 Sam’s broken thermometer 262
8.12 Comments 264
8.12.1 The fallacy of sample re-use 264
8.12.2 A folk theorem 266
8.12.3 Effect of prior information 267
8.12.4 Clever tricks and gamesmanship 267
9 Repetitive experiments: probability and frequency 270
9.1 Physical experiments 271
9.2 The poorly informed robot 274
9.3 Induction 276
9.4 Are there general inductive rules? 277
9.5 Multiplicity factors 280
9.6 Partition function algorithms 281
9.6.1 Solution by inspection 282
9.7 Entropy algorithms 285
9.8 Another way of looking at it 289
9.9 Entropy maximization 290
9.10 Probability and frequency 292
9.11 Significance tests 293
9.11.1 Implied alternatives 296

9.12 Comparison of psi and chi-squared 300
9.13 The chi-squared test 302
9.14 Generalization 304
9.15 Halley's mortality table 305
9.16 Comments 310
9.16.1 The irrationalists 310
9.16.2 Superstitions 312
10 Physics of 'random experiments' 314
10.1 An interesting correlation 314
10.2 Historical background 315
10.3 How to cheat at coin and die tossing 317
10.3.1 Experimental evidence 320
10.4 Bridge hands 321
10.5 General random experiments 324
10.6 Induction revisited 326
10.7 But what about quantum theory? 327
10.8 Mechanics under the clouds 329
10.9 More on coins and symmetry 331
10.10 Independence of tosses 335
10.11 The arrogance of the uninformed 338

Part II Advanced applications

11 Discrete prior probabilities: the entropy principle 343
11.1 A new kind of prior information 343
11.2 Minimum $\sum p_i^2$ 345
11.3 Entropy: Shannon's theorem 346
11.4 The Wallis derivation 351
11.5 An example 354
11.6 Generalization: a more rigorous proof 355
11.7 Formal properties of maximum entropy distributions 358
11.8 Conceptual problems – frequency correspondence 365
11.9 Comments 370
12 Ignorance priors and transformation groups 372
12.1 What are we trying to do? 372
12.2 Ignorance priors 374
12.3 Continuous distributions 374
12.4 Transformation groups 378
12.4.1 Location and scale parameters 378
12.4.2 A Poisson rate 382
12.4.3 Unknown probability for success 382
12.4.4 Bertrand's problem 386
12.5 Comments 394

13 Decision theory, historical background 397
- 13.1 Inference vs. decision 397
- 13.2 Daniel Bernoulli's suggestion 398
- 13.3 The rationale of insurance 400
- 13.4 Entropy and utility 402
- 13.5 The honest weatherman 402
- 13.6 Reactions to Daniel Bernoulli and Laplace 404
- 13.7 Wald's decision theory 406
- 13.8 Parameter estimation for minimum loss 410
- 13.9 Reformulation of the problem 412
- 13.10 Effect of varying loss functions 415
- 13.11 General decision theory 417
- 13.12 Comments 418
  - 13.12.1 'Objectivity' of decision theory 418
  - 13.12.2 Loss functions in human society 421
  - 13.12.3 A new look at the Jeffreys prior 423
  - 13.12.4 Decision theory is not fundamental 423
  - 13.12.5 Another dimension? 424

14 Simple applications of decision theory 426
- 14.1 Definitions and preliminaries 426
- 14.2 Sufficiency and information 428
- 14.3 Loss functions and criteria of optimum performance 430
- 14.4 A discrete example 432
- 14.5 How would our robot do it? 437
- 14.6 Historical remarks 438
  - 14.6.1 The classical matched filter 439
- 14.7 The widget problem 440
  - 14.7.1 Solution for Stage 2 443
  - 14.7.2 Solution for Stage 3 445
  - 14.7.3 Solution for Stage 4 449
- 14.8 Comments 450

15 Paradoxes of probability theory 451
- 15.1 How do paradoxes survive and grow? 451
- 15.2 Summing a series the easy way 452
- 15.3 Nonconglomerability 453
- 15.4 The tumbling tetrahedra 456
- 15.5 Solution for a finite number of tosses 459
- 15.6 Finite vs. countable additivity 464
- 15.7 The Borel–Kolmogorov paradox 467
- 15.8 The marginalization paradox 470
  - 15.8.1 On to greater disasters 474

15.9 Discussion 478
  15.9.1 The DSZ Example #5 480
  15.9.2 Summary 483
15.10 A useful result after all? 484
15.11 How to mass-produce paradoxes 485
15.12 Comments 486
16 Orthodox methods: historical background 490
16.1 The early problems 490
16.2 Sociology of orthodox statistics 492
16.3 Ronald Fisher, Harold Jeffreys, and Jerzy Neyman 493
16.4 Pre-data and post-data considerations 499
16.5 The sampling distribution for an estimator 500
16.6 Pro-causal and anti-causal bias 503
16.7 What is real, the probability or the phenomenon? 505
16.8 Comments 506
  16.8.1 Communication difficulties 507
17 Principles and pathology of orthodox statistics 509
17.1 Information loss 510
17.2 Unbiased estimators 511
17.3 Pathology of an unbiased estimate 516
17.4 The fundamental inequality of the sampling variance 518
17.5 Periodicity: the weather in Central Park 520
  17.5.1 The folly of pre-filtering data 521
17.6 A Bayesian analysis 527
17.7 The folly of randomization 531
17.8 Fisher: common sense at Rothamsted 532
  17.8.1 The Bayesian safety device 532
17.9 Missing data 533
17.10 Trend and seasonality in time series 534
  17.10.1 Orthodox methods 535
  17.10.2 The Bayesian method 536
  17.10.3 Comparison of Bayesian and orthodox estimates 540
  17.10.4 An improved orthodox estimate 541
  17.10.5 The orthodox criterion of performance 544
17.11 The general case 545
17.12 Comments 550
18 The A_p distribution and rule of succession 553
18.1 Memory storage for old robots 553
18.2 Relevance 555
18.3 A surprising consequence 557
18.4 Outer and inner robots 559

18.5 An application 561
18.6 Laplace's rule of succession 563
18.7 Jeffreys' objection 566
18.8 Bass or carp? 567
18.9 So where does this leave the rule? 568
18.10 Generalization 568
18.11 Confirmation and weight of evidence 571
  18.11.1 Is indifference based on knowledge or ignorance? 573
18.12 Carnap's inductive methods 574
18.13 Probability and frequency in exchangeable sequences 576
18.14 Prediction of frequencies 576
18.15 One-dimensional neutron multiplication 579
  18.15.1 The frequentist solution 579
  18.15.2 The Laplace solution 581
18.16 The de Finetti theorem 586
18.17 Comments 588
19 Physical measurements 589
19.1 Reduction of equations of condition 589
19.2 Reformulation as a decision problem 592
  19.2.1 Sermon on Gaussian error distributions 592
19.3 The underdetermined case: K is singular 594
19.4 The overdetermined case: K can be made nonsingular 595
19.5 Numerical evaluation of the result 596
19.6 Accuracy of the estimates 597
19.7 Comments 599
  19.7.1 A paradox 599
20 Model comparison 601
20.1 Formulation of the problem 602
20.2 The fair judge and the cruel realist 603
  20.2.1 Parameters known in advance 604
  20.2.2 Parameters unknown 604
20.3 But where is the idea of simplicity? 605
20.4 An example: linear response models 607
  20.4.1 Digression: the old sermon still another time 608
20.5 Comments 613
  20.5.1 Final causes 614
21 Outliers and robustness 615
21.1 The experimenter's dilemma 615
21.2 Robustness 617
21.3 The two-model model 619
21.4 Exchangeable selection 620
21.5 The general Bayesian solution 622

21.6 Pure outliers 624
21.7 One receding datum 625
22 Introduction to communication theory 627
22.1 Origins of the theory 627
22.2 The noiseless channel 628
22.3 The information source 634
22.4 Does the English language have statistical properties? 636
22.5 Optimum encoding: letter frequencies known 638
22.6 Better encoding from knowledge of digram frequencies 641
22.7 Relation to a stochastic model 644
22.8 The noisy channel 648
Appendix A Other approaches to probability theory 651
A.1 The Kolmogorov system of probability 651
A.2 The de Finetti system of probability 655
A.3 Comparative probability 656
A.4 Holdouts against universal comparability 658
A.5 Speculations about lattice theories 659
Appendix B Mathematical formalities and style 661
B.1 Notation and logical hierarchy 661
B.2 Our ‘cautious approach’ policy 662
B.3 Willy Feller on measure theory 663
B.4 Kronecker vs. Weierstrasz 665
B.5 What is a legitimate mathematical function? 666
B.5.1 Delta-functions 668
B.5.2 Nondifferentiable functions 668
B.5.3 Bogus nondifferentiable functions 669
B.6 Counting infinite sets? 671
B.7 The Hausdorff sphere paradox and mathematical diseases 672
B.8 What am I supposed to publish? 674
B.9 Mathematical courtesy 675
Appendix C Convolutions and cumulants 677
C.1 Relation of cumulants and moments 679
C.2 Examples 680
References 683
Bibliography 705
Author index 721
Subject index 724

1

Plausible reasoning

> The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.
>
> *James Clerk Maxwell (1850)*

Suppose some dark night a policeman walks down a street, apparently deserted. Suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a gentleman wearing a mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman doesn't hesitate at all in deciding that this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion? Let us first take a leisurely look at the general nature of such problems.

1.1 Deductive and plausible reasoning

A moment's thought makes it clear that our policeman's conclusion was not a logical deduction from the evidence; for there may have been a perfectly innocent explanation for everything. It might be, for example, that this gentleman was the owner of the jewelry store and he was coming home from a masquerade party, and didn't have the key with him. However, just as he walked by his store, a passing truck threw a stone through the window, and he was only protecting his own property.

Now, while the policeman's reasoning process was not logical deduction, we will grant that it had a certain degree of validity. The evidence did not make the gentleman's dishonesty *certain*, but it did make it extremely *plausible*. This is an example of a kind of reasoning in which we have all become more or less proficient, necessarily, long before studying mathematical theories. We are hardly able to get through one waking hour without facing some situation (e.g. will it rain or won't it?) where we do not have enough information to permit deductive reasoning; but still we must decide immediately what to do.

In spite of its familiarity, the formation of plausible conclusions is a very subtle process. Although history records discussions of it extending over 24 centuries, probably nobody has ever produced an analysis of the process which anyone else finds completely satisfactory. In this work we will be able to report some useful and encouraging new progress, in which conflicting intuitive judgments are replaced by definite theorems, and *ad hoc* procedures are replaced by rules that are determined uniquely by some very elementary – and nearly inescapable – criteria of rationality.

All discussions of these questions start by giving examples of the contrast between deductive reasoning and plausible reasoning. As is generally credited to the *Organon* of Aristotle (fourth century BC)[1] deductive reasoning (*apodeixis*) can be analyzed ultimately into the repeated application of two strong syllogisms:

$$\begin{array}{c} \text{if } A \text{ is true, then } B \text{ is true} \\ A \text{ is true} \\ \hline \text{therefore, } B \text{ is true,} \end{array} \tag{1.1}$$

and its inverse:

$$\begin{array}{c} \text{if } A \text{ is true, then } B \text{ is true} \\ B \text{ is false} \\ \hline \text{therefore, } A \text{ is false.} \end{array} \tag{1.2}$$

This is the kind of reasoning we would like to use all the time; but, as noted, in almost all the situations confronting us we do not have the right kind of information to allow this kind of reasoning. We fall back on weaker syllogisms (*epagoge*):

$$\begin{array}{c} \text{if } A \text{ is true, then } B \text{ is true} \\ B \text{ is true} \\ \hline \text{therefore, } A \text{ becomes more plausible.} \end{array} \tag{1.3}$$

The evidence does not prove that A is true, but verification of one of its consequences does give us more confidence in A. For example, let

$$\begin{aligned} A &\equiv \text{it will start to rain by 10 AM at the latest;} \\ B &\equiv \text{the sky will become cloudy before 10 AM.} \end{aligned}$$

Observing clouds at 9:45 AM does not give us a logical certainty that the rain will follow; nevertheless our common sense, obeying the weak syllogism, may induce us to change our plans and behave *as if* we believed that it will, if those clouds are sufficiently dark.

This example shows also that the major premise, 'if A then B' expresses B only as a *logical* consequence of A; and not necessarily a causal physical consequence, which could be effective only at a later time. The rain at 10 AM is not the physical cause of the clouds at

---

[1] Today, several different views are held about the exact nature of Aristotle's contribution. Such issues are irrelevant to our present purpose, but the interested reader may find an extensive discussion of them in Lukasiewicz (1957).

9:45 AM. Nevertheless, the proper logical connection is not in the uncertain causal direction (clouds $\Longrightarrow$ rain), but rather (rain $\Longrightarrow$ clouds), which is certain, although noncausal.

We emphasize at the outset that we are concerned here with *logical* connections, because some discussions and applications of inference have fallen into serious error through failure to see the distinction between logical implication and physical causation. The distinction is analyzed in some depth by Simon and Rescher (1966), who note that all attempts to interpret implication as expressing physical causation founder on the lack of contraposition expressed by the second syllogism (1.2). That is, if we tried to interpret the major premise as 'A is the physical cause of B', then we would hardly be able to accept that 'not-B is the physical cause of not-A'. In Chapter 3 we shall see that attempts to interpret plausible inferences in terms of physical causation fare no better.

Another weak syllogism, still using the same major premise, is

$$\begin{array}{c} \text{If } A \text{ is true, then } B \text{ is true} \\ A \text{ is false} \\ \hline \text{therefore, } B \text{ becomes less plausible.} \end{array} \tag{1.4}$$

In this case, the evidence does not prove that B is false; but one of the possible reasons for its being true has been eliminated, and so we feel less confident about B. The reasoning of a scientist, by which he accepts or rejects his theories, consists almost entirely of syllogisms of the second and third kind.

Now, the reasoning of our policeman was not even of the above types. It is best described by a still weaker syllogism:

$$\begin{array}{c} \text{If } A \text{ is true, then } B \text{ becomes more plausible} \\ B \text{ is true} \\ \hline \text{therefore, } A \text{ becomes more plausible.} \end{array} \tag{1.5}$$

But in spite of the apparent weakness of this argument, when stated abstractly in terms of A and B, we recognize that the policeman's conclusion has a very strong convincing power. There is something which makes us believe that, in this particular case, his argument had almost the power of deductive reasoning.

These examples show that the brain, in doing plausible reasoning, not only decides whether something becomes more plausible or less plausible, but that it evaluates the *degree* of plausibility in some way. The plausibility for rain by 10 AM depends very much on the darkness of those clouds at 9:45. And the brain also makes use of old information as well as the specific new data of the problem; in deciding what to do we try to recall our past experience with clouds and rain, and what the weatherman predicted last night.

To illustrate that the policeman was also making use of the past experience of policemen in general, we have only to change that experience. Suppose that events like these happened several times every night to every policeman – and that in every case the gentleman turned out to be completely innocent. Very soon, policemen would learn to ignore such trivial things.

Thus, in our reasoning we depend very much on *prior information* to help us in evaluating the degree of plausibility in a new problem. This reasoning process goes on unconsciously, almost instantaneously, and we conceal how complicated it really is by calling it *common sense*.

The mathematician George Pólya (1945, 1954) wrote three books about plausible reasoning, pointing out a wealth of interesting examples and showing that there are definite rules by which we do plausible reasoning (although in his work they remain in qualitative form). The above weak syllogisms appear in his third volume. The reader is strongly urged to consult Pólya's exposition, which was the original source of many of the ideas underlying the present work. We show below how Pólya's principles may be made quantitative, with resulting useful applications.

Evidently, the deductive reasoning described above has the property that we can go through long chains of reasoning of the type (1.1) and (1.2) and the conclusions have just as much certainty as the premises. With the other kinds of reasoning, (1.3)–(1.5), the reliability of the conclusion changes as we go through several stages. But in their quantitative form we shall find that in many cases our conclusions can still approach the certainty of deductive reasoning (as the example of the policeman leads us to expect). Pólya showed that even a pure mathematician actually uses these weaker forms of reasoning most of the time. Of course, on publishing a new theorem, the mathematician will try very hard to invent an argument which uses only the first kind; but the reasoning process which led to the theorem in the first place almost always involves one of the weaker forms (based, for example, on following up conjectures suggested by analogies). The same idea is expressed in a remark of S. Banach (quoted by S. Ulam, 1957):

> Good mathematicians see analogies between theorems; great mathematicians see analogies between analogies.

As a first orientation, then, let us note some very suggestive analogies to another field – which is itself based, in the last analysis, on plausible reasoning.

1.2 Analogies with physical theories

In physics, we learn quickly that the world is too complicated for us to analyze it all at once. We can make progress only if we dissect it into little pieces and study them separately. Sometimes, we can invent a mathematical model which reproduces several features of one of these pieces, and whenever this happens we feel that progress has been made. These models are called *physical theories*. As knowledge advances, we are able to invent better and better models, which reproduce more and more features of the real world, more and more accurately. Nobody knows whether there is some natural end to this process, or whether it will go on indefinitely.

In trying to understand common sense, we shall take a similar course. We won't try to understand it all at once, but we shall feel that progress has been made if we are able to construct idealized mathematical models which reproduce a few of its features. We expect that any model we are now able to construct will be replaced by more complete ones in the future, and we do not know whether there is any natural end to this process.

The analogy with physical theories is deeper than a mere analogy of method. Often, the things which are most familiar to us turn out to be the hardest to understand. Phenomena whose very existence is unknown to the vast majority of the human race (such as the difference in ultraviolet spectra of iron and nickel) can be explained in exhaustive mathematical detail – but all of modern science is practically helpless when faced with the complications of such a commonplace fact as growth of a blade of grass. Accordingly, we must not expect too much of our models; we must be prepared to find that some of the most familiar features of mental activity may be ones for which we have the greatest difficulty in constructing any adequate model.

There are many more analogies. In physics we are accustomed to finding that any advance in knowledge leads to consequences of great practical value, but of an unpredictable nature. Röntgen's discovery of X-rays led to important new possibilities of medical diagnosis; Maxwell's discovery of one more term in the equation for curl H led to practically instantaneous communication all over the earth.

Our mathematical models for common sense also exhibit this feature of practical usefulness. Any successful model, even though it may reproduce only a few features of common sense, will prove to be a powerful extension of common sense in some field of application. Within this field, it enables us to solve problems of inference which are so involved in complicated detail that we would never attempt to solve them without its help.

1.3 The thinking computer

Models have practical uses of a quite different type. Many people are fond of saying, 'They will never make a machine to replace the human mind – it does many things which no machine could ever do.' A beautiful answer to this was given by J. von Neumann in a talk on computers given in Princeton in 1948, which the writer was privileged to attend. In reply to the canonical question from the audience ('But of course, a mere machine can't really *think*, can it?'), he said:

> You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that!

In principle, the only operations which a machine cannot perform for us are those which we cannot describe in detail, or which could not be completed in a finite number of steps. Of course, some will conjure up images of Gödel incompleteness, undecidability, Turing machines which never stop, etc. But to answer all such doubts we need only point to the existence of the human brain, which *does* it. Just as von Neumann indicated, the only real limitations on making ‘machines which think’ are our own limitations in not knowing exactly what ‘thinking’ consists of.

But in our study of common sense we shall be led to some very explicit ideas about the mechanism of thinking. Every time we can construct a mathematical model which reproduces a part of common sense by prescribing a definite set of operations, this shows us how to ‘build a machine’, (i.e. write a computer program) which operates on incomplete information and, by applying quantitative versions of the above weak syllogisms, does plausible reasoning instead of deductive reasoning.

Indeed, the development of such computer software for certain specialized problems of inference is one of the most active and useful current trends in this field. One kind of problem thus dealt with might be: given a mass of data, comprising 10 000 separate observations, determine in the light of these data and whatever prior information is at hand, the relative plausibilities of 100 different possible hypotheses about the causes at work.

Our unaided common sense might be adequate for deciding between two hypotheses whose consequences are very different; but, in dealing with 100 hypotheses which are not very different, we would be helpless without a computer *and* a well-developed mathematical theory that shows us how to program it. That is, what determines, in the policeman’s syllogism (1.5), whether the plausibility for A increases by a large amount, raising it almost to certainty; or only a negligibly small amount, making the data B almost irrelevant? The object of the present work is to develop the mathematical theory which answers such questions, in the greatest depth and generality now possible.

While we expect a mathematical theory to be useful in programming computers, the idea of a thinking computer is also helpful psychologically in developing the mathematical theory. The question of the reasoning process used by actual human brains is charged with emotion and grotesque misunderstandings. It is hardly possible to say anything about this without becoming involved in debates over issues that are not only undecidable in our present state of knowledge, but are irrelevant to our purpose here.

Obviously, the operation of real human brains is so complicated that we can make no pretense of explaining its mysteries; and in any event we are not trying to explain, much less reproduce, all the aberrations and inconsistencies of human brains. That is an interesting and important subject; but it is not the subject we are studying here. Our topic is the *normative principles of logic*, and not the principles of psychology or neurophysiology.

To emphasize this, instead of asking, ‘How can we build a mathematical model of human common sense?’, let us ask, ‘How could we build a machine which would carry out useful plausible reasoning, following clearly defined principles expressing an idealized common sense?’

1.4 Introducing the robot

In order to direct attention to constructive things and away from controversial irrelevancies, we shall invent an imaginary being. Its brain is to be designed *by us*, so that it reasons according to certain definite rules. These rules will be deduced from simple desiderata which, it appears to us, would be desirable in human brains; i.e. we think that a rational person, on discovering that they were violating one of these desiderata, would wish to revise their thinking.

In principle, we are free to adopt any rules we please; that is our way of *defining* which robot we shall study. Comparing its reasoning with yours, if you find no resemblance you are in turn free to reject our robot and design a different one more to your liking. But if you find a very strong resemblance, and decide that you want and trust this robot to help you in your own problems of inference, then that will be an accomplishment of the theory, not a premise.

Our robot is going to reason about propositions. As already indicated above, we shall denote various propositions by italicized capital letters, $\{A, B, C, \text{etc.}\}$, and for the time being we must require that any proposition used must have, to the robot, an unambiguous meaning and must be of the simple, definite logical type that must be either true or false. That is, until otherwise stated, we shall be concerned only with two-valued logic, or Aristotelian logic. We do not require that the truth or falsity of such an 'Aristotelian proposition' be ascertainable by any feasible investigation; indeed, our inability to do this is usually just the reason why we need the robot's help. For example, the writer personally considers both of the following propositions to be true:

$$
\begin{aligned}
A &\equiv \text{Beethoven and Berlioz never met.} \\
B &\equiv \text{Beethoven's music has a better sustained quality than that of} \\
& \text{Berlioz, although Berlioz at his best is the equal of anybody.}
\end{aligned}
$$

Proposition B is not a permissible one for our robot to think about at present, whereas proposition A is, although it is unlikely that its truth or falsity could be definitely established today.[2] After our theory is developed, it will be of interest to see whether the present restriction to Aristotelian propositions such as A can be relaxed, so that the robot might help us also with more vague propositions such as B (see Chapter 18 on the A_p-distribution).[3]

1.5 Boolean algebra

To state these ideas more formally, we introduce some notation of the usual symbolic logic, or Boolean algebra, so called because George Boole (1854) introduced a *notation* similar to the following. Of course, the principles of deductive logic itself were well understood centuries before Boole, and, as we shall see, all the results that follow from Boolean algebra were contained already as special cases in the rules of plausible inference given by (1812). The symbol

---

[2] Their meeting is a chronological possibility, since their lives overlapped by 24 years; my reason for doubting it is the failure of Berlioz to mention any such meeting in his memoirs – on the other hand, neither does he come out and say definitely that they did *not* meet.

[3] The question of how one is to make a machine in some sense 'cognizant' of the conceptual meaning that a proposition like A has to humans, might seem very difficult, and much of the subject of artificial intelligence is devoted to inventing *ad hoc* devices to deal with this problem. However, we shall find in Chapter 4 that for us the problem is almost nonexistent; our rules for plausible reasoning automatically provide the means to do the mathematical equivalent of this.

$$AB, \tag{1.6}$$

called the *logical product* or the *conjunction*, denotes the proposition 'both A and B are true'. Obviously, the order in which we state them does not matter; AB and BA say the same thing. The expression

$$A + B, \tag{1.7}$$

called the *logical sum* or *disjunction*, stands for 'at least one of the propositions, A, B is true' and has the same meaning as $B + A$. These symbols are only a shorthand way of writing propositions, and do not stand for numerical values.

Given two propositions A, B, it may happen that one is true if and only if the other is true; we then say that they have the same *truth value*. This may be only a simple tautology (i.e. A and B are verbal statements which obviously say the same thing), or it may be that only after immense mathematical labor is it finally proved that A is the necessary and sufficient condition for B. From the standpoint of logic it does not matter; once it is established, by any means, that A and B have the same truth value, then they are logically equivalent propositions, in the sense that any evidence concerning the truth of one pertains equally well to the truth of the other, and they have the same implications for any further reasoning.

Evidently, then, it must be the most primitive axiom of plausible reasoning that two propositions with the same truth value are equally plausible. This might appear almost too trivial to mention, were it not for the fact that Boole himself (Boole, 1854, p. 286) fell into error on this point, by mistakenly identifying two propositions which were in fact different – and then failing to see any contradiction in their different plausibilities. Three years later, Boole (1857) gave a revised theory which supersedes that in his earlier book; for further comments on this incident, see Keynes (1921, pp. 167–168); Jaynes (1976, pp. 240–242).

In Boolean algebra, the equal sign is used to denote not equal numerical value, but equal truth value: $A = B$, and the 'equations' of Boolean algebra thus consist of assertions that the proposition on the left-hand side has the same truth value as the one on the right-hand side. The symbol '$\equiv$' means, as usual, 'equals by definition'.

In denoting complicated propositions we use parentheses in the same way as in ordinary algebra, i.e. to indicate the order in which propositions are to be combined (at times we shall use them also merely for clarity of expression although they are not strictly necessary). In their absence we observe the rules of algebraic hierarchy, familiar to those who use hand calculators: thus $AB + C$ denotes $(AB) + C$; and not $A(B + C)$.

The *denial* of a proposition is indicated by a bar:

$$\overline{A} \equiv A \text{ is false.} \tag{1.8}$$

The relation between A, $\overline{A}$ is a reciprocal one:

$$A = \overline{A} \text{ is false,} \tag{1.9}$$

and it does not matter which proposition we denote by the barred and which by the unbarred letter. Note that some care is needed in the unambiguous use of the bar. For example, according to the above conventions,

$$\overline{AB} = AB \text{ is false;} \tag{1.10}$$

$$\overline{A}\,\overline{B} = \text{both } A \text{ and } B \text{ are false.} \tag{1.11}$$

These are quite different propositions; in fact, $\overline{AB}$ is not the logical product $\overline{A}\,\overline{B}$, but the logical sum: $\overline{AB} = \overline{A} + \overline{B}$.

With these understandings, Boolean algebra is characterized by some rather trivial and obvious basic identities, which express the properties of:

$$\begin{array}{ll}
\text{Idempotence:} & \begin{cases} AA = A \\ A + A = A \end{cases} \\[2ex]
\text{Commutativity:} & \begin{cases} AB = BA \\ A + B = B + A \end{cases} \\[2ex]
\text{Associativity:} & \begin{cases} A(BC) = (AB)C = ABC \\ A + (B + C) = (A + B) + C = A + B + C \end{cases} \\[2ex]
\text{Distributivity:} & \begin{cases} A(B + C) = AB + AC \\ A + (BC) = (A + B)(A + C) \end{cases} \\[2ex]
\text{Duality:} & \begin{cases} \text{If } C = AB, \text{ then } \overline{C} = \overline{A} + \overline{B} \\ \text{If } D = A + B, \text{ then } \overline{D} = \overline{A}\,\overline{B} \end{cases}
\end{array} \tag{1.12}$$

but by their application one can prove any number of further relations, some highly nontrivial. For example, we shall presently have use for the rather elementary theorem:

$$\text{if } \overline{B} = AD \text{ then } A\overline{B} = \overline{B} \text{ and } B\overline{A} = \overline{A}. \tag{1.13}$$

Implication

The proposition

$$A \Rightarrow B \tag{1.14}$$

to be read as 'A implies B', does not assert that either A or B is true; it means only that $A\overline{B}$ is false, or, what is the same thing, $(\overline{A} + B)$ is true. This can be written also as the logical equation $A = AB$. That is, given (1.14), if A is true then B must be true; or, if B is false then A must be false. This is just what is stated in the strong syllogisms (1.1) and (1.2).

On the other hand, if A is false, (1.14) says nothing about B: and if B is true, (1.14) says nothing about A. But these are just the cases in which our weak syllogisms (1.3), (1.4) do say something. In one respect, then, the term 'weak syllogism' is misleading. The theory of plausible reasoning based on weak syllogisms is not a 'weakened' form of logic; it is an *extension* of logic with new content not present at all in conventional deductive logic. It will become clear in the next chapter (see (2.69) and (2.70)) that our rules include deductive logic as a special case.

A tricky point

Note carefully that in ordinary language one would take 'A implies B' to mean that B is logically deducible from A. But, in formal logic, 'A implies B' means only that the propositions A and AB have the same truth value. In general, whether B is logically deducible from A does not depend only on the propositions A and B; it depends on the totality of propositions $(A, A', A'', \ldots)$ that we accept as true and which are therefore available to use in the deduction. Devinatz (1968, p. 3) and Hamilton (1988, p. 5) give the truth table for the implication as a binary operation, illustrating that $A \Rightarrow B$ is false only if A is true and B is false; in all other cases $A \Rightarrow B$ is true!

This may seem startling at first glance; however, note that, indeed, if A and B are both true, then $A = AB$ and so $A \Rightarrow B$ is true; in formal logic every true statement implies every other true statement. On the other hand, if A is false, then AQ is also false for all Q, thus $A = AB$ and $A = A\overline{B}$ are both true, so $A \Rightarrow B$ and $A \Rightarrow \overline{B}$ are both true; a false proposition implies all propositions. If we tried to interpret this as logical deducibility (i.e. both B and $\overline{B}$ are deducible from A), it would follow that every false proposition is logically contradictory. Yet the proposition: 'Beethoven outlived Berlioz' is false but hardly logically contradictory (for Beethoven did outlive many people who were the same age as Berlioz).

Obviously, merely knowing that propositions A and B are both true does not provide enough information to decide whether either is logically deducible from the other, plus some unspecified 'toolbox' of other propositions. The question of logical deducibility of one proposition from a set of others arises in a crucial way in the Gödel theorem discussed at the end of Chapter 2. This great difference in the meaning of the word 'implies' in ordinary language and in formal logic is a tricky point that can lead to serious error if it is not properly understood; it appears to us that 'implication' is an unfortunate choice of word, and that this is not sufficiently emphasized in conventional expositions of logic.

1.6 Adequate sets of operations

We note some features of deductive logic which will be needed in the design of our robot. We have defined four operations, or 'connectives', by which, starting from two propositions A, B, other propositions may be defined: the logical product or conjunction AB, the logical sum or disjunction $A + B$, the implication $A \Rightarrow B$, and the negation $\overline{A}$. By combining these operations repeatedly in every possible way, one can generate any number of new propositions, such as

$$C \equiv (A + \overline{B})(\overline{A} + A\,\overline{B}) + \overline{A}\,B(A + B). \tag{1.15}$$

Many questions then occur to us: How large is the class of new propositions thus generated? Is it infinite, or is there a finite set that is closed under these operations? Can every proposition defined from $A,\ B$ be thus represented, or does this require further connectives beyond the above four? Or are these four already overcomplete so that some might be dispensed with? What is the smallest set of operations that is adequate to generate all such 'logic functions' of A and B? If instead of two starting propositions A, B we have an arbitrary number $\{A_1, \ldots, A_n\}$, is this set of operations still adequate to generate all possible logic functions of $\{A_1, \ldots, A_n\}$?

All these questions are answered easily, with results useful for logic, probability theory, and computer design. Broadly speaking, we are asking whether, starting from our present vantage point, we can (1) increase the number of functions, (2) decrease the number of operations. The first query is simplified by noting that two propositions, although they may appear entirely different when written out in the manner (1.15), are not different propositions from the standpoint of logic if they have the same truth value. For example, it is left for the reader to verify that C in (1.15) is logically the same statement as the implication $C = (B \Rightarrow \overline{A})$.

Since we are, at this stage, restricting our attention to Aristotelian propositions, any logic function $C = f(A, B)$ such as (1.15) has only two possible 'values', true and false; and likewise the 'independent variables' A and B can take on only those two values.

At this point, a logician might object to our notation, saying that the symbol A has been defined as standing for some fixed proposition, whose truth cannot change; so if we wish to consider logic functions, then instead of writing $C = f(A, B)$ we should introduce new symbols and write $z = f(x, y)$, where $x, y, z,$ are 'statement variables' for which various specific statements A, B, C may be substituted. But if A stands for some fixed but unspecified proposition, then it can still be either true or false. We achieve the same flexibility merely by the understanding that equations like (1.15) which define logic functions are to be true for all ways of defining A, B; i.e. instead of a statement variable we use a variable statement.

In relations of the form $C = f(A, B)$, we are concerned with logic functions defined on a discrete 'space' S consisting of only $2^2 = 4$ points; namely those at which A and B take on the 'values' {TT, TF, FT, FF}, respectively; and, at each point, the function $f(A, B)$ can take on independently either of two values {T, F}. There are, therefore, exactly $2^4 = 16$ different logic functions $f(A, B)$, and no more. An expression $B = f(A_1, \ldots, A_n)$ involving n propositions is a logic function on a space S of $M = 2^n$ points; and there are exactly 2^M such functions.

In the case $n = 1$, there are four logic functions $\{f_1(A), \ldots, f_4(A)\}$, which we can define by enumeration, listing all their possible values in a truth table:

A	T	F
$f_1(A)$	T	T
$f_2(A)$	T	F
$f_3(A)$	F	T
$f_4(A)$	F	F

But it is obvious by inspection that these are just

$$\begin{aligned} f_1(A) &= A + \overline{A} \\ f_2(A) &= A \\ f_3(A) &= \overline{A} \\ f_4(A) &= A\ \overline{A}, \end{aligned} \tag{1.16}$$

so we prove by enumeration that the three operations: conjunction, disjunction, and negation are adequate to generate all logic functions of a single proposition.

For the case of general n, consider first the special functions, each of which is true at one and only one point of S. For $n = 2$ there are $2^n = 4$ such functions,

A, B	TT	TF	FT	FF
$f_1(A, B)$	T	F	F	F
$f_2(A, B)$	F	T	F	F
$f_3(A, B)$	F	F	T	F
$f_4(A, B)$	F	F	F	T

It is clear by inspection that these are just the four basic conjunctions,

$$\begin{aligned} f_1(A, B) &= A\ B \\ f_2(A, B) &= A\ \overline{B} \\ f_3(A, B) &= \overline{A}\ B \\ f_4(A, B) &= \overline{A}\ \overline{B}. \end{aligned} \tag{1.17}$$

Consider now any logic function which is true on certain specified points of S; for example, $f_5(A, B)$ and $f_6(A, B)$, defined by

A, B	TT	TF	FT	FF
$f_5(A, B)$	F	T	F	T
$f_6(A, B)$	T	F	T	T

We assert that each of these functions is the logical sum of the conjunctions (1.17) that are true on the same points (this is not trivial; the reader should verify it in detail). Thus,

$$\begin{aligned} f_5(A, B) &= f_2(A, B) + f_4(A, B) \\ &= A\,\overline{B} + \overline{A}\,\overline{B} \\ &= (A + \overline{A})\,\overline{B} \\ &= \overline{B}, \end{aligned} \tag{1.18}$$

and, likewise,

$$\begin{aligned} f_6(A, B) &= f_1(A, B) + f_3(A, B) + f_4(A, B) \\ &= AB + \overline{A}\,B + \overline{A}\,\overline{B} \\ &= B + \overline{A}\,\overline{B} \\ &= \overline{A} + B. \end{aligned} \tag{1.19}$$

That is, $f_6(A, B)$ is the implication $f_6(A, B) = (A \Rightarrow B)$, with the truth table discussed above. Any logic function $f(A, B)$ that is true on at least one point of S can be constructed in this way as a logical sum of the basic conjunctions (1.17). There are $2^4 - 1 = 15$ such functions. For the remaining function, which is always false, it suffices to take the contradiction, $f_{16}(A, B) \equiv A\,\overline{A}$.

This method (called 'reduction to *disjunctive normal form*' in logic textbooks) will work for any n. For example, in the case $n = 5$ there are $2^5 = 32$ basic conjunctions,

$$\{ABCDE,\ ABCD\overline{E},\ ABC\overline{D}E, \ldots, \overline{A}\,\overline{B}\,\overline{C}\,\overline{D}\,\overline{E}\}, \tag{1.20}$$

and $2^{32} = 4\,294\,967\,296$ different logic functions $f_i(A, B, C, D, E)$; of which 4 294 967 295 can be written as logical sums of the basic conjunctions, leaving only the contradiction

$$f_{4294967296}(A, B, C, D, E) = A\,\overline{A}. \tag{1.21}$$

Thus one can verify by 'construction in thought' that the three operations

$$\{\text{conjunction, disjunction, negation}\}, \qquad \text{i.e.} \qquad \{\text{AND, OR, NOT}\}, \tag{1.22}$$

suffice to generate all possible logic functions; or, more concisely, they form an *adequate set*.

The duality property (1.12) shows that a smaller set will suffice; for disjunction of A, B is the same as denying that they are both false:

$$A + B = \overline{(\overline{A}\,\overline{B})}. \tag{1.23}$$

Therefore, the two operations (AND, NOT) already constitute an adequate set for deductive logic.[4] This fact will be essential in determining when we have an adequate set of rules for plausible reasoning; see Chapter 2.

---

[4] For you to ponder: Does it follow that these two commands are the only ones needed to write any computer program?

It is clear that we cannot now strike out either of these operations, leaving only the other; i.e. the operation ‘AND’ cannot be reduced to negations; and negation cannot be accomplished by any number of ‘AND’ operations. But this still leaves open the possibility that both conjunction and negation might be reducible to some third operation, not yet introduced, so that a single logic operation would constitute an adequate set.

It comes as a pleasant surprise to find that there is not only one but two such operations. The operation ‘NAND’ is defined as the negation of ‘AND’:

$$A \uparrow B \equiv \overline{AB} = \overline{A} + \overline{B} \tag{1.24}$$

which we can read as ‘A NAND B’. But then we have at once

$$\begin{aligned} \overline{A} &= A \uparrow A \\ AB &= (A \uparrow B) \uparrow (A \uparrow B) \\ A + B &= (A \uparrow A) \uparrow (B \uparrow B). \end{aligned} \tag{1.25}$$

Therefore, every logic function can be constructed with NAND alone. Likewise, the operation NOR defined by

$$A \downarrow B \equiv \overline{A + B} = \overline{A}\,\overline{B} \tag{1.26}$$

is also powerful enough to generate all logic functions:

$$\begin{aligned} \overline{A} &= A \downarrow A \\ A + B &= (A \downarrow B) \downarrow (A \downarrow B) \\ AB &= (A \downarrow A) \downarrow (B \downarrow B). \end{aligned} \tag{1.27}$$

One can take advantage of this in designing computer and logic circuits. A ‘logic gate’ is a circuit having, besides a common ground, two input terminals and one output. The voltage relative to ground at any of these terminals can take on only two values; say +3 volts, or ‘up’, representing ‘true’; and 0 volts or ‘down’, representing ‘false’. A NAND gate is thus one whose output is up if and only if at least one of the inputs is down; or, what is the same thing, down if and only if both inputs are up; while for a NOR gate the output is up if and only if both inputs are down.

One of the standard components of logic circuits is the ‘quad NAND gate’, an integrated circuit containing four independent NAND gates on one semiconductor chip. Given a sufficient number of these and no other circuit components, it is possible to generate any required logic function by interconnecting them in various ways.

This short excursion into deductive logic is as far as we need go for our purposes. Further developments are given in many textbooks; for example, a modern treatment of Aristotelian logic is given by Copi (1994). For non-Aristotelian forms with special emphasis on Gödel incompleteness, computability, decidability, Turing machines, etc., see Hamilton (1988).

We turn now to our extension of logic, which is to follow from the conditions discussed next. We call them ‘desiderata’ rather than ‘axioms’ because they do not assert that anything is ‘true’ but only state what appear to be desirable goals. Whether these goals are attainable without contradictions, and whether they determine any unique extension of logic, are matters of mathematical analysis, given in Chapter 2.

1.7 The basic desiderata

To each proposition about which it reasons, our robot must assign some degree of plausibility, based on the evidence we have given it; and whenever it receives new evidence it must revise these assignments to take that new evidence into account. In order that these plausibility assignments can be stored and modified in the circuits of its brain, they must be associated with some definite physical quantity, such as voltage or pulse duration or a binary coded number, etc. – however our engineers want to design the details. For present purposes, this means that there will have to be some kind of association between degrees of plausibility and real numbers:

(I) *Degrees of plausibility are represented by real numbers.* (1.28)

Desideratum (I) is practically forced on us by the requirement that the robot's brain must operate by the carrying out of some definite physical process. However, it will appear (Appendix A) that it is also required theoretically; we do not see the possibility of any consistent theory without a property that is equivalent functionally to desideratum (I).

We adopt a natural but nonessential convention: that a greater plausibility shall correspond to a greater number. It will also be convenient to assume a continuity property, which is hard to state precisely at this stage; to say it intuitively: an infinitesimally greater plausibility ought to correspond only to an infinitesimally greater number.

The plausibility that the robot assigns to some proposition A will, in general, depend on whether we told it that some other proposition B is true. Following the notation of Keynes (1921) and Cox (1961), we indicate this by the symbol

$$A|B, \tag{1.29}$$

which we may call 'the conditional plausibility that A is true, given that B is true' or just 'A given B'. It stands for some real number. Thus, for example,

$$A|BC \tag{1.30}$$

(which we may read as 'A given BC') represents the plausibility that A is true, given that both B and C are true. Or,

$$A + B|CD \tag{1.31}$$

represents the plausibility that at least one of the propositions A and B is true, given that both C and D are true; and so on. We have decided to represent a greater plausibility by a greater number, so

$$(A|B) > (C|B) \tag{1.32}$$

says that, given B, A is more plausible than C. In this notation, while the symbol for plausibility is just of the form $A|B$ without parentheses, we often add parentheses for clarity of expression. Thus, (1.32) says the same thing as

$$A|B > C|B, \tag{1.33}$$

but its meaning is clearer to the eye.

In the interest of avoiding impossible problems, we are not going to ask our robot to undergo the agony of reasoning from impossible or mutually contradictory premises; there could be no 'correct' answer. Thus, we make no attempt to define $A|BC$ when B and C are mutually contradictory. Whenever such a symbol appears, it is understood that B and C are compatible propositions.

Also, we do not want this robot to think in a way that is directly opposed to the way you and I think. So we shall design it to reason in a way that is at least *qualitatively* like the way humans try to reason, as described by the above weak syllogisms and a number of other similar ones.

Thus, if it has old information C which gets updated to C' in such a way that the plausibility for A is increased:

$$(A|C') > (A|C); \tag{1.34}$$

but the plausibility for B given A is not changed:

$$(B|AC') = (B|AC). \tag{1.35}$$

This can, of course, produce only an increase, never a decrease, in the plausibility that both A and B are true:

$$(AB|C') \geq (AB|C); \tag{1.36}$$

and it must produce a decrease in the plausibility that A is false:

$$(\overline{A}|C') < (\overline{A}|C). \tag{1.37}$$

This qualitative requirement simply gives the 'sense of direction' in which the robot's reasoning is to go; it says nothing about *how much* the plausibilities change, except that our continuity assumption (which is also a condition for qualitative correspondence with common sense) now requires that if $A|C$ changes only infinitesimally, it can induce only an infinitesimal change in $AB|C$ and $\overline{A}|C$. The specific ways in which we use these qualitative requirements will be given in the next chapter, at the point where it is seen why we need them. For the present we summarize them simply as:

(II) *Qualitative correspondence with common sense.* (1.38)

Finally, we want to give our robot another desirable property for which honest people strive without always attaining: that it always reasons *consistently*. By this we mean just the three common colloquial meanings of the word ‘consistent’:

(IIIa) *If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.* (1.39a)

(IIIb) *The robot always takes into account all of the evidence it has relevant to a question. It does not arbitrarily ignore some of the information, basing its conclusions only on what remains. In other words, the robot is completely nonideological.* (1.39b)

(IIIc) *The robot always represents equivalent states of knowledge by equivalent plausibility assignments. That is, if in two problems the robot’s state of knowledge is the same (except perhaps for the labeling of the propositions), then it must assign the same plausibilities in both.* (1.39c)

Desiderata (I), (II), and (IIIa) are the basic ‘structural’ requirements on the inner workings of our robot’s brain, while (IIIb) and (IIIc) are ‘interface’ conditions which show how the robot’s behavior should relate to the outer world.

At this point, most students are surprised to learn that our search for desiderata is at an end. The above conditions, it turns out, uniquely determine the rules by which our robot must reason; i.e. there is only one set of mathematical operations for manipulating plausibilities which has all these properties. These rules are deduced in Chapter 2.

(At the end of most chapters, we insert a section of informal Comments in which are collected various side remarks, background material, etc. The reader may skip them without losing the main thread of the argument.)

1.8 Comments

As politicians, advertisers, salesmen, and propagandists for various political, economic, moral, religious, psychic, environmental, dietary, and artistic doctrinaire positions know only too well, fallible human minds are easily tricked, by clever verbiage, into committing violations of the above desiderata. We shall try to ensure that they do not succeed with our robot.

We emphasize another contrast between the robot and a human brain. By Desideratum I, the robot’s mental state about any proposition is to be represented by a real number. Now, it is clear that our attitude toward any given proposition may have more than one ‘coordinate’. You and I form simultaneous judgments about a proposition not only as to whether it is plausible, but also whether it is desirable, whether it is important, whether it is useful, whether it is interesting, whether it is amusing, whether it is morally right, etc. If we assume that each of these judgments might be represented by a number, then a fully adequate description of a human state of mind would be represented by a vector in a space of a rather large number of dimensions.

Not all propositions require this. For example, the proposition ‘The refractive index of water is less than 1.3’ generates no emotions; consequently the state of mind which it produces has very few coordinates. On the other hand, the proposition, ‘Your mother-in-law just wrecked your new car’ generates a state of mind with many coordinates. Quite generally, the situations of everyday life are those involving many coordinates. It is just for this reason, we suggest, that the most familiar examples of mental activity are often the most difficult to reproduce by a model. Perhaps we have here the reason why science and mathematics are the most successful of human activities: they deal with propositions which produce the simplest of all mental states. Such states would be the ones least perturbed by a given amount of imperfection in the human mind.

Of course, for many purposes we would not want our robot to adopt any of these more ‘human’ features arising from the other coordinates. It is just the fact that computers do *not* get confused by emotional factors, do *not* get bored with a lengthy problem, do *not* pursue hidden motives opposed to ours, that makes them safer agents than men for carrying out certain tasks.

These remarks are interjected to point out that there is a large unexplored area of possible generalizations and extensions of the theory to be developed here; perhaps this may inspire others to try their hand at developing ‘multidimensional theories’ of mental activity, which would more and more resemble the behavior of actual human brains – not all of which is undesirable. Such a theory, if successful, might have an importance beyond our present ability to imagine.[5]

For the present, however, we shall have to be content with a much more modest undertaking. Is it possible to develop a consistent ‘one-dimensional’ model of plausible reasoning? Evidently, our problem will be simplest if we can manage to represent a degree of plausibility uniquely by a single real number, and ignore the other ‘coordinates’ just mentioned.

We stress that we are in no way asserting that degrees of plausibility in actual human minds have a unique numerical measure. Our job is not to postulate – or indeed to conjecture about – any such thing; it is to *investigate* whether it is possible, in our robot, to set up such a correspondence without contradictions.

But to some it may appear that we have already assumed more than is necessary, thereby putting gratuitous restrictions on the generality of our theory. Why must we represent degrees of plausibility by real numbers? Would not a ‘comparative’ theory based on a system of qualitative ordering relations such as $(A|C) > (B|C)$ suffice? This point is discussed further in Appendix A, where we describe other approaches to probability theory and note that some attempts have been made to develop comparative theories which it was thought would be logically simpler, or more general. But this turned out not to be the case; so, although it is quite possible to develop the foundations in other ways than ours, the final results will not be different.

---

[5] Indeed, some psychologists think that as few as five dimensions might suffice to characterize a human personality; that is, that we all differ only in having different mixes of five basic personality traits which may be genetically determined. But it seems to us that this must be grossly oversimplified; identifiable chemical factors continuously varying in both space and time (such as the distribution of glucose metabolism in the brain) affect mental activity but cannot be represented faithfully in a space of only five dimensions. Yet it may be that five numbers can capture enough of the truth to be useful for many purposes.

1.8.1 Common language vs. formal logic

We should note the distinction between the statements of formal logic and those of ordinary language. It might be thought that the latter is only a less precise form of expression; but on examination of details the relation appears different. It appears to us that ordinary language, carefully used, need not be less precise than formal logic; but ordinary language is more complicated in its rules and has consequently richer possibilities of expression than we allow ourselves in formal logic.

In particular, common language, being in constant use for other purposes than logic, has developed subtle nuances – means of implying something without actually stating it – that are lost on formal logic. Mr A, to affirm his objectivity, says, 'I believe what I see.' Mr B retorts: 'He doesn't see what he doesn't believe.' From the standpoint of formal logic, it appears that they have said the same thing; yet from the standpoint of common language, those statements had the intent and effect of conveying opposite meanings.

Here is a less trivial example, taken from a mathematics textbook. Let L be a straight line in a plane, and S an infinite set of points in that plane, each of which is projected onto L. Now consider the following statements:

(I) The projection of the limit is the limit of the projections.
(II) The limit of the projections is the projection of the limit.

These have the grammatical structures 'A is B' and 'B is A', and so they might appear logically equivalent. Yet in that textbook, (I) was held to be true, and (II) not true in general, on the grounds that the limit of the projections may exist when the limit of the set does not.

As we see from this, in common language – even in mathematics textbooks – we have learned to read subtle nuances of meaning into the exact phrasing, probably without realizing it until an example like this is pointed out. We interpret 'A is B' as asserting first of all, as a kind of major premise, that A exists; and the rest of the statement is understood to be conditional on that premise. Put differently, in common grammar the verb 'is' implies a distinction between subject and object, which the symbol '$=$' does not have in formal logic or in conventional mathematics. (However, in computer languages we encounter such statements as '$\mathrm{J} = \mathrm{J} + 1$', which everybody seems to understand, but in which the '$=$' sign has now acquired that implied distinction after all.)

Another amusing example is the old adage 'knowledge is power', which is a very cogent truth, both in human relations and in thermodynamics. An ad writer for a chemical trade journal[6] fouled this up into 'power is knowledge', an absurd – indeed, obscene – falsity.

These examples remind us that the verb 'is' has, like any other verb, a subject and a predicate; but it is seldom noted that this verb has two entirely different meanings. A person whose native language is English may require some effort to see the different meanings in the statements: 'The room is noisy' and 'There is noise in the room'. But in Turkish these meanings are rendered by different words, which makes the distinction so clear that a visitor who uses the wrong word will not be understood. The latter statement is ontological, asserting the physical existence of something, while the former is epistemological, expressing only the speaker's personal perception.

---

[6] *LC-CG Magazine*, March 1988, p. 211.

Common language – or, at least, the English language – has an almost universal tendency to disguise epistemological statements by putting them into a grammatical form which suggests to the unwary an ontological statement. A major source of error in current probability theory arises from an unthinking failure to perceive this. To interpret the first kind of statement in the ontological sense is to assert that one's own private thoughts and sensations are realities existing externally in Nature. We call this the 'mind projection fallacy', and note the trouble it causes many times in what follows. But this trouble is hardly confined to probability theory; as soon as it is pointed out, it becomes evident that much of the discourse of philosophers and Gestalt psychologists, and the attempts of physicists to explain quantum theory, are reduced to nonsense by the author falling repeatedly into the mind projection fallacy.

These examples illustrate the care that is needed when we try to translate the complex statements of common language into the simpler statements of formal logic. Of course, common language is often less precise than we should want in formal logic. But everybody expects this and is on the lookout for it, so it is less dangerous.

It is too much to expect that our robot will grasp all the subtle nuances of common language, which a human spends perhaps 20 years acquiring. In this respect, our robot will remain like a small child – it interprets all statements literally and blurts out the truth without thought of whom this may offend.

It is unclear to the writer how difficult – and even less clear how desirable – it would be to design a newer model robot with the ability to recognize these finer shades of meaning. Of course, the question of principle is disposed of at once by the existence of the human brain, which does this. But, in practice, von Neumann's principle applies; a robot designed by us cannot do it until someone develops a theory of 'nuance recognition', which reduces the process to a definitely prescribed set of operations. This we gladly leave to others.

In any event, our present model robot is quite literally real, because today it is almost universally true that any nontrivial probability evaluation is performed by a computer. The person who programmed that computer was necessarily, whether or not they thought of it that way, designing part of the brain of a robot according to some preconceived notion of how the robot should behave. But very few of the computer programs now in use satisfy all our desiderata; indeed, most are intuitive *ad hoc* procedures that were not chosen with any well-defined desiderata at all in mind.

Any such adhockery is presumably usable within some special area of application – that was the criterion for choosing it – but as the proofs of Chapter 2 will show, any adhockery which conflicts with the rules of probability theory must generate demonstrable inconsistencies when we try to apply it beyond some restricted area. Our aim is to avoid this by developing the general principles of inference once and for all, directly from the requirement of consistency, and in a form applicable to any problem of plausible inference that is formulated in a sufficiently unambiguous way.