The Material Theory of Induction

John D. Norton
Department of History and Philosophy of Science
University of Pittsburgh
http://www.pitt.edu/~jdnorton

April 24, 2020

Version
April 24, 2020
Minor typo corrections April 30, 2020

Preface

The project for this volume started modestly. It was classified as the "little induction book" in my original notes. The plan was to write a short and easy introduction to the main ideas of the material theory of induction. As the writing proceeded, those modest ambitions were supplanted by increasingly ambitious ones until the project had ballooned into something enormous. There were three parts. The first dealt with qualitative notions of inductive inference and the second with quantitative notions. They correspond roughly to the chapters 1-9 and 10-16 of the present work. There was no space for the third part that dealt with the global structure of inductive support. It will be the subject of another volume. Readers anxious for a taste of its content should consult the Epilog here.

The principal idea of the material theory of induction is that background facts obtaining in some domain tell us which are the good and which are the bad inductive inferences in that domain. This conception differs fundamentally from virtually all approaches to inductive inference in the present literature. There the good inductive inferences are distinguished from the bad by checking whether the inference has appropriate formal properties, such as fitting to an approved inferential template or preferred calculus. Because the divergence from the present literature occurs at such a fundamental level, my experience is that philosophers of science who work in inductive inference have trouble approaching the theory. The difficulty, I conjecture, is that we approach new ideas by trying to assimilate them into our existing conceptual system, which has in turn been tailored to the details of our own research agendas. What are we to do when an idea arrives that does not neatly fit into any of our existing conceptual pigeon holes? Is this material theory just another variant of enumerative induction? Is it inference to the best explanation with some alternative notion of explanation? Is it the proposal of a non-probabilistic, mathematical calculus of inductive inference? Or is it another tiresome skeptical assault on inductive inference and the evidential grounding of science?

It is none of these. The slogans "All induction is local." and "No universal rules of induction." may appear skeptical. They are not. They are an attempt to diagnose why inductive inference has, for thousands of years, been a locus of trouble for philosophers. The words "induction" and "problem" are nearest neighbors in any philosophical lexicon. This enduring, troublesome character derives, I believe, from a foundational mistake that was made at the outset. We tried and continue to try to understand inductive inference using the formal methods that have proven so fertile for deductive inference. While different formal approaches may work in

different domains, the formal conception is the wrong approach for understanding inductive inference overall. Choosing it is responsible for the enduring trouble. The material approach offers an alternative foundation for inductive inference that repairs the trouble.

A prominent corollary of the material approach is that probabilistic methods do not provide a universally applicable account of inductive inference. For those enamored by Bayesianism, it will be tempting to drop the material theory into the pigeon hole occupied by formal luddites whose opposition to all mathematical approaches is grounded in a visceral antipathy to them. I do not belong in that company, as readers will see if they consult Chapter 16. My work elsewhere in history and philosophy of physics is very hospitable to mathematical methods, whose power continues to astonish me. I am especially impressed with the power of probabilistic methods in statistical physics. When they are applicable, they are wonders.

My advocacy and defense of probabilistic approaches extends to inductive inference, but only on a case by case basis. When probabilistic methods are warranted in some domain, they work and they work very well. Where Bayesians err is in their belief that probabilistic methods are a universal default that can be applied everywhere, automatically. Instead, my view is that probabilistic methods can be applied only in some domain when the background facts of that domain authorize it. We cannot just assume that they apply in some new domain. We have a positive obligation to show that they are warranted by background facts in each case.

A consequence is that I wilt every time I see yet another paper that promises a Bayesian analysis of fiddle-de-dee, especially when fiddle-de-dee is some aspect of inductive inference or evidential support. The pretense is that the Bayesian analysis will provide universal understanding. It cannot do this since Bayesian analysis cannot be applied everywhere. Instead we are given a few elementary results in the probability calculus. The terms of these formulae are then matched tendentiously with terms of art from fiddle-de-dee. The relabeled formulae are supposed to provide insight, but they only give us the illusion of understanding.

The style of analysis of this work falls within my conception of history and philosophy of science. It begins by taking the pertinent science seriously. That is especially important when it comes to inductive inference since the evidential successes of modern science are extraordinary. That we philosophers of science are struggling to vindicate these successes is more a commentary on our failures than any failure of the sciences. The chapters that follow are rich in examples from science. I lean towards grasping the science by exploring its history, for an emphasis on the history provides some protection from the inevitable, modern textbook simplifications of relations of inductive support. The presence of the history is not mere decoration. It is essential to understanding of the evidential relations in the science.

It is customary in a preface to acknowledge those who have been helpful in the book's project. This project has many distinct parts, commonly divided naturally by chapter. Rather than

delivering here a long but opaque list of names, I have acknowledged in individual chapters those who have been especially helpful in those parts. Those acknowledgments fall short of naming all those who have provided support, encouragement or helpful critical responses. To all those I have failed to name, I offer apologies and thanks.

On October 27-28, 2018, there was a conference on the material theory of induction at the Center for Philosophy of Science, University of Pittsburgh, called "Norton for Everyone: The Material Theory of Induction and Beyond." It was beyond extraordinary and humbling for me to have the material theory of induction scrutinized by so many talented and accomplished philosophers of science. May I thank once again all those who participated? Its organizers were John Earman, Bryan W. Roberts and Elay Shech. Speakers and discussion leaders were Jonathan Bain, Nora Boyd, Jeremy Butterfield, Richard Dawid, Siska De Baerdemaeker, Balazs Gyenis, Eric Hatleback, Leah Henderson, Michel Janssen, Molly Kao, Jonathan Livengood, Wendy Parker, Dasha Pruss, Bryan W. Roberts, Elay Shech and David Wallace. Many more were present and contributed most valuably. With apologies to those omitted, my faulty memory means that this list is only partial: Harvey Brown, Hasok Chang, Pat Corvini, Nick Huggett, Shahin Kaveh, Edouard Machery, John McCaskey, Tom Pashby, Willy Penn, Mike Tamir, Jennifer Whyte and Jim Woodward.

Subsequently Elay Shech and Wendy Parker have solicited contributions from the speakers and elsewhere for a special issue on the material theory of induction in *Studies in History and Philosophy of Science*. At the time of this writing, many papers have been published in advance on the journal website; and there are more to come. Once again, I thank the contributors for their interest and efforts. I reserve special thanks for Elay and Wendy for having undertaken the burden of organizing this special issue and shepherding its contributions through to completion.

Finally I offer the most profound gratitude to my wife Eve who has provided a happy home for my body and heart through the years of the writing of this work and many before it. Those who know the joy of true and enduring love will understand what that means. No combination of words can properly express it.

Contents

November 26, 2017; June 12, 2018

Prolog

1. The Wonder of Science

Our best science tells us wonderful things. The cold and dark skies of our universe were not so long ago in their entirety in a state of unimaginably high energy and temperature. The detritus that exploded from it congealed into stars, planets and galaxies. These systems of celestial masses are in turn held together by a curvature of the geometry of space and time itself. On a most minute scale, the matter of these systems and the light they radiate consist of neither waves nor particles but a curious amalgam that is, at once, both and neither. The organisms that walk on one of these planets, complete with their intricate eyes and thinking brains, emerged incrementally from crude matter, in tiny steps over eons of time. They were shaped only by the fact that a small, random change in one organism might give it a slight advantage over its rivals. The design specification of these accumulated advantages is recorded and transmitted through the generations of the organisms by its encoding in hundreds of millions of base pairs of a chemical found in every cell of each organism.

These, and many more ideas of science like them, are extraordinary. Their contemplation must eventually overwhelm with wonder even the most curious and flexible of minds. Only the dullest of wit or the most soured of skeptics could resist their charms.

For me, there is a still greater wonder. These ideas are not the inventions of writers of myth and fiction. They could not be so, for their content far outstrips our meager human imaginations. Rather they are the result of careful, painstaking, systematic investigations of nature, guided solely by inventive insight and cautious reasoning. They are discoveries. When these efforts go past the early speculative stages and succeed, their products are distinguished by a special relation with what we experience of the world. Those experiences provide the inductive support for successful science. They tell us that this is how the world is.

The explosive expansion of the universe is supported by the reddening of light from distant galaxies. That the curvature of the geometry of space and time keeps the planets in their orbits is supported by the most delicate measurements of slight anomalies in planetary motions. The curious quantum nature of matter in the small is supported by how light from excited gases is concentrated into just a few quite specific frequencies. The evolution of humans from simpler organisms is supported by fossilized bones, whose chronology is recorded by their positions in

layers of rock strata. The double spiral geometry of the molecules of deoxyribonucleic acid is supported by the patterns formed when X-rays diffract off material extracted from the nuclei of cells.

In all this, the essential relation is inductive support. It obtains between the propositions of science and those that express the evidence on which it rests. It enables us to assign an authority to the ideas of science that no other narrative can match. Without it, science becomes just another "way of knowing," to use a popular oxymoron of the skeptics. Without this relation, we do not know anything of the world. We "know" but do not know. Without it, the ideas of science are no better than the fanciful creation stories of primitive mythologies.

2. Where the Philosophy of Science Literature Falls Short

If we are to understand how science succeeds where these other narratives fail, we must understand how this relation of inductive support works. That is a core task for philosophy of science. Its efforts reside in the expansive literature on induction or inductive inference. The project of this book results from an enduring dissatisfaction with this literature.

There is no shortage of approaches in this literature. However, what is distinctive about these approaches is that they are fractured. There are many of them. They rise and fall with the generations and even with the particular philosopher consulted. Each has its successes and each has its failures. None, it seems to me, is by itself fully adequate to the task.

Loosely speaking, there are two traditions.[1] One is qualitative and a few examples illustrate its pervasive problems. Evidence supports those hypotheses that, in various senses, generalize the evidence; or deductively entail the evidence; or explain the evidence; or provide a severe test of the evidence. Each case is troubled. There are so many ways one item of evidence can be generalized that most generalizations cannot be supported. Most applications of the simple scheme must fail. Similarly there are very many hypotheses that entail one item of evidence. The same problem arises. Most applications of this scheme will fail. The problem of proliferation is ameliorated if the hypothesis must not just entail the evidence but explain it. The meagerness of the gain is revealed when we realize that we have no general account of explanation precise enough to support a theory of inductive inference. The account rests ultimately on dubious intuitive judgments of what explains what and how well it does it. Severe testing requires a judgment that the evidence would likely not come about were the favored hypothesis false. To apply the scheme we must know what is likely in the case of this falsity.

[1] This is a hasty dissection of an enormous literature. See Norton (2005) for a more careful dissection and categorization.

Excepting contrived situations like controlled studies, such judgments are at best speculative and at worst self-serving inventions.

The second tradition is quantitative. We assign a numerical measure to the support. The measure used almost universally is probability. The approach is, initially, appealing since we replace a vague "weakly supports" or "strongly supports" by precise numbers that must be combined by quite specific rules. Now we can calculate! My enthusiasm for this approach dampened when I found that its central theoretical tool, Bayes' theorem, has a voracious appetite for prior probabilities and likelihoods. The trouble is that their values must be specified by considerations outside the calculation itself. Prudent or malicious choices for their values, more than the niceties of mathematical theorems, control the final result. Worse, as this Bayesian approach ascended to the momentary dominance it presently enjoys in the literature, its analyses became more and more separated from real applications to inductive inference in the sciences. They have drifted towards self-contained exercise in recreational probability theory. That separation is disguised by tendentious labeling of terms. A calculation best adapted to the accumulated results of many coin tosses is represented as giving some sort of understanding of how the accumulation of intricate and diverse evidence in science can support a univocal result.

The situation has not been improved by a rash decision to conceive of the prior probabilities of Bayes' theorem subjectively, that is, as freely chosen opinions that can vary from person to person. For once one has let arbitrary opinion into the system, the probabilities cease to measure strengths of inductive support, but only some indissoluble amalgam of them with arbitrary opinion. These problems are not resolved but compounded with dubious analogies. We are told a fable of a punter at a racetrack making monetary bets with bookies who are determined to take every advantage possible. This epistemic situation is supposed sufficiently close to that of scientists weighing evidence for big bang cosmology or a neural basis for cognition that all should conform to the same principles of rationality.

3. The Material Approach

The upshot of these accumulated woes is that philosophy of science as a discipline cannot now offer those outside it a univocal account of inductive support. My goal in this book and in the larger program of research it embodies is to solve this problem. The clue to its solution is found in the observation that each of the accounts sketched above do work somewhere. If we are investigating controlled trials, then ideas about severe testing are apt. If we are interested in matching DNA from blood samples with that of accused offenders, then we can use Bayesian methods. When Einstein found that his new general theory of relativity "explained" (as he put it)

the anomalous motion of Mercury, he could claim a wonderful "confirmation" (as he wrote) of his theory.

The clue in all this is that the application of the various approaches works when we add factual conditions that limit the domain in which they are to be applied. The stronger the factual restriction, the more successful the application. The material approach simply asks us to "take the limit." That is, what warrants the successful application of the particular inference is found *entirely* in the background factual conditions that delimit the domain of application.

This last assertion is the key idea of the material theory. It distinguishes it from all other approaches. They use the standard literature in deductive inference as the model for analyzing inductive inference. It provides them a formal model. According to it, we distinguish the good from the bad inferences by checking whether the candidate inference fits in its form with some universal template or schema. For example, take the inference

All men are mortal.

Therefore, some men are mortal.

This is a valid, deductive inference since it is derived from the universally applicable schema that I will call "*all-some*":

All *A*'s are *B*.

Therefore, some *A*'s are *B*.

We are allowed to make any substitution for *A* and *B* and we are assured that what results will be a good inference in its form. The schema is universally applicable. Its use is not restricted, for example, to inferences about human mortality.

Since antiquity, philosophers have sought to recover similar schemas for inductive inference. The successes have always been partial. One of the earliest attempts was "enumerative induction":

Some *A*'s are *B*.

Therefore, all *A*'s are *B*.

The trouble is all too clear. It will almost never work. With obvious substitutions, we might be happy to infer:

Some men are mortal.

Therefore, all men are mortal.

But we would be unhappy with almost every other variant of it, such as:

Some men are Greeks.

Therefore, all men are Greeks.

All of the approaches sketched briefly above lie within this formal tradition. If we just focus on simple examples like these, it becomes quite apparent that they must fail to have universal scope.

The schema *all-some* does have universal scope since it is fully self-contained. Its cogency derives completely from the meanings of the words "all" and "some." If someone doubts the cogency of the inferences it authorizes, we would gently inquire of them whether they understood the meaning of the words.

In contrast, enumerative induction is not self-contained. It can work, but only when we restrict the substitutions for *A* and *B* to terms hospitable to the induction. When *A* is "men," successful substitutions for *B* include biological properties like "is mortal," "is borne of a mother," "has a blood circulation system," and so on. That is, if we restrict the domain in which the schema is applied, it can warrant good inferences. However its success is entirely dependent on the restriction. The facts comprising that restriction are the ultimate source of its warrant. They are biological facts about people. The inference is warranted, in the last analysis, because that is the way people are biologically.

Further, the inference is a good inference only in so far as the warranting facts are true. If science advances so far that we can create people entirely in the test tube from synthetic DNA without the need for a gestating mother, some of these facts would cease to be true and one of the inferences would become an inductive fallacy.

It is easy to see how these conclusions about inductive inference generalize. All inductive inferences lead to conclusions that go beyond what is necessitated logically by their premises. It follows that they are only good in so far as the inferences are carried out in domains that are factually hospitable to the inferences. The facts that make the domain hospitable are the facts that warrant the inference. Here it is helpful to remember that a commonplace of deductive inference is that propositions can both state factual matters and also serve as warrants for deductive inference. The proposition "If *A* then *B*." is both a factual proposition and also a warrant that authorizes a deductive inference from *A* to *B*. The material theory asserts that, ultimately, this dual role for factual propositions is the only way that inductive inferences are warranted.

This applies even to Bayesian analysis, in so far as it has any ambitions of providing an account of inductive inference. It is true that the manipulations of Bayes' theorem itself are deductive inferences lying within the probability calculus. We deduce a value near unity for the probability of Newton's universal law of gravitation, conditioned on the motion of the sun's planets and their moons. An essential background fact is that these deductions are implemented in a domain in which distributions of inductive support are properly represented by probabilities. In the second half of this book, we shall explore domains in which this presumption fails.

These last considerations constitute the core of the material approach to inductive inference. It provides a single, unified approach that incorporates all the different approaches presently in the literature; or at least it incorporates them all in so far as they are sufficiently precisely defined to be viable in some domain.

Its core ideas can be encapsulated in some slogans: "All induction is local." This slogan reminds us that any regularity we may find among inductive inferences is restricted to some domain and dependent for its warrant on the particular facts that obtain there. Another slogan is "There are no universal rules for inductive inference." It reflects the core posit that the warrant of an inductive inference is not traced back, ultimately, to some universal schema, but to facts that obtain only locally.

If one hears only this latter slogan in isolation, one might mistake it for a skeptical thesis akin to Feyerabend's notorious "anything goes." That is very far from its import. It is merely a part of the relocating of the warrant of inductive inferences from rules to facts. The material theory does not seek to undermine inductive inference. It seeks to save it. For the formal approaches that dominate the literature have simply failed in their most important functions. None gives us a successful system, applicable universally, for discerning which are the good inductive inferences. None gives an account of why the inferences it does authorize are appropriate. This last failure stands in stark contrast with standard examples of deductive inference. Inferences warranted by the deductive schema *all-some* are good inferences simply in virtue of the meaning of "all" and "some." These last considerations pose two problems that the material theory solves.

First, inference schemas in the present literature cannot be used universally. While their writings are curiously silent on the question, Bayesians will concede to me in conversation that their system does not apply everywhere. That invites the key questions of where are the limits and how we identify them. The material theory answers: one must locate the facts that can warrant the schema, Bayesian or otherwise. The schemas can be applied only in domains in which those facts obtain.

Second, merely stating an inference schema does not automatically make it a good one. In familiar deductive cases, we discern that they are good because of the meaning of the connectives. We cannot do the same for inductive schemas. Instead, the material theory tells us that certain inference schemas are good since they depend on factual matters in the domain of application. Biological predicates, like "is mortal" and "has a blood circulation system" are facts common to all people and that fact of commonality authorizes the inferences sketched earlier.

Adopting the material approach to inductive inference leads one to approach problems in inductive inference differently. There is no default scheme that can be applied mechanically and automatically. If one wants to employ some mode of inductive inference in some context, one must be able to supply positive reasons for why that mode is applicable in that circumstance. This applies especially to probabilistic inference. One should not assume by default that it always applies. If it is to be used in some domain, we have a positive obligation to provide the foundations for its applicability. Otherwise it cannot be used.

While this book is largely not concerned with beliefs (credences) as opposed to objective relations of inductive support, the moral carries over. There should not be a default presumption that credences are probabilities. If credences are to be represented as probabilities in some circumstance, then positive reasons must be given for why they are appropriate in that circumstance.

4. The Chapters

The chapters of this book are divided into two parts. The earlier Chapters 1-9 are devoted to laying out the basic ideas of the material theory and applying it to what are identified above as the qualitative approaches to inductive inference. The later Chapters 10-16 concern quantitative approaches, most notably the probabilistic approaches of Bayesianism.

Chapter 1 states the basic propositions of the material theory of induction. The vehicle to develop them is Marie Curie's inference from the crystallographic properties of her sample of radium chloride to those of all possible samples. It is an instance of enumerative induction of breathtaking scope. It depends on the evidence of just a few specks of the only sample of radium chloride then known. This chapter also shows how the material theory can warrant successful inferences of this form, even if of breathtaking scope, by displaying the underlying facts that warrant them. In this case the pertinent fact is Haüy's principle. It lies at the core of extensive investigations into the properties of crystals in the nineteenth century and solves the vexing problem of discerning just which of the many properties of crystals are projectable, that is, suitable for enumerative inductions.

Chapter 2 elaborates the argument stated briefly above in Section 3 that justifies the material theory of induction. The essential ideas of the justification are these. No extant formal scheme of inductive inference has proven to be applicable universally. The successes of all these schemes can be explained by the material facts within the restricted domains in which they succeed. Most importantly, inductive inference is by its nature ampliative. That means that its conclusions are logically stronger than its premises. Hence an inductive inference can only succeed in domains in which further background facts are hospitable to it. This chapter also poses the inductive puzzle "1, 3, 5, 7. What's next?" The puzzle is, of course, insoluble nontrivially without some indication of the background facts that can serve to warrant an inductive inference that answer "what's next?" The chapter reports the underappreciated and ingenious way Galileo solved the problem.

Subsequent Chapters 3 to 9 address specific rules and schemes proposed in the literature for inductive inference. The goal of these chapters is to show that, when these rules or schemes work, they do so because of identifiable background facts; and that they can only work in

domains with such hospitable facts. We also find in each case that the apparent unity of application of the candidate rule survives only as long as we do not look too closely at the details of the examples. As we consider those details more thoroughly, we find the specific background facts taking on the primary burden of warranting the inferences. The original rule survives only as a superficial similarity among the examples.

In writing these chapters, I have tried as much as possible to use examples of inductive inference from real science. This literature can suffer when commonplace, non-scientific examples are used to guide our inductive inferences in science. The material theory predicts the problem: since the background facts of ordinary life differ from those of abstruse scientific contexts, there is no basis for expecting the same inferential schemes to work in both contexts.

Chapter 3 looks at the idea of replication of experiment. It is routinely touted in the scientific literature as the "scientific gold standard." We find that merely a useful, but defeasible rule of thumb. It has not been given a precise enough formulation, comparable to those of the schemas of deductive logic, that would enable its mechanical application. Through a series of case studies, we find that the rule is defeasible and has been overruled in every possible combination. Successful replications (intercessionary prayer) and failures of replication (Miller experiment) have both been discarded as evidentially inert. However, on a case by case basis, warrants for the strong inferences associated with individual replications can be found in particular facts in their domains. A general principle of replication is superfluous.

Chapter 4 investigates analogy. It is a traditionally recognized argument form whose history extends back to Aristotle. However, as a review of the recent literature shows, efforts to express the form precisely as a universal rule devolve into an explosion of divisions into special cases and further qualifying clauses. Each expansion produces new problems that require further expansions and, paradoxically, carries us farther from any final formulation. This conception of analogy as an argument form is contrasted with how analogies are treated by scientists. For them, analogies are facts. This fits with a material analysis, for it allows analogies to be both facts and warrants for inductive inferences. Among these warrants, there can be no universal, formal rules. Efforts to adapt a candidate analogical rule to real examples will force a proliferation of conditions, while the rules seek a unity not present in the details of the examples. Instead, the inferences we label analogical are warranted by the facts of analogy identified by the scientists. In the examples explored in the chapter, Galileo infers analogically to mountains on the moon. His inferences are justified by the fact that the dark patches visible on the moon's surface are formed by the same processes that produce shadows on the earth. The same factual basis for inference is found in two further case studies: Reynolds analogy in transport phenomena in fluid engineering and the liquid drop model of the nucleus of an atom.

Chapter 5 takes an unflinching look at the now fashionable talk of "epistemic values" or "epistemic virtues." An early twentieth century quantum physicist who prefers the logically inconsistent old quantum theory does so, we are to suppose, because that physicist values simplicity over the competing virtue of logical consistency. The latter, however, is valued more highly by the classical physicist who then finds a different import for the same evidence. If the terms "virtue" and "value" have their usual meanings, they are ends in themselves and can be freely chosen by us. With this understanding, the physicists' inferences cease to be objective. The bearing of evidence merely reflects the physicists' freely chosen biases and prejudices. This, I maintain, is not how notions of simplicity and logical consistency are used. They are not values, but criteria, whose use is justified by their heuristic ability to lead us to the truth. They are defeasible and can be discarded when they cease to serve this end. Unless we wish to endorse an inductive skepticism by our use of tendentious language, we should stop using the misleading language of virtue and value. The term "criterion" serves better.

Chapter 6 examines the inductive criterion of simplicity in greater detail. There is no precise rule that tells us when to prefer simpler hypotheses. The later misattribution to William of Ockham, "entities must not be multiplied beyond necessity," is vacuous without specification of what counts as an entity and which are the necessities. We are bluffed into allowing its vacuity to pass because of the faux dignity of its expression in Latin. Instead appeals to parsimony in real evidential situation are abbreviated appeals to specific background facts that tell us which are the simplest cases. In curve fitting, for example, straight lines are not necessarily the simplest starting point. If we are fitting trajectories to the observed positions of comets, background facts tell us to start with parabolas, then ellipses and then hyperbolas. For tidal data, we start with an elaborate set of sinusoidal curves whose periods are adapted to the physical parameters of the tidal processes.

Chapter 7 probes the Akaike information criterion, which has been offered as a vindication through statistical theory of a general principle of parsimony. Closer scrutiny reveals that the criterion neither employs a presumption of parsimony in its derivation, nor does it entail any such general principle. Its celebrated formula merely adds a term that corrects for the overfitting of data in curve fitting problems. We, not the statistics, illicitly interpret this narrowly applicable term as a vindication of a broader principle of parsimony. The presence of the term itself depends upon strong background assumptions, most notably that the true curve lies within the model under test. Assumptions like these are the material facts that warrant inferences that use the Akaike information criterion.

Chapter 8 addresses the popular argument form, inference to the best explanation. The hope of its proponents is that there is some feature, peculiar to explanation, that can power inductive inferences. The analysis proves unable to find such a feature. Indeed notions of

explanation are so varied that instances of inferences to the best explanation may bear only superficial similarity to one another. At this superficial level, these arguments share a rudimentary common form. Real examples in science commonly begin as comparative arguments. One hypothesis is favored over another because the first entails the evidence. The competing hypothesis fails the evidence. It is either refuted deductively by the evidence or must take on a substantial evidential debt in the form of further unsupported assumptions, if it is to remain compatible with the evidence. The success of the favored hypothesis does not rest on any peculiar explanatory prowess, but merely on its adequacy to the evidence and, more importantly, the failure of the competitor. The more fraught subsequent step of the inference must show that the favored hypothesis prevails over not just this one explicit competitor, but against all. It is often left tacit in real cases in science.

Chapter 9 seeks to reverse a decline in the literature on inference to the best explanation. This literature began rich in real examples drawn from science. The most notable is Darwin's self-conscious use of the argument form in his *Origin of Species*. Since then, proper study of scientific examples has been replaced gradually by imperfect mentions of them that often oversimplify and misinterpret them; and by prosaic illustrations drawn from everyday life. The entirety of Peter Lipton's canonical monograph, *Inference to the Best Explanation*, contains only one example from real science that is developed at length. It is Semmelweis' identification of the cause of childbed fever. The example is poorly chosen since it one of few that happens to be treated more precisely by the simple thinking of Mill's methods.

This literature is increasingly dominated by superficial examples. The best explanation for footprints in the snow is that someone walked past. This example is unlike those in science, for the human explanation of a person making distinctive marks has no serious competitors. Worse, it encourages explanation by intelligent intervention. That would be an unwelcome encouragement to Darwin. He sought to overthrow intelligent creation as an explanation for biological features. My contribution is provide a somewhat more detailed exposition of eight cases in science, to which the loose pattern of inference to the best explanation can be fitted. I show in each case how some powerful, primitive notion of explanation plays no role. These examples illustrate and support the general claims made in Chapter 8 for the structure of inferences to the best explanation in real science.

With Chapters 10 to 16, the narrative takes a different turn. The Bayesian approach presently dominates thinking about inductive inference in the philosophy of science. According to it, relations of inductive support are recoverable in some manner from probabilistic relations among propositions. I have no quarrel with the use of these probabilistic methods in domains where the background facts specifically authorize them. There are many such domains. Where I differ from the Bayesians is over their ambitions of providing a universally applicable

understanding of inductive relations. It is not, contrary to the title of Jaynes' Bayesian manifesto, "The Logic of Science." It is only the logic of certain special cases. My arguments against those ambitions of universality are laid out in these chapters.

Chapter 10 is entitled "Why Not Bayes." It is a statement, not a question. I illustrate how background conditions can lead us to non-probabilistic representations of evidential relations using the extreme illustration of completely neutral evidence. For this case, application of simple invariances leads to a highly non-additive representation of inductive support. It is quite contrary to the additivity of a probability measure. I argue that even the contrivances of the new literature in "imprecise probability" can sometimes fail to do justice to it.

Bayesian analysis is distinctive in that, laudably, it has taken seriously the burden of proving the uniqueness of its probabilistic representations. This chapter argues that all these efforts must fail since they all have the same structure. Whether they are Dutch book arguments or employ representation theorems, they proceed from some set of assumptions and then *deduce* that the targeted beliefs or relations of inductive support must conform to the probability calculus. This last conclusion is a contingent proposition. It follows that it can only be deduced from assumptions that are at least as strong as it logically. Hence, necessarily, the assumption of probabilities must be hidden within the starting assumptions. The proofs are not demonstrations of the necessity of probabilities, but merely a restatement of a preference encoded in its premises. Once one realizes this, it merely becomes a mechanical exercise to identify and expose the hidden assumptions. I carry out the exercise for Dutch book arguments and representation theorems and note that all similar arguments will fail in the same way.

Chapter 11 contains an extended example of this last exercise. The scoring rule or "accuracy" based vindication of probabilism is based on a dominance theorem. If our credences are not probabilistic, then the theorem tells us we can always improve the accuracy of our credences, no matter what the true situation may be, merely by shifting our credences to a probability. The chapter shows that the theorem is sensitively dependent on the particular scoring rule used to measure the inaccuracy of credences. It develops a family of scoring rules such that any desired deviation from additivity in the credences can be secured merely by choosing the requisite rule from the family. Then a variant theorem shows the dominance of credences with the specified deviation from additivity. The literature in accuracy-based vindications has sought to parry such possibilities by seeking further reasons for why only those rules that deliver probabilities are admissible. These efforts cannot succeed since they still seek to derive probabilities deductively from further assumptions. I continue the exercise of displaying how these further assumptions still have hidden within them the presumption of probabilities.

Chapter 12 addresses a more general problem facing all efforts to devise a mathematical calculus for strengths of inductive support. Applications of Bayes' theorem require specification

of prior probabilities. They make a difference to the resulting posterior probabilities. Since they must be determined by factors external to this application of Bayes' theorem, it follows that this specific computation is not inductively self-contained. One might hope to eliminate this dependence on external considerations by a suitable expansion of the scope of the application of Bayes' theorem. The present prior probabilities would then be recovered as posterior probabilities of antecedent applications of Bayes' theorem. Continued expansion might, we hope, eventually eliminate the intrusion of external considerations. It is well known that these hopes fail. No matter how large the scope of the application, one is never freed from the need to use external considerations to fix prior probabilities.

It turns out that this inductive incompleteness of the Bayesian system is not a failure unique to the Bayesian system. Rather, it is an instance of a broader incompleteness that afflicts all candidate calculi of inductive inference. That is, a theorem demonstrated elsewhere shows that this incompleteness must arise in all such calculi that conform with weak and broadly acceptable conditions. This chapter does not develop the theorem in all its mathematical details, but presents its core ideas and some illustrations of it. The theorem gives a precise instantiation of the more nebulous slogan, "there are no universal rules of inductive inference." It shows that there are no inductively complete calculi of inductive inference.

The remaining Chapters 13 to 16 display further situations in which the background facts warrant formal treatments of inductive support that are not probabilistic. They illustrate the locality of inductive inference. In each case, we must first find the facts prevailing in some domain and then read from those facts the particular logic that would apply to it.

Chapter 13 considers an infinite lottery machine that chooses without favor among a countable infinity of outcomes, labeled 1, 2, 3, 4, …. The condition that the lottery machine chooses without favor is expressed as an invariance, "label independent." According to it, the support accrued to any individual outcome, or set of outcomes, remains the same no matter how we may permute the labels. This independence exercises a profound restriction on the formal behavior of strengths of support. For example, all infinite sets of outcomes whose complements are also infinite must accrue the same support. This sector of the logic is highly non-additive. A corollary is that the relative frequency of even-numbered outcomes does not stabilize towards one half in many, repeated drawings. Rather, all relative frequencies continue to accrue equal support. The factual conditions characteristic of the infinite lottery machine turn out to arise in a particular problem in recent inflationary cosmology. The infinite lottery machine logic is the applicable logic.

Chapter 14 undertakes the same exercise for an uncountably infinite outcome set, such as the continuum-sized set of outcomes formed by the real numbers between zero and one. One might think that choosing without favor among outcomes in this set is easily achieved

probabilistically by a uniform probability distribution. That is a misleading impression since, by foundational design, such a probability distribution neglects to assign probabilities to very many subsets of outcomes of the space. If we require a representation that covers all subsets, we arrive at a logic similar to that of the infinite lottery machine logic, but with more sectors. The chapter then considers successive restrictions that would move the logic towards a probabilistic logic. With each restriction we find a variant of the non-probabilistic inductive logic warranted. One application of these intermediate logics is the continuous creation of matter in the steady state cosmology of Bondi, Gold and Hoyle. The most interesting cases technically arise with paradoxical decompositions of measure spaces. They show the existence of outcome sets not measurable by additive measures such as a probability measure. To make their character more concrete, the chapter develops nonmeasurable sets derived from coin tosses. It turns out that a variant, but weak inductive logic—an "ultrafilter logic"—applies to these sets.

Chapter 15 investigates the inductive logic warranted in two sorts of indeterministic physical systems. The first are those whose temporal behavior is indeterministic. They are quiescent for an arbitrary time and then, without any specific triggering event, spontaneously move. The chapter develops the especially simple example of the infinite domino cascade, which is new in the literature. The second type of indeterministic system is those in which specification of one part of the system fails to fix the remainder. Fixing the mass distribution in Newtonian cosmology fails to fix the gravitational potential. It is then shown that no probability measure can represent the indeterminacy. The infinite dimensionality of the space of Newtonian potentials presents especially intractable problems for additive measures. Instead, it is shown that the background facts of the systems realize the invariance that led to the completely neutral support elaborated in Chapter 10. This is the logic applicable to these indeterministic systems.

The alternative inductive logics explored so far all tend to be simpler in their structures than the additive measures of probability theory. Chapter 16 shows that this need not be so. The system considered is the spin of electrons in quantum theory. While probabilities arise in the process of quantum measurement, they turn out not to be the structure representing inductive support that is warranted by the physical facts of quantum theory. Rather that structure is the density operator that also represents states in quantum theory. The chapter explains what these operators are, how they come about and how they represent inductive support. The development is written at a level that presumes no special knowledge of quantum theory, but assumes a little comfort with abstract mathematics. We learn from the example that background facts in some domains can warrant an inductive logic of some complexity that is quite different in its structure from a probabilistic logic.

5. “A” or “The”?

Finally a note on terminology: is it *a* material theory of induction or *the* material theory of induction? I use both expressions. The first refers to the general idea of finding the warrants for inductive inferences in background facts. There is no presumption in this usage of a particular way of proceeding beyond just the general idea. The second expression—*the* material theory of induction—refers to the particular instantiation of the general idea found in this book and my relevant papers.

References

Norton, John D. (2005) "A Little Survey of Induction," in P. Achinstein, ed., *Scientific Evidence: Philosophical Theories and Applications.* Baltimore: The Johns Hopkins University Press. pp. 9-34.

Chapter 1

The Material Theory of Induction Stated and Illustrated

The primary goal of this first part is to argue for a view of induction that I call the "material theory of induction."[2] This first chapter will give a synopsis and illustrations of the theory. Later chapters will elaborate and support the view.

0. The Terms "Induction" and "Inductive Inference"

This is a book about induction and inductive inference. Since these terms may mean different things to different people, it is worth fixing at the outset what is meant by them here. Traditionally, induction has had a narrow meaning. At its narrowest, it refers to "induction by simple enumeration," the inference from "Some A's are B" to "All A's are B." This is an example of what is known as "ampliative inference," for we have amplified the instances to which our knowledge applies. The premise applies just to the few cases of A's at hand; the conclusion applies to all. I take this idea of ampliation in its most general sense to be what induction is about. I shall use "induction" and "inductive inference" as the general terms for any sort of ampliative inference. That is, they are licit inferences that lead to conclusions stronger deductively than the premises or even just conclusions that differ from those that can be inferred deductively from the premises. Therefore the terms embrace what is sometimes called "abductive inference," which is an inference to something that explains an otherwise puzzling phenomenon.

A still broader form of induction commonly goes under the name of "confirmation theory." It typically has no inferences with premises and conclusions. Rather it looks at degrees of support between propositions. The best-known and dominant form is probabilistic support. The conditional probability, P(H|E), represents the total inductive support an hypothesis accrues from all evidence, including our background knowledge, written as E. One then tracks how the support between hypothesis and evidence changes as the evidence is changed. This form of analysis will be included under the terms "induction" and "inductive inference."

---

[2] For earlier accounts, see Norton (2003, 2005).

My use of the terms "inference" and "infer" will follow what I take to be the traditional usage and the one that is still most common. That is, an inference from proposition A to proposition B is a logical relation between the two propositions as sanctioned by some logic. When we infer from A to B we merely trace through that logical relation. The usage is analogous to that of "add." When we add 7 to 5 to arrive at 12, we are merely tracing through the relation 5+7=12 among the three numbers as authorized by ordinary arithmetic.

This usage is to be contrasted with a psychologized notion of the term "inference" that will *not* be employed here. Under this alternative view, to say that we infer from proposition A to proposition B merely records a fact of our psychology: that we proceed from a belief in A to a belief in B, without a requirement that this transition is authorized by some logic. While I understand the distinction is important to those who work in the psychology of belief, it seems to me a troublesome redefinition of a term when its normal usage is already well established. Could not another word have been found? Perhaps the redefinition is supported by the grammar of the use of the term that presupposes an agent that infers. A similar redefinition might insist that saying "we add 7 to 5 to arrive at 12" merely reports our belief in the summation with no supposition that it conforms with arithmetic. I would find that redefinition equally troublesome.[3]

Throughout this volume, unless some context demands a momentary exception, I will restrict notions of inference and logic to relations of deductive and inductive support between propositions, independently of our beliefs and thought processes.

1. The Formal Approach to Induction

My contention is that the broad literature on induction is built on faulty foundations. It has long sought as its most basic goal to develop inductive inference as a formal system akin to deductive logic and even ordinary arithmetic. What is distinctive about these systems is that they are non-contextual, universal and governed by simple rules. If we have six cartons of a dozen eggs each, arithmetic tells us that we have 72 eggs overall. It also tells us that if we have six troupes of a dozen acrobats, then we have 72 acrobats overall. Arithmetic tells us that when it

[3] Harman (2002, p. 173) gives a clear statement of the psychologized notion of inference that is *not* employed in this text: "Inference and reasoning are psychological processes leading to possible changes in belief (theoretical reasoning) or possible changes in plans and intentions (practical reasoning). Implication is more directly a relation among propositions." This usage is incompatible with the long-standing and pervasive usage of "rules of inference" as designating licit manipulations and argument schemas, such as modus ponens and various syllogisms. See, for example, Boole (1854, Ch. XV) and Copi (1967, p.36 and inside back cover).

comes to counting problems like this we can ignore almost everything except the numbers appearing in the descriptions. We extract those numbers and then see if our arithmetic provides a schema that covers them. In this case, we find in our multiplication tables that

$$6 \times 12 = 72$$

That is really a schema that says (amongst other things)

If you have 6 *groupings* of 12 *individuals*, then you have 72 *individuals* overall.

It is a schema or template since it has empty slots, indicated by the words "grouping" and "individuals" in italics; and we generate truths about specific systems by inserting appropriate, specific terms into the slots. Insert "carton" and "egg" and we generate a numerical fact about eggs. Insert "troupe" and "acrobat" and we have a numerical fact about acrobats.

This example illustrates the key features typically sought in an inductive logic. It is to be non-contextual, universal and formal. The numerical facts of arithmetic are non-contextual—that is, independent of the context. In abstracted form, they hold for eggs, acrobats and every other sort of individual. The rules are universal; they don't come with restrictions to particular domains. It is the same arithmetic for eggs and acrobats. And the rules are formal in the sense that they attend only to the form of the sentence asserting the data: six …. of 12 ….. The matter—eggs or acrobats—is ignored.

Deductive logic has developed similarly as a universal, non-contextual formal theory; and it enjoys extraordinary success. It has been a reasonable and attractive project to try to find a similar account of inductive inference. A universal formal theory of induction would enable us to focus attention just on the specifically inductive-logical parts, ignoring all the material complications of the much larger inductive enterprise. And we would hope eventually to generate great theorems of tremendous power and scope, perhaps rivaling those of arithmetic and deductive metalogic.

2. Problems of the Formal Approach

However it is a failed project. The simple formal rules that worked so well for deductive inference have no counterpart in inductive inference. In antiquity, we were quite confident of the deductive schema

All A's are B.

Therefore, some A's are B.

Yet its inductive counterpart, enumerative induction,

Some A's are B.

Therefore, all A's are B.

was already the subject of doubt and even ridicule in antiquity. Inductive logic never really caught up. While deductive inference has settled into the grey maturity of arcane theorem proving, inductive inference has remained an erratic child. For philosophers, the words "induction" and "problem" are routinely coupled.

There are, as we shall see later, a plethora of modern accounts of induction. But none succeed with the simple clarity of deductive logic. We should infer inductively, we are told, to the best explanation. But we are given no comparably precise account of what makes an explanation better or even what an explanation is. Efforts to make these notions precise open more problems than they solve. Or we are told that all of inductive logic is subsumed by probability theory. A later part of this book is devoted to arguing that the resulting theory has failed to provide a universal account of inductive inference. The probabilistic enterprise has become so many-headed that no single formula captures the difficulty. The account is sometimes too strong and imposes properties on inductive inference it should not have. It is sometimes used too permissively so that any inductive manipulation one might conceive is somehow embraced by it. It is almost always too precise, fitting exact numbers to relations that are not that exact.

So how are we to think about inductive inference? A formal theory of induction distinguishes the good inductive inferences from the bad by means of universal schemas. In its place, I urge a material theory of induction. According to it, what separates the good from the bad inductive inferences are background facts, the *matter* of the inference of the inference, as opposed to its *form.* Or, to put it another way, we locate what authorizes an inductive inference not in some universal, formal schema, but in facts that prevail in the domain of the inference.

3. Inductions on Crystal Forms[4]

An example will make the problems of the formal accounts clearer and the idea of a material theory of induction more concrete. We shall consider an elementary inductive inference in science that is so routine that we may even fail to notice that it is an induction. Let us say that a chemist prepares a new salt of some metal and notes its particular crystalline form. It is routine for the chemist to report the form not merely as the form of this sample, but as the form of this salt generally. For crystals have quite regular properties and crystals of different substances have characteristic differences. Nonetheless, it is an inductive inference from the one sample to all.

[4] My thanks to Pat Corvini for correcting errors in an earlier version of this section and also Section 8 below; and later for providing an extensive list of typographical errors in the Prolog and Chapters 1 and 2.

Even if it is easy to overlook its inductive character, we should expect a good treatment of it from an account of inductive inference.

To develop the example, we need to appreciate that adequate reporting of the crystalline structure of a new salt is somewhat delicate. For the individual crystals of one salt may have many different shapes. In the early history of work on crystals, it proved to be quite hard to find a simple and robust system of classification. That complication will prove to be important for inductive inference.

Crystallographic analysis now categorizes crystal forms according to the axes characteristic of the shape. The simplest of the seven crystallographic systems is the cubic or regular system. The crystals of common table salt, sodium chloride, fall into this system. It is characterized by three perpendicular axes of equal length. A cube conforms to this system; it takes no great geometrical insight to see that a cube has these three perpendicular axes of equal length. The same is true of a regular octahedron, which also conforms to the system. Sodium chloride normally crystallizes in cubes. However in special environments, such as in the presence of urea, it can crystalize as octahedra.

One might imagine that the cube and the octahedron are the only shapes that crystals in the cubic system can adopt. Matters are more complicated. There are many more shapes in this system. The mineral spinel lies within the cubic family and forms octahedral crystals. However spinel can also form many misshapen octahedral crystals, as shown in Figure 1.

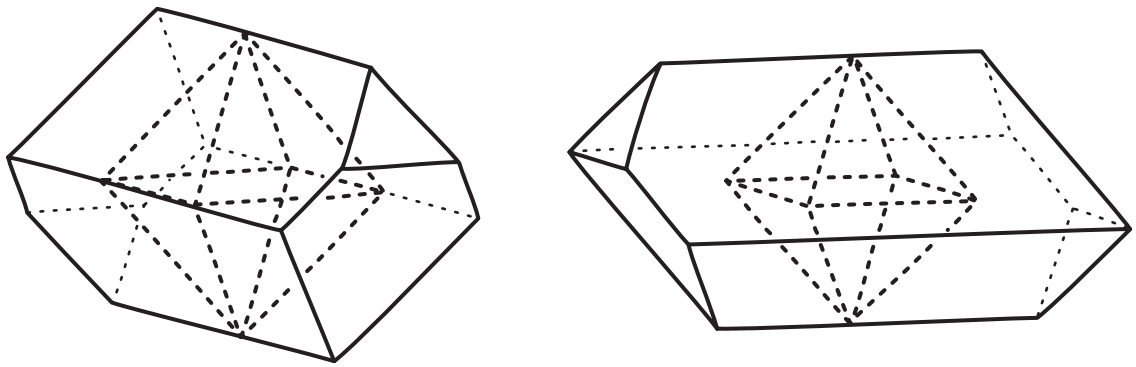


Figure 1. Mis-shapen Octahedra[5]

Their octahedral character arises from their faces being parallel to those of a fictional regular octahedron, which we might imagine secretly buried within the crystal.

Crystals have natural cleavage planes. A crystal cube of sodium chloride will cleave along planes parallel to cube's surfaces. The mineral fluorspar represents an unusual case. It is in

[5] Redrawn after Miers (1902, p. 11, Fig. 9 and 10).

the cubic family and crystalizes in cubes. However it cleaves along planes that eventually expose an octahedral shape. Figure 2 shows successive cleavages.

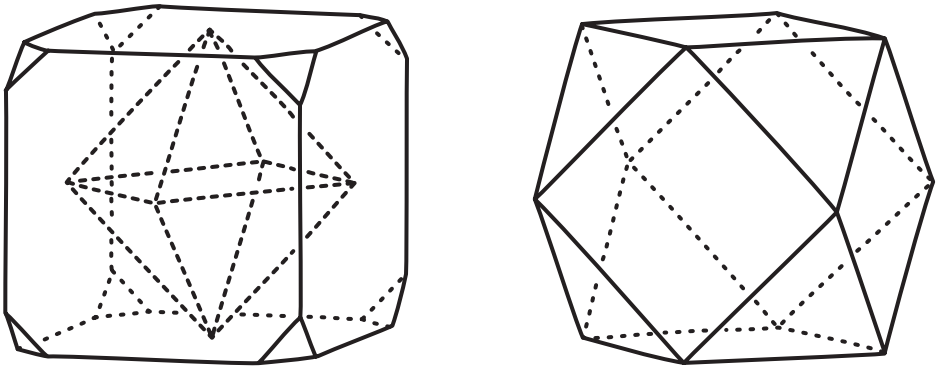


Figure 2. Cleaving Fluorspar[6]

During the process, we pass through many, more complicated shapes of cubes with corners removed to different extents. The shape on the right of Figure 2 is such an intermediate form. These multi-faceted shapes and many more are licit forms for certain crystalline substances within the cubic system.

All these shapes are different from the crystalline shapes permitted to barium chloride, for barium chloride is monoclinic. That means that its crystals are characterized by three unequal axes, two of which intersect at an oblique angle and the third is perpendicular to them. Instead of a cube, its primitive form, the simplest crystal shape, is a right prism with a parallelogram base. This is shown in Figure 3, where the parallelogram is the rearmost face. Alternatively, one may generate the shape by starting with a right prism with a rectangular base and inclining it to one side (hence "mono-cline"). In Figure 3, the inclination is towards the right of the figure.

[6] Redrawn after Miers (1902, p. 14, Fig. 17 and 18).

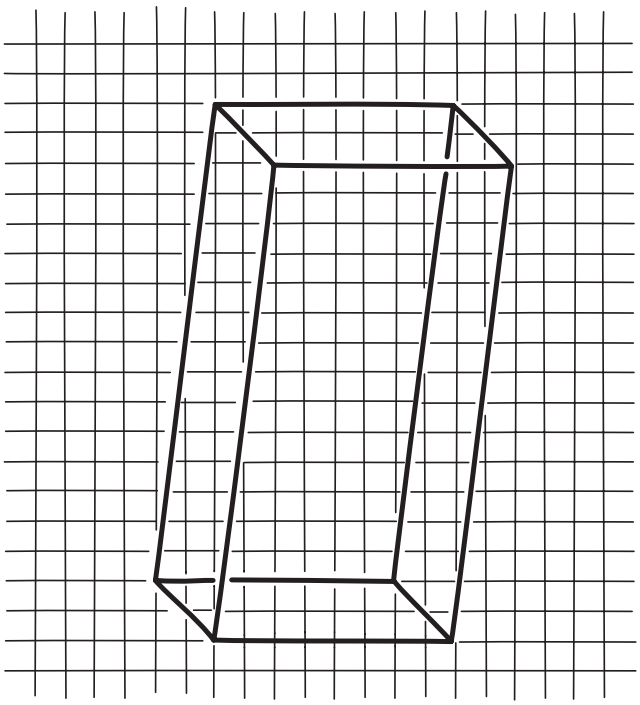


Figure 3. Primitive Form of the Monoclinic System

The range of crystal shapes allowed in the monoclinic system is related to this form, in the same way that those allowed in the cubic system are related to a cube.

When a new metallic salt is prepared, the chemist will simply assert that such-and-such is the form of the salt's crystals. This is an inductive inference and one of breathtaking scope. On the strength of just a few samples, the chemist is quite prepared to infer the crystal system of all samples of the salt:

This sample of salt A belongs to crystallographic system B.

Therefore, all samples of salt A belong to crystallographic system B.

4. Curie and Radium

Perhaps the most famous of all episodes in crystal formation was Marie Curie's separation of radium by fractional crystallization from uranium ore. The massive labor of extracting radium from the pitchblende ore is the stuff of scientific legends, Nobel Prizes and a 1943 MGM movie. The radioactive elements, polonium, radium and actinium, exist in such trace quantities that several tons of uranium ore residue had to be treated to recover just a few decigrams. A decigram, a tenth of a gram, is a mere speck. The process of recovering the radium was arduous. From each ton of ore, after much processing, about eight kilograms of barium

chloride was recovered. Radium chloride is present in it as a trace impurity, revealed by its great radioactivity.

The final separation of the radium chloride from the barium chloride is difficult to achieve since radium and barium behave in similar ways chemically. The separation depends on the fact that radium chloride is less soluble in water than barium chloride. If the barium chloride in solution is concentrated by boiling and cooling until it forms crystals, those crystals will harbor more radium chloride. The solution remaining above the crystals has a fifth the radioactivity of the original, Curie reported. While that seems like a large increase, the quantity of radium present in the crystals is so tiny that it falls far short of what is required for substantial separation. Curie needed to repeat the process over and over; redissolving and recrystallizing to form more fractions; recombining them according to their radioactivity; and doing it again and again. In all she needed to carry out several thousand crystallizations.

All this is described in her doctoral dissertation (Curie, 1904), presented to the Faculté des Sciences de Paris in June 1903. There, she reported on the analytic work carried out in the few years before, with her husband, Pierre Curie. The feature of the radium chloride that attracted most attention was its powerful radioactivity. In spite of the thousands of crystallizations performed, the crystallographic properties of radium chloride barely rated a mention. In the ninety-four pages of the dissertation, there are only a few complete sentences on the crystallographic form (Curie, 1904, p. 26) and they bleed off into less certain reports on the colors of the crystals that, she suspects, may prove of practical use in the separation:

> The crystals, which form in very acid solution, are elongated needles, those of barium chloride having exactly the same appearance as those of radium chloride. Both show double refraction. Crystals of barium chloride containing radium are colourless, but when the proportion of radium becomes greater, they have a yellow colouration after some hours, verging on orange, and sometimes a beautiful pink. This colour disappears in solution. Crystals of pure radium chloride are not coloured, so that the colouration appears to be due to the mixture of radium and barium. The maximum colouration is obtained for a certain degree of radium present, and this fact serves to check the progress of the fractionation.
>
> I have sometimes noticed that formation of a deposit composed of crystals of which one part remained uncoloured, whilst the other was coloured, and it seems possible that the colourless crystals might be sorted out.

Curie and, soon, others separated out only minuscule quantities of radium. Yet, that radium chloride forms crystals just like those of barium chloride entered the literature quite quickly. In his 1913 survey of radioactive substances, Rutherford (1913, p. 470) reported:

> Radium salts crystallise in exactly the same form as the corresponding salts of barium. The crystals of radiferous barium chloride several hours after preparation usually assume a yellow or rose tint. The intensity of this colouration depends on the relative proportions of barium and radium present in the crystal. Nearly pure radium chloride crystals do not show this colouration, indicating that the presence of barium is necessary.

The facts are reported as having quite general scope, even though the instances of observed radium chloride crystals must have been very few, given the enormous labors needed to create them in tiny quantities. Nonetheless, both Curie and Rutherford seem quite certain of the generalization. Rutherford's report looks like little more than a shorter paraphrase of Curie's remark.

5. A Formal Analysis

If we approach inductive inference formally, how are we to accommodate this induction? We need only investigate a few simple formal attempts to see just how poor is the formal analysis. The inference looks like a type of enumerative induction with the schema:

> Some (few) A's are B.
>
> Therefore, all A's are B.

Yet this alone cannot be what authorizes the induction. For almost every substitution for the As and Bs would yield a feeble induction. To get an induction of the strength seen by Curie and Rutherford, we have to be very selective in what is substituted for A and B. The As have to be specific chemical types, such as radium chloride or barium chloride, as opposed to the hundred and one other types of stuff that Curie found in her vats. More importantly, the induction works only for very carefully chosen properties B. There are very many ways of describing crystal forms. Virtually none of them support a strong inductive inference.

To revert to the simpler example, one may find some particular crystal of common salt is a perfect cube. However no chemist would risk the induction to all crystals of common salt having exactly that shape. It was only after serviceable systems of crystallography were introduced that the right property was found. Individual crystals of common salt fall into the cubic or regular system and that property can be inserted into the schema of enumerative induction to form the generalization.

This problem of finding the right descriptions challenged generations of crystallographers. Indeed, for a long time, many held that crystal forms admit no simple systematization so that exactly this sort of induction would be denied. The scientist, historian and philosopher of science William Whewell published in the mineralogical literature. His *History of the Inductive Sciences*

(1837, Vol. III, Book XV, Ch.1-2) gives a lively account of these hesitations and their clarifications by Romé de l'Isle and René Just Haüy after 1780.

These difficulties make it a matter of some delicacy to specify in formal terms just what property of the radium chloride crystals can be generalized. Curie and Rutherford above used parasitic locutions: the crystals of radium chloride are the same as those of barium chloride. Hence Marie Curie, in her 1911 Nobel Prize address, chose a technical locution to describe the crystal form of radium chloride.

> In chemical terms radium differs little from barium; the salts of these two elements are isomorphic, while those of radium are usually less soluble than the barium salts.

Isomorphism is a term of art then and now used to describe the circumstance in which two different substances have very close chemical and crystalline properties. (See, Miers, 1902, p. 213.) It saved Curie the need of describing in more detail the precise structure possessed by the salts of radium. It was familiar knowledge for chemists that barium chloride has such and such a monoclinic crystalline form. The declaration of isomorphism tells us that radium chloride has it too.

If the schema of enumerative induction is to function as a general logic, these restrictions on just what may be substituted for A and B have to be abstracted, regularized and formalized and then included in the schema. The problem is that the restrictions that must be added are so specific that one despairs of finding a general formulation. Presumably a general logic cannot append clauses of the form:

> "…and, if A is a substance that manifests in crystalline form,
> then B must be one of the known crystal forms
> as sanctioned by modern crystallography."

This is a little short of offering a huge list in which we inventory the specific inferences that are allowed. That is not a logic, but merely a catalog whose guiding rationale is hidden.

A more promising approach is to draw on a popular philosophical notion devised for this sort of application: we require that A and B must be natural kind terms. These are terms adapted to the divisions arising in nature ("is crystallographically regular"); as opposed to artificial divisions introduced by humans ("looks like a cubist sculpture"). The hope is that we succeed in delimiting good inductive inferences by restricting the schema explicitly to natural kind terms.

The approach fails at multiple levels. First it fails because the good inductions on crystal forms are still narrower. It is surely a natural kind term for a crystal to be a perfect cube, one of the five Platonic solids. Yet an induction on common salt that uses the property fails to be a good induction by the standards of the crystallographers. Second, the schema is only viable if one can give a general formula that specifies what is a natural kind term. The familiar characterizations

of natural kind terms include that the terms support induction. (Bird and Tobin, 2010, Section 1.1) This means that we are allowed to generalize relations found in a few cases to hold between natural kind terms. If we append this characterization of natural kind terms to the schema of enumerative induction, the schema is rendered circular. For to require that the schema can only be used on terms A and B that support induction is to say in fancy words that the schema only works when it works. Another common characterization of natural kind terms is that they appear in natural laws. If we try to include this characterization in the specification of the schema, we face similar circularities when we try to state just what we mean by "law." Are they true relations that obtain between natural kinds?

6. A Bayesian Attempt[7]

This last section sought to embellish the simple scheme of enumerative induction to convert it into a serviceable scheme with universal application. These efforts were unsuccessful. Might a different approach that employs probabilistic analysis fare better? What if we seek help from Bayesian analysis? We seek a vindication of the inference from some few A's are B to all A's are B that relies essentially on the probabilistic character of relations of support. It should not merely adopt antecedently some version of the idea that the proposition "all A's are B" accrues support from the proposition that "some A's are B"; and then just restate it in probabilistic language. We saw that precisely that idea proved unsustainable in the last section. Merely translating the idea into probabilistic language would only serve to hide the difficulties behind a veil of numbers and formulae. In addition we should like the probabilistic analysis to show us that some *few* A's are B can provide strong support for all A's are B.

There are many ways that one can give Bayesian analyses of this problem. Let me sketch just one. We write H for the hypothesis that a newly prepared salt belongs to some particular crystallographic system. We write E for the evidence that a number of samples are each observed to belong to that class. If there are n samples, we can write $E = E_1 \& E_2 \& \ldots \& E_n$, where E_i asserts the evidence in the i-th case. The probability of interest is P(H|E), the probability of the hypothesis H given the evidence E. It represents the inductive support afforded to H by E, if we think of the probabilities objectively. Or it is the belief we have in H given that we know E, if we interpret the probabilities subjectively. We are interested in seeing how the posterior probability P(H|E) compares with the prior probability, P(H); that is, we seek how the probability of H changes when we incorporate our learning of evidence E. Those changes will tell us the

[7] I thank Nick Huggett for helping me to think through revisions to this and the next section.

evidential import of E. An increase in probability is favorable evidence; a decrease is unfavorable.

We can compute these changes by means of Bayes' celebrated theorem. In a form suitable for this application, it asserts

$$\frac{P(H|E)}{P(\sim H|E)} = \frac{P(E|H)}{P(E|\sim H)} \frac{P(H)}{P(\sim H)}$$

We will not compute P(H|E) directly, but only how incorporating E alters the balance of probability between the hypothesis H and its negation, ~H. That is, we can see how the ratio of prior probabilities, P(H)/P(~H) changes to P(H|E)/P(~H|E) = r. From this last ratio, P(H|E) can be recovered as

$$P(H|E) = \frac{r}{r+1}$$

Bayes' theorem tells us that the controlling quantities are the two likelihoods, P(E|H) and P(E|~H). The first is easy to compute. It expresses the probability that we have the evidence E if the hypothesis H is true. The hypothesis H says that all samples belong to a particular crystallographic system. Hence the n samples at hand must belong to that system. So the probability is unity that we have evidence E: P(E|H) = 1.

The other likelihood P(E|~H) is much harder to determine. How probable, it asks us, is the evidence if the hypothesis is false? Answering that requires some creative imagination for we have no precise prescription for the ways that the hypothesis might fail. The likelihood will vary depending on how we judge the hypothesis might fail. If the only possibility for failure is that the salt belongs to one of the other crystallographic classes, then there is no possibility of the evidence E obtaining. Then P(E|~H)=0. Inserting this into Bayes' theorem leads to P(H|E) = 1; the hypothesis is maximally probable.

However things are not quite so simple. E can be reported if there are observational errors, so that the evidence is misreported. Or it may turn out that the salt is dimorphous or even polymorphous. That means that the salt can crystallize into two or more the systems. So there is some chance, perhaps small, perhaps large, that the evidence E_i obtains, even if H is false.

We will set these concerns aside. Let us set that probability to q so that $P(E_i|\sim H)=q$ and suppose that each of the samples is taken under independent conditions, under the supposition of the falsity of H. Then the obtaining of each E_i is probabilistically independent of the others and the probability of the conjunction is just a simple product of terms:

$$P(E|\sim H) = P(E_1 \& E_2 \& \ldots \& E_n|\sim H) = P(E_1|\sim H).P(E_2|\sim H). \ldots .P(E_n|\sim H) = q^n$$

Bayes' theorem now becomes

$$\frac{P(H|E)}{P(\sim H|E)} = \frac{1}{q^n} \frac{P(H)}{P(\sim H)}$$

Here we have a nice limit result. As n becomes large, q^n can be brought arbitrarily close to 0, as long as $q<1$. Hence the ratio of likelihoods $1/ q^n$ becomes arbitrarily large, so that the ratio $r = P(H|E)/P(\sim H|E)$ also grows arbitrarily large. That corresponds to the posterior, $P(H|E) = r/(r+1)$ coming arbitrarily close to unity. And that means that the support for or belief in H approaches certainty. This limiting result is comforting, for it means that we do not need to worry about the particular values that we might assign to the priors. Whatever influence their values may have had on the final result is "washed out" by the limit process. That is for the better, since the prior probabilities P(H) and P(~H) would have to be plucked from the air.

7. What is Wrong With It

If one inclines to numerical and algebraic thinking, this may seem like a very satisfactory analysis. It has brought mathematical precision to what first seemed like an intractable problem. There is even a little limit theorem in which priors are washed out. All that is an illusion. There are few if any gains in the analysis. However the harm done is great, since we have convinced ourselves that we have solved a great problem, when we have not. Rather any positive result achieved has little to do with the probabilistic properties supposed for relations of inductive support, but everything to do with choices we make externally to the analysis. We shall see that the long term results are determined by our antecedent choice of prior probabilities, which prove to be narrowly constrained to two extreme, dogmatic possibilities. The short term results depend critically on arbitrarily chosen numbers. Finally the necessary condition for any successful result is choosing a description of the hypotheses and evidence delicately tuned to the properties of the system. Without this description, inductive success is impossible. With it, success is assured for virtually any approach.

7.1 External Inductive Content

First, the analysis is heavily dependent on judgments of probability that are supplied externally to the analysis. That is, we must set prior probabilities that presume either a dogmatic skepticism or an unreasonable credulity concerning the universal hypothesis H. There is no other option.

To avoid the danger that these externally specified assumptions prejudge the result, we might require a prior probabilistic independence of the individual items of evidence, $E_1, E_2, \ldots, E_n$. That avoids an antecedent assumption that they are connected by the universal hypothesis H. That is, we would have

$$P(E) = P(E_1 \& E_2 \& \ldots \& E_n) = P(E_1)\, P(E_2) \,\& \ldots \& P(E_n) = s^n$$

where for simplicity I have assumed an equal probability $0<s<1$ for each $P(E_i)$. The result is immediately disastrous. A version of Bayes' theorem now tells us that

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H) = \frac{1}{s^n} P(H)$$

As the number of instances n increases, s^n decreases and can be brought arbitrarily close to zero; which means that $1/s^n$ can be made arbitrarily large. Since P(H|E) can never exceed unity, probabilistic consistency requires that we can no longer choose our prior probability P(H) freely. We must have $P(H) \leq s^n$. Since s^n can be brought arbitrarily close to zero with large enough n, we must somehow choose a prior probability P(H) close enough to zero that anticipates in advance the number of items of evidence that may appear. The only secure value is a zero prior probability, $P(H) = 0$. In this worst case, we preclude learning from evidence, since $P(H) = 0$ forces $P(H|E) = 0$ no matter what evidence E is presented. We must commit to a prior skepticism about the universal hypothesis H.

It is entirely reasonable to respond that this shows that presuming prior probabilistic independence of the individual items of evidence, $E_1, E_2, \ldots, E_n$ is not benign after all. The assumption of independence encodes a dogmatic skepticism concerning the universal hypothesis H. However the alternative is equally troublesome. If we now admit the possibility of a prior probabilistic dependence among the items of evidence, we commit to unreasonable credulity concerning the universal hypothesis H. Here is why.

To avoid prior skepticism about H, we must free ourselves of the need to set P(H) arbitrarily close to zero. We do this by ensuring that $P(E) = P(E_1 \& E_2 \& \ldots \& E_n)$ does not become arbitrarily small as n grows large. Expand P(E) as

$$\begin{aligned} P(E) &= P(E_1 \& E_2 \& \ldots \& E_n) \\ &= P(E_n \mid E_1 \& E_2 \& \ldots \& E_{n-1})\, P(E_{n-1} \mid E_1 \& E_2 \& \ldots \& E_{n-2}) \ldots P(E_2 \mid E_1)\, P(E_1) \end{aligned}$$

We preclude P(E) becoming arbitrarily small by requiring that $P(E_n \mid E_1 \& E_2 \& \ldots \& E_{n-1})$ approaches unity in the limit as n grows large. This requirement says that conditioning on the evidence $E_1, E_2, \ldots, E_{n-1}$ requires the limiting probability of E_n to be arbitrarily close to unity. That is close to assuming H itself. For informally it says that being an instance of H is projectable in this sense: if we have seen n-1 instances of H, with increasing n, we approach probabilistic certainty that the next, nth item will also be an instance with H.

The credulity toward H lies in the permissiveness of this result. It turns out that we approach probabilistic certainty not just for the next instance of H, but for the next N instances of H after it, no matter how large N is. For a simple variant of the last calculation shows that the conditional probability

$$P(E_n \,\&\, E_{n+1} \,\&\, \ldots \,\&\, E_{n+N} \mid E_1 \& E_2 \& \ldots \& E_{n-1})$$

must also approach unity as n and N grow large. Our confidence in projectability is not limited just to the universal hypothesis H, but to any hypothesis of which the items of evidence are an instance, no matter how curious the hypotheses. The hypothesis may be that all samples of radium chloride are prepared by Marie Curie; or all are in Paris; or that all are in the Northern hemisphere.

In sum, we cannot simply present the evidence as bare data and have the Bayesian analysis tell us its import. We have to add prior probabilities and there is no benign way to set them. We must choose antecedently between those that commit us to a dogmatic skepticism or to an unreasonable credulity. This difficulty of Bayesian analysis has long been recognized.[8] Jeffrey (1983, p. 194) was sufficiently disturbed by it that he concluded:

> …willingness to attribute positive [prior] probability to a universal generalization is tantamount to willingness to learn from experience at so great a rate as to tempt one to speak of "jumping to conclusions."

This example illustrates a quite general result reviewed in Chapter 12 below: formal analyses within a calculus of inductive inference cannot be freed from their dependence on externally supplied inductive content.

7.2 Curie Did Not Take a Large n Limit

Second, the analysis has solved the wrong problem. Curie was sure of the result already from just a few samples. She did not need to look at n samples and ponder the result as this n grew arbitrarily large. This "small n" result can be addressed in the Bayesian system, but it requires us to insert numbers. We need concrete values for q and for the priors P(H) and P(~H) in order to see if the analysis supports Curie's analysis. Which are the right values? Can we find them? Or are our selections just hunches driven by dim feelings of what is reasonable?

We now must face the awkward problem of all Bayesian analysis: it introduces specific probability numbers, while no such numbers are in evidence in the inductive practice. Just which value is appropriate for $P(E_i|\sim H)$? Is it 0.1? Or 0.5? What of the prior probabilities? If we think of the probabilities as measuring objective degrees of support, then we have no good basis for assigning the prior probabilities and the whole small n calculation will rest on a fabrication. If we think of probabilities subjectively so that they are merely reflections of our freely chosen opinion, we are no better off. The hope, in this case, is that the accumulation of evidence will wash out the individual prejudices we introduced by arbitrary stipulation of our prior belief. This washing out does not happen precisely because we are limited to the small n analysis.

---

[8] For a brief review, see Norton (2011, pp. 430-31).

More generally, this "solving the wrong problem" is an infraction committed repeatedly in Bayesian analyses. There are a few simple, exemplar computations and the exercise in Bayesian analysis is to modify the problem actually posed in successive steps until it resembles one of them. In this case the original problem is transformed into the problem of distinguishing a double-headed coin (hypothesis H) from a coin that has probability q of showing a head (hypothesis ~H). We are given the evidence E of n independent tosses all of which show heads.

These first two problems are familiar and generally addressed by making the analysis more complicated. If selecting appropriate likelihoods or prior probabilities is troublesome, then reassure a skeptical reader that further Bayesian analysis would surely vindicate exactly the selections needed to get the result promised. My prediction, however, is that this maneuver will not solve the problem. It will merely enlarge the analysis and exile these same problems to remote corners, where more will appear to accompany them. The problems will just be harder to see because the analysis will have become so much enlarged and so much more complicated.

7.3 Finding the Right Description

The third problem is, in my view, the most serious. The Bayesian analysis began by declaring the hypothesis that the salt has crystals belonging to a certain crystallographic system and that the observed instances all conformed to this system. Once that description is given, the most important part of the inductive analysis is over. Once we know that these are the terms in which the problem should be described, then almost any analysis will succeed. Enumerative induction will quickly return something like Curie's result. Or, looking ahead to other accounts of induction, we can declare the evidence a severe test of the hypothesis; or best explained by the hypothesis.

Until we are able to describe things in these terms, no analysis will work, including the Bayesian. The alternative descriptions will either be too coarse or too fine. If they are too coarse, the sorts of hypotheses investigated and affirmed under Bayesian analysis will likely end up as banal. We may affirm that radium chloride forms crystals, for example. If the descriptions are too fine, we will likely find that no hypothesis is well supported by the evidence. If, for example, we give too detailed a description of the crystal form, then the several cases at hand will differ sufficiently so that no single description fits and so that we do not even have a compatible hypothesis to set for H in the analysis.

The damage done by the Bayesian analysis is that it obscures exactly the most important part of the inductive analysis with a smokescreen of numbers and theorems. The essential part of the analysis was the recognition that the hypothesis and the evidence need to be described in terms of a narrow and hard won vocabulary of crystallographic theory. The elaborate computations of the Bayesian analysis mislead us into thinking that inductive problems are

solved by manipulating probabilities and by proving theorems in the probability calculus. It is a seductive aura of precision that is to be resisted if we are to understand inductive inference.

It is widely acknowledged that the real work lies in finding the appropriate system of classification. In introducing crystallography as a "classificatory science," Whewell (1837, pp. 212-13, his emphasis) stresses that finding this appropriate description is the object of the science:

> Our classification of objects must be made consistent and systematic, in order to be scientific; we must discover marks and characters, properties and conditions, which are constant in their occurrence and relations; we must form our classes, we must impose our names, according to such marks. We can thus, and thus alone, arrive at that precise, certain, and systematic knowledge, which we seek; that is, at science. The object, then, of the classificatory sciences is to obtain *fixed characters* of the kinds of things; the criterion of the fitness of names is, that *they make general propositions possible.*

Finding the right system of classification is what makes generalization possible.[9]

8. A Material Analysis

Formal analysis presumes that one isolates the transition from knowledge of a single case to all cases as a problem in inductive logic; and that we establish the cogency of the transition by displaying its conformity with formal principles. Examples are conforming the transition with an abstract schema of enumerative induction or, in the probabilistic case, with Bayes' theorem. Hence the inference from a single sample to all, is immediately beset with the familiar problems that have troubled induction for millennia. They sustain the weary sense among philosophers that induction, trouble and woe all go together.

Chemists at the start of the 20th century, pondering the crystalline structure of matter, would likely not have sensed that their passage from one sample to all was problematic. Indeed they are unlikely to have thought of it in the abstract terms of theories of inductive inference at all. The century before had seen vigorous investigation into the question of just how properly to

---

[9] Looking ahead, a probabilistic analysis could avail itself of the Weakened Haüy's Principle that I argue warrants the inference materially. The analysis would derive directly from the principle that there is a high probability that all samples of radium chloride crystals are monoclinic, conditioned on the fact that Curie's few samples are monoclinic. This is merely a probabilistic restatement of the final result already achieved. Probabilistic analysis has added nothing beyond the illusion of quantitative precision.

characterize the crystalline forms so that the passage from properties of one sample to all may be effected. Curie and Rutherford, if called on to defend this transition, would not have recited passages from logic books. They would have pointed to background knowledge then shared by all competent chemists.

The foundations of the successful approach were laid by René Just Haüy in the late 18th and early 19th century. His approach was based on the idea that each distinct substance that forms crystals is built up from many, primitive geometrical nuclei, all of the same geometric shape. The mineral galena, in this theory, is built from minute cubes. In his treatise published at the time Curie was working on radium, Henry Miers (1902, p. 21) illustrated Haüy's account as in Figure 4:[10]

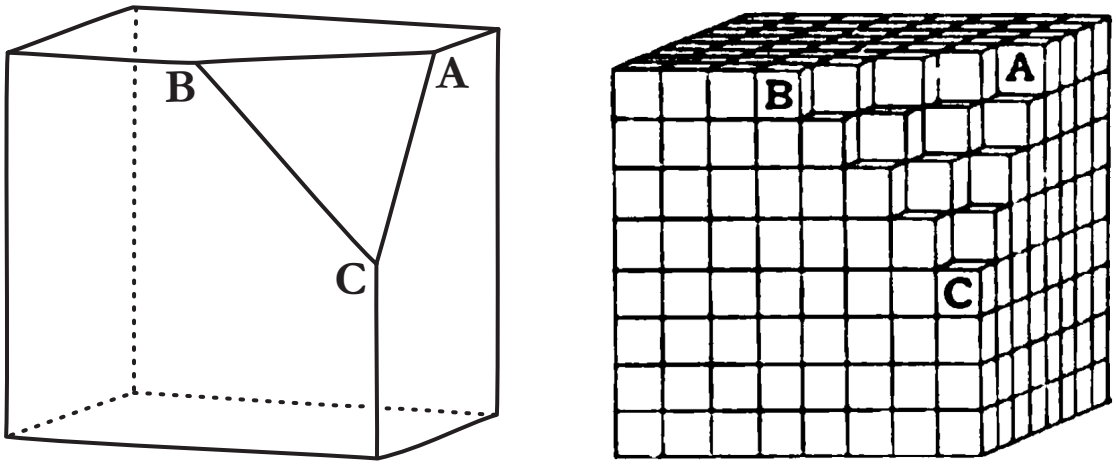


Figure 4. Haüy's Account of Crystalline Shapes

The oblique face ABC of a galena crystal in Figure 4 is, at the smallest scale, really many staircases of these cubes. But that scale is so small that we perceive a perfectly smooth surface.

An account[11] contemporary to Haüy summarized the theory (Accum, 1813, p. 110):

> He [Haüy] has also shewn that *all* crystals, however complicated their form may be, contain within them a primitive geometrical nucleus, which has an invariable form in each chemical species of crystallisable material.

From this theory came the essential result that every substance was characterized by a unique primitive form (Accum, p. 117)

> The diversity of primitive forms ought therefore to be regarded as a certain indication of a difference in nature between two substances and the identity of primitive form indicates identity of composition, unless the nucleus is one of

[10] The figure on the left is redrawn from Miers' Fig. 38 and the figure on the right is a reproduction of Miers' Fig. 37.

[11] This account is more succinct than Haüy's own synopsis of Haüy (1807, pp. 86-101).

> those solids which have a marked character of regularity; such as the cube, the regular octahedron, &c.

The essential qualification is that sometimes two substances may be composed of nuclei of the same form; this was likely to happen for crystals built from regular solids like cubes. This was a quite essential qualification since Accum could list numerous cases of substances with the same crystalline form. Accum (1813, p. liv) listed ten substances based on the cube, for example. Among them are native gold, native silver, native copper, gray cobalt ore, leucite, common salt, galena and iron pyrites.

A century later, Haüy's system had received multiple adjustments and his basic supposition was commonly Bowdlerized as (Anon, p. 365 under "crystallography"):

> The Abbë Reny Just Hauy [sic], whom Dr Tutton designates the "father of modern crystallography," has enunciated the great principle that to every specific substance of definite chemical composition capable of existing in the solid condition there appears a crystallizing form peculiar to and characteristics of that substance.

The view outlined was not so much a principle as a simple consequence of his theory; and Haüy's theory, as outlined by Accum, did not insist that each crystalline substance had its own "peculiar," that is, unique, form.

For our purposes, the essential point is that, if a chemist were to accept Haüy's theory, then one good sample of a crystalline substance is sufficient to identify the crystallographic system to which all crystals of that substance must belong. We have the inference:

> (Haüy's Principle) Each crystalline substance has a single characteristic crystallographic form.
>
> This sample of salt A has crystallographic form B.
>
> (Therefore, deductively) All samples of salt A have crystallographic form B.

This is the crudest version of how chemists pass from a single sample to all. What is notable is that it is no inductive inference at all. The inference is deductive and authorized by early crystallographic theory.

Of course this is an extreme case and a purely deductive passage was possible only during a brief window of a few decades of the early years of Haüy's crystallographic theory. The theory soon encountered anomalies. The shapes Haüy postulated for his nuclei could not always be stacked so as to properly fill space. Whewell (1837, p. 235) reports the collapse of Haüy's physical theory:

> …and when Haüy, pressed by this difficulty, as in the case of fluor-spar, put his integrant molecules together, touching by the edges only, his method became an empty geometrical diagram, with no physical meaning.

A still more serious problem was the recognition mentioned above that one crystalline substance may form crystals belonging to two, three or many crystallographic systems—called "dimorphism," "trimorphism" and "polymorphism" respectively. It was not clear how merely stacking the nuclei of the same shape could yield these different shapes. Mineralogy texts of the early 20th century routinely reported examples. Ford's (1912, p. 80) list is presented more as a reminder of what everyone supposedly knew, than as a surprising novelty:

> Carbon in the forms of graphite and diamond, calcium carbonate as calcite and aragonite, iron sulphide as pyrite and marcasite, are familiar examples of dimorphism. The two minerals in each case differ from each other in such physical properties as crystallization, hardness, specific gravity, color, reactions with acids, etc. Titanium oxide, TiO_2, is trimorphous, since it occurs in the three distinct minerals, rutile, octahedrite and brookite.

This means that Haüy's principle of the earlier deduction was not true, for there were cases of one substance routinely manifesting in several different crystalline forms.

However the idea of a strict regularity in the crystal forms manifested by one substance remained. So we might render a corrected version of the earlier inference as

> (Weakened Haüy's Principle) *Generally*, each crystalline substance has a single characteristic crystallographic form.
>
> This sample of salt A has crystallographic form B.
>
> (Therefore, *inductively*) All samples of salt A have crystallographic form B.

We now have an inductive inference. The warranting principle is what I have called the "weakened Haüy's Principle." What makes it inductive is the insertion of the word "generally." It licenses us to proceed from one sample to all, but not with certainty.

One might imagine that this "generally" is, finally, a manifestation of some universal inductive logic. Perhaps its schema is:

> Generally X.
>
> Therefore X in this case.

While we may find many instances of propositions of the form "Generally…," they are not manifestations of a unique inductive logic. In each case, the word "generally" will have a meaning peculiar to the context. In this case "generally" means "in so far as polymorphism does not interfere." So the nature of the risk one takes in accepting the conclusion will differ with each context.[12]

---

[12] While the inferences may look formally similar, they will be quite different if applied to crystals or to astronomy. Take the proposition: Generally, orbiting objects in our solar system orbit in the same direction as the earth. From it, we may infer with a small risk, that this recently

This is one illustration of how background knowledge drives inductive inferences and how that background knowledge is deeply entangled with inductive practices. Once one knows to look for it, the extent of the entanglement is quite profound. Another notion that was well established at the time Curie worked was the isomorphism, mentioned earlier. This was then defined more precisely as (Ford, 1912, p. 79):

> A series of compounds which have analogous chemical compositions and closely similar crystal forms are said to make an isomorphous group.

An early celebrated instance was a triumph of crystallographic analysis. Whewell (1837, pp. 226-28) reports confusion over the crystalline substance "heavy spar." Haüy found that its cleavage angles varied by three and a half degrees, according to the origin of sample. One was from Sicily and one from Derbyshire. That was a great perplexity and a dire threat to Haüy's theory since the same nuclei could not accommodate even such a small change of angle. It turned out that the two samples were of different substances. The Sicilian was barium sulphate; and the one from Derbyshire strontium sulfate. Barium and strontium are both alkaline earth metals in the same column of the periodic table and have similar chemistry. They also form crystals that are very similar although, as was essential to this story, not perfectly identical. They are a classic case of isomorphism.

When Curie remarked that the radium chloride formed crystals having "exactly the same appearance" as barium chloride, it would have been with full knowledge that the chemistry of radium mimicked closely that of barium. Indeed that mimicry is what made the separation of the two hard. Hence the familiar idea of isomorphism would have indicated that the crystals of the two chlorides should be similar. All that was really left to affirm was how close the similarity would be. It was, Curie found, "exactly the same."

Immediately after Curie's work, the chemical and crystallographic similarity of radium and barium was immediately investigated and affirmed. Runge and Precht (1903) used spectrographic and atomic weight measurements to locate radium with the other alkaline earth metals, magnesium, calcium, strontium and barium. The expected similarity of crystalline forms was found by direct measurement of the bromides of barium and radium. (Soddy, 1907, p.332) reported

> F. Rinne … has published a careful comparison of the crystallographic relation between the bromides of radium and barium and has shown that radium

---

discovered asteroid will orbit in that same direction. The risk we are taking is distinct from that taken in crystallography. We are risking the possibility that this asteroid was not formed by the same processes that formed most other objects in our solar system.

bromide crystallises in the monoclinic system and is isomorphous with and crystallographically closely related to barium bromide.

To report the isomorphism of barium and radium became standard in the literature.

We can now appreciate the great subtlety of Curie's inference. As long as the background theories of crystallography are to be trusted, the possibility of polymorphism was the principal risk taken in generalizing the crystalline form of radium chloride from one sample to many. Hence Curie and Rutherford were quite sanguine to report the radium salts' crystalline form as an isomorphism with barium salts. For, if there was any polymorphism of the radium salt, they could reasonably expect a similar polymorphism to arise with the barium salt. So, with or without polymorphism, their result would stand. With that canny formulation, the result could be asserted with the confidence they showed. The only real danger was a failure of the isomorphism and, given the multiple points of agreement between barium and radium, that was easy to discount.

Let us take stock. Our starting point was a simple inductive inference from a few crystal samples to all samples. It is the sort of simple induction that should be explicated easily by an inductive logic. In particular, we would expect the logical analysis to tell us why this particular inference from a "some" to an "all" is so strong as to be essentially unquestioned. On closer inspection we found that appearance quite deceptive. The strength of the passage from "some" to "all" in this particular case had little to do with issues identifiable by a formal logic. It had all to do with background chemical knowledge. The confidence the chemists had for the inference resulted from the care with which Curie and Rutherford located it within a complicated network of chemical ideas that had been devised over the previous century precisely to admit such generalizations.

9. Main Ideas of a Material Theory of Induction

This last case exemplifies how I believe we should understand inductive inference. Let me collect the main ideas:

Inductive inferences are warranted by facts not by formal schema.

What makes the inductive inference a good and strong one is not conformity with some universal formal schema. It is facts pertaining to the subject matter of the induction; hence the warrant is "material" and not formal. Curie already knew of the closeness of the chemical properties of barium and radium. She knew of the well-established isomorphism that arose in such cases and indicated a closeness of the corresponding crystalline structures. Those facts assured her that the few cases she had observed of similarity of radium and barium chloride crystals could be generalized.

The essential idea here is that facts can serve a dual role, both as statements of fact and warrants of inference. That idea is actually quite familiar. In deductive logic, the conditional "If A then B." serves that dual role. It can serve as a factual premise in an argument; or we can take the same argument and understand its role as warranting a deductive inference from A to B.

In chemistry, the facts that play this dual role look, loosely, like "Generally, X." For example, "Generally, salts that are chemically analogous have similar crystalline structures." This is both a fact in chemistry and an authorization to infer that radium salts and barium salts will have similar crystalline structures because of their chemical similarity. The inference is authorized all the more strongly when Curie found a single sample of radium chloride crystals that, as expected, exactly resembles barium chloride crystals. That diminished the possibility of smaller but superficially detectable differences. The inference is inductive since the chemical facts do not deductively entail Curie's inference. That is the import of the "generally." It accommodates the ways the inference can still fail that are peculiar to this particular chemical example.

All induction is local. It is contextual.

The chemical facts that authorize these inductive inferences are truths of a particular domain of chemistry. They warrant a local mini-logic, peculiar to the context, in which evidence of chemical similarity and of a few samples warrants the generalizations indicated. This local mini-logic resembles the universal schema of enumerative induction. But the resemblance is superficial. There will, no doubt, be other domains in which other facts will warrant inferences that also resemble enumerative induction. The inferences of each domain will be distinct, carrying their own unique restrictions that do not derive from a universal schema, and bearing their own unique form of inductive risk.

Inductive inference is generically variegated and imprecise.

The imprecision here designates a lack of formal properties such as appear in mathematical theories of inductive inference. The inductive inferences on crystalline structure surveyed above are characterized as "strong" or "reliable" or "very certain." These terms have a meaning only within the crystallographic context. Inferences to a unique crystallographic system are prone to failure if the salt displays polymorphism. The inference is "strong" just to the extent that polymorphism can be discounted.

Terms like these are variegated in that they have meanings peculiar to their contexts. The term "strong" will have one meaning in crystallography, another in some branch of physics and another in some subfield of astronomy. What is missing generically is any precise means of comparing the strengths of inferences deemed as "strong" in crystallography and in some disjoint domain, such as physics or astronomy. We also lack precise means for calibrating the difference between, say, "strong" and "very strong" within a single domain. This stands in contrast with

contexts in which probabilities are applicable. The probability of at least one head in ten coin tosses is 1/1024 = 0.99902. In another domain, we may find that the probability that a parent passes on some specific genetic trait is 0.99. The two probabilities are comparable. The first exceeds the second by 1% and this slight difference will manifest eventually in slight frequency differences among many repeated trials.

The qualification "generically" allows that there are important exceptions. Background facts may sometimes authorize a precise, mathematical calculus of inductive inference. The most familiar case arises when we perform inductive inferences specifically on systems governed by probabilistic facts. Such systems include those undergoing radioactive decay, or the forensics of DNA or games of chance in a casino. Later chapters will describe systems in which other, non-probabilistic calculi of inductive inference are warranted. These precise calculi are only applicable when definite background facts warrant them.

The material theory does not authorize the default application of numbers to measure strengths of inductive support. It may be appealing to some to presume such numbers as a default. A probabilistic analysis can supply a definite number—say 0.99—whose closeness to unity gives the sought-for quantitative measure. As satisfying as it may be, without specific background facts to authorize them, applying these numbers is an exercise in spurious precision. It forces variegated notions of strength of support into a single, uniform mold that supposedly enables comparisons across domains. It neglects the domain-specific meaning for the strength of inductive support in each domain. To demand a single number or a single universal term to characterize inductive strengths across all domains invents a uniformity that is not found in the variegated character of inductive inference.

Inductive risk is assessed and controlled by factual investigation.

When one makes an inductive inference, one takes an inductive risk and one seeks both to assess and to minimize the risk taken. In a formal theory of induction, that assessment of the risk becomes an assessment of the reliability of the inference schema used. If we infer to the best explanation, we then need to ask how reliable it is to do that. We are faced immediately with an intractable problem. There is no simple answer to this question; and likely no serviceable, complicated answer either.

In a material theory of induction, things are quite different. The warrant for an induction is a fact and we assess and then control the inductive risk by exploring and developing the fact. Let us imagine that we notice only that a few radium chloride crystals resemble those of barium chloride. The inference to a broader resemblance might then be warranted by a chemical fact that salts manifest only a few crystalline forms. The strength of the inductive inference depends essentially on the correctness of that fact and just how many forms are admitted by the "few." All that can be checked by further investigation and just that is the normal business of research

chemists. They developed theories of how crystals are constituted to enable a better understanding of which crystalline forms will appear in which circumstances. These investigations assure us that two salts will manifest similar crystalline forms if they are chemically similar; and this conclusion is in turn grounded in both other observations and a theoretical argument. Since radium and barium are chemically very similar, the chlorine atoms in a barium chloride crystal will permit the barium atoms to be replaced by radium atoms with minimum alteration to crystal structure.

We assess and control inductive risk by learning more facts. These new facts both provide new premises for inductive inference and also new warranting facts. What was an intractable problem for a formal theory of induction has become a routine problem in exploring the factual realm of chemistry for a material theory.

Inductive inference is material at all levels.

The crystallographic example explored here looks at particular sorts of inductive inferences at a specific level of refinement. One may wonder what happens if we take a more fine-grained view that looks even more narrowly at very specific inferences; or if we take a coarser view that looks at inductive practice at a more general level. Might we find a formal account of inductive inference succeeding there? Might we find that, at levels of great refinement, the glue that inductively binds the corpuscles of analysis is formal? Or that, at a very general level, a universal, formal theory emerges that can unify the diversity of the particular cases?

The claim is that a material theory prevails at all levels. Of course, at all levels there will be inferences that loosely fit with one or other formal theory. We have seen in the crystallographic case that the inferences resemble enumerative induction. We should expect such loose fits, else the formal theories could not have survived at all in the literature. However, they will always be loose fits and, I maintain, closer examination will reveal that material facts are warranting them.

10. Does the Material Theory Say that Inductive Inferences are Really Deductive? No!

No. *No. NO.* It does not say that. This is perhaps the most frequent misreading of the material theory and it can be put to rest here. The material theory maintains the distinction between the two forms of inference. In deductive inference, the truth of the premises assures the truth of the conclusion. In inductive inference, understood materially or otherwise, the premises only lend support to the conclusion. Inductive inference is not deductive inference.

The misreading of the material theory has it affirming that inductive inference is really some form of disguised deductive inference. My sense is that this misreading comes from a similarity between the material theory and another approach to inductive inference. In this other approach, we note that good inductive inferences are also deductive fallacies. For example, we take as a premise:

This sample of salt A has crystallographic form B.

and from it infer

All samples of salt A have crystallographic form B.

This is a deductive fallacy. We could imagine that the argument is really, secretly, a valid deductive argument, but we do not see it because one or more the premises are unstated. That would make the argument an "enthymeme," a valid inference with unstated premises. In this case, a suitable unstated premise would be the strong form of Haüy's Principle:

Each crystalline substance has a single characteristic crystallographic form.

With this added premise, the inference becomes deductively valid. In this other approach, all inductive inference is treated this way. They are treated as failed deductions that are repaired by supplying missing or unstated premises It is not how the material theory treats inductive inference, however.

If we transform the inductive inference to a deductive inference by adding such premises, we have generated what is known as a "deduction from the phenomena." The best-known examples are given in Book III of Newton's *Principia*, where he shows how to infer deductively from the phenomena of celestial motions to the basic ideas of his theory of gravitation. His examples are so important that inferences of this type are often called "Newtonian deduction from the phenomena."

In admitting these cases, the material theory does allow that some inductive inferences may turn out to have been deductive inferences all along, once we make the background facts explicit.[13] However—and here is the key observation—this deductive outcome is an extreme and relatively rare case. Most commonly, it does not arise. When we identify the warranting facts, they supply an inductive warrant only. The strong form of Haüy's Principle is false. The correct weakened form of Haüy's Principle merely asserts that:

Generally, each crystalline substance has a single characteristic crystallographic form.

---

[13] That is not a bad outcome at all. We thought that we must take an inductive risk in accepting the conclusion of the original inference. However we learn that background facts assure us that no inductive risk is taken in accepting the conclusion. The inference has become deductive and, in effect, we have already taken any needed inductive risk when we accepted the background assumptions.

That crucial word "generally" makes all the difference. It reminds us that the original principle fails if there is polymorphism. In accepting the conclusion we take the risk that polymorphism, if present after all, will undo the conclusion. That is, the warrant supplied by the weakened form of the principle is not strong enough to assure us of the conclusion with deductive certainty. The distinction between deductive and inductive inference is maintained.

---oOo---

The chapters to come will elaborate and illustrate these claims further through examination of a sequence of inductive inference forms employed in the literature: the replication of experiment, analogical inferences, inferences grounded in notions of simplicity and inference to the best explanation. They will be followed by an extensive investigation into the limitations of the Bayesian approach. Where the present chapter has developed the material theory of induction by means of an example, the next chapter will develop the general arguments for it.

Note added March 15, 2020.
Commentaries on the draft chapters of this book have been collected for an issue of *Studies in History and Philosophy of Science*. It has become apparent from those commentaries that the draft chapters had not adequately distinguished two question that arise within the material theory of induction. They are:

(inductive-logical)
Question: Which inductive inferences are licit?
Answer: Those that are warranted by a (true) fact.

(epistemic)
Question: How can we know that a specific inductive inference is licit?
Answer: We must be assured of the truth of the appropriate warranting fact.

The first question is answered by matters of fact that obtain independently of any human beliefs, knowledge or awareness. The answer to the second question depends on the answer to the first question. To know that some candidate inference is licit, we need to know the warranting fact. Gaining that knowledge can sometimes be troublesome. We may have to conjecture which is the warranting fact. In that case, we cannot be assured that the associated inference is licit until further investigation assures us that we have conjectured a factual truth.

References

Accum, Frederick (1813) *Elements of Crystallography After the Method of Haüy.* London: Longman, Hurst, Rees, Orme and Brown.

Anon (1911), *The Americana Supplement: A Comprehensive Record of the Latest Knowledge and Progress of the World Compiled by the Editorial Staff of the Americana assisted by expert authorities Complete in Two Volumes.* The Scientific American Compiling Department.

Bird, Alexander and Tobin, Emma, (2010) "Natural Kinds", *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/sum2010/entries/natural-kinds.

Boole, George (1854) *An Investigation of the Laws of Thought.* London: Walton and Maberly.

Copi, Irving M. (1967) *Symbolic Logic.* 3rd ed. New York: MacMillan.

Ford, William E. (1912) *Dana's Manual of Mineralogy.* 13th ed. New York: John Wiley & Sons.

Haüy, René Just (1807), *An Elementary Treatise on Natural Philosophy.* Vol. 1.Trans. Olinthus Gregory. London: George Kearsley.

Jeffrey, Richard C. (1983) *The Logic of Decision.* Chicago: University of Chicago Press.

Harman, Gilbert (2002) "Internal Critique: A Logic is not a Theory of Reasoning and a Theory of Reasoning is not a Logic," pp. 171-86, in, R.H. Johnson, H.J. Ohlbach, Dov M. Gabbay, John Woods, eds., *Handbook of the Logic of Argument and Inference: The Turn Towards the Practical.* Amsterdam: Elsevier Science.

Miers, Henry A. (1902) Mineralogy: An Introduction to the Scientific Study of Minerals. London: MacMillan.

Norton, John D. (2003) "A Material Theory of Induction" *Philosophy of Science*, **70**, pp. 647-70.

Norton, John D. (2005) "A Little Survey of Induction," in P. Achinstein, ed., *Scientific Evidence: Philosophical Theories and Applications.* Johns Hopkins University Press, 2005. pp. 9-34.

Norton, John D. (2011) "Challenges to Bayesian Confirmation Theory," pp. 391 - 439 in *Philosophy of Statistics, Vol. 7: Handbook of the Philosophy of Science.* Prasanta S. Bandyopadhyay and Malcolm R. Forster (eds.) Amsterdam: Elsevier.

Runge, C and Precht, J, (1903), "The Position of Radium in the Periodic System according to its Spectrum," *Philosophical Magazine*, **5**, pp. 476-81.

Rutherford, Ernest (1913) *Radioactive Substances and their Radiations*. Cambridge: Cambridge University Press.

Sklodowska Curie, Marie (1904) *Radioactive Substances*. 2nd ed. London: Chemical News Office.

Sklodowska Curie, Marie (1911) "Radium and the New Concepts in Chemistry," pp. 202-212 in *Nobel Lectures: Chemistry 1901-1921*. Singapore: World Scientific, 1999.

Whewell, William (1837) *History of the Inductive Sciences. Vol. III*. London: John W. Parker

August 7, 8, 13; October 31, November 21, 28, 2014. March 21, 2016. June 13, 25, November 9, 2018.

Chapter 2

What Powers Inductive Inference?[14]

1. Introduction

This chapter summarizes the case for the material theory of induction, drawing on material in other parts of this text. There are three arguments for the material theory of induction. The first two are:

1. Failure of universal schemas. Through many examples in this text, we see that no attempt to produce a universally applicable formal theory of induction has succeeded.

2. Accommodation of standard inferences. These same examples show that the successes of many exemplars of good inductive inferences can be explained by the material theory of induction.

These first two arguments suffice, I believe, to make a solid case for the material theory. They are developed in Sections 2 and 3. They make the case without giving an intuitive grounding for why the material approach is the right one. They establish *that* it is, not *why* it is. For the arguments succeed by showing that the other, formal approach fails and that the material approach does work where its competitor fails. The third argument, however, is grounded in the foundational question developed in Section 4 of why any inductive inference should work at all. That is, it asks "what powers inductive inference?" The question presumes that we cannot take the success of inductive inference for granted. If it works, it does so for an identifiable reason. The material theory answers that inductive inference is powered by facts. For emphasis:

3. Inductive inference is powered by facts. The ampliative character of inductive inference precludes universal schemas.

---

[14] My thanks to Fellows and a visitor to the Center for Philosophy of Science for discussion of a draft of this chapter on November 30, 2011: Yuichi Amatani, Ari Duwell, Uljana Feest, Leah Henderson, Gabor Hofer-Szabo, Soazig LeBihan, Dana Tulodziecki, Adrian Wuethrich; and again on November 23, 2014: Adele Abrahamsen, Joshua Alexander, William Bechtel, Ingo Brigandt (presenter), Sara Green, Nicholaos Jones, Maria Serban and Raphael Scholl.

There are two steps in the argument for this conclusion and they are developed more fully in Section 5.

Briefly, the first step notes that inductive inference is, by its nature, ampliative. That is, unlike deductive inference, the conclusion asserts more than do the premises. It amplifies what the premises say. For each sort of inductive inference, there will be worlds hostile to its success. Generalizing chemical properties of samples, for example, is futile in a world without stable chemical properties. Using an inductive inference presupposes that, as a factual matter, we are not in one of those hostile worlds. If the notion of these facts is construed broadly enough, commitment to them is all there is to accepting the logic. These are the facts warranting the inductive inference.

The second step specifies the character of these facts. They are not universal contingencies such as would warrant a universally applicable inductive logic. That is shown by our failure to identify a universally applicable inductive logic and our failure to exhibit such a universally warranting fact explicitly. Rather the facts hold true only in limited domains, so that there are many of them and the inductive logic each warrants has local applicability only.

The two sections following Section 5 illustrate these two steps. Sections 6 and 7 consider the inductive problem of extending the series 1, 3, 5, 7. It is insoluble without background facts to warrant the inference. Section 8 displays some more examples of warranting facts. Finally, our predisposition for treating inference formally is strong. Section 9 will seek to weaken the presumption that all theories of inference must be formal by indicating limitations in the formal, non-contextual treatment of the most favorable case, deductive inference.

2. Failure of Universal Schemas

Formal approaches to inductive inference depend upon supplying a universal template or schema. For example, in the last chapter, we saw the schema of enumerative induction:

Some (few) A's are B.

Therefore, all A's are B.

These templates are then used to generate the licit inductive inferences by substitution of content for the placeholders A and B. The enduring difficulty for formal theories is that no general account of inductive inference has provided a clearly articulated, exceptionless schema. Therefore, all formal accounts fail and, by elimination of its only known rival, we gain support for the material theory.

That all the schemas fail is hard to show directly since there are many of them. What can be shown, however, is the failure of a representative sample, as is done in various chapters of this

text.[15] The mode of failure displayed by this sample is sufficiently straightforward to make it likely that it will afflict all candidate schemas.

In the preceding chapter, we saw in the example of crystalline forms that the schema of enumerative induction fails. For it to be applied successfully to crystalline forms, we needed to add additional, formal conditions contrived to rule out all but the very small set of properties of crystals that support inductive generalization. The sequence of additional conditions seemed to have no discernible end. Once even a few were added, however, it was already clear that the schema had lost all semblance of generality.

In the next chapter, we will look at the requirement of the reproducibility of experiments, which is often introduced as a gold standard of evidence. On closer examination, however, it proves to be something less. It is a guide whose verdict is sometimes accepted and sometimes discarded. There is no formal rule that tells us when the principle is to be upheld and when not. It is a principle that holds except when it does not. The following chapter looks at reasoning by analogy. It is a form of inductive inference whose use has pervaded science from antiquity to the present. Once again we find that the bare schema is too impoverished to be used exceptionlessly. Efforts over the past century to augment the schema have led to supplements of monumental size while still not delivering a self-contained formal schema.

This pattern of failure continues in subsequent chapters. While considerations of simplicity are often invoked in discerning the bearing of evidence, they do not rest upon a factual principle of parsimony in nature. Notions of simplicity prove sufficiently elusive that there is no clear formulation of such a principle. Similarly the slogan "inference to the best explanation" is so familiar that one might presume that there is some hidden inductive power in explanation. The presumption fails on closer examination. Our notions of explanation are too varied and vague to harbor powers sufficient to support a universal scheme of inductive inference.

Finally, a series of chapters investigates what is, momentarily, the favored account of inductive inference in the literature in philosophy of science, Bayesian inference. Any aspirations of universal applicability fail. Several chapters develop cases in which a probabilistic logic cannot apply since such a logic would contradict symmetries inhering in the cases. There is a rich literature that seeks to establish the necessity of probabilities in representations of belief and inductive support. An examination of these arguments shows them all to be circular. This circularity is developed at length in a chapter devoted to the scoring rule approach. Finally, any

[15] In earlier work (Norton, 2003, 2005), I sought to be more systematic. I showed how virtually all accounts of inductive inference fell into one of three families, each being powered inductively by a single idea. Since the failures in the earlier work are here are spread over the three families, we have some assurance that they are adequately representative of the range of accounts.

Bayesian analysis is inductively incomplete in the sense that it always requires inductively potent prior probabilities to be specified externally. I report work elsewhere that shows that this incompleteness is not specific to the Bayesian system but troubles any calculus meeting certain weak requirements. It follows that no single calculus can cover all the inductive inferences of science. To repeat an earlier conclusion: all induction is local.

These examples embody modes of failure that afflict, I believe, all candidates for universal schemas of inductive inference. The schemas may simply be too vaguely specified at the outset to count as a logic of induction, as is the case with inference to the best explanation. Or, if they are precisely specified, they prove too permissive and authorize far too much, such as enumerative induction. Efforts to restrict them may specialize them so narrowly to one particular domain that they lose their universality. Or these efforts may burden them with more conditions. In adding them, we may need to import new notions—natural kinds, explanation, lawfulness—and these in turn require further conditions for their explication; and so on without termination.

3. Accommodation of Standard Inferences

The last section reviews the failure of familiar, formal schemas for inductive inference. These schemas were devised because, to some degree, each fits some collection of inductive inferences we deem licit. The second argument for the material theory is merely the other side of the coin of this failure. Where the formal approach fails for these repositories of licit examples, the material theory succeeds.[16]

Once again this can be read from the analyses of the surrounding chapters. Curie inferred inductively from the crystalline form of mere specks of radium chloride to all samples of radium chloride. What licensed the inference was a fact hard won from the preceding century's work on crystals. It is what I have called the Weakened Haüy's Principle: *Generally*, each crystalline substance has a single characteristic crystallographic form.

In the next chapter, we visit the requirement of the reproducibility of experiments. The requirement proves not to be a universal inductive principle. Rather it arises in connection with a loosely affiliated but irregular collection of inductive inferences concerning repeated experiments. Their otherwise inexplicable irregularity becomes intelligible when we recognize

[16] Norton (2003) works through the three families of accounts of inductive inference showing briefly how the inferences of each account are materially warranted. The treatment of so many accounts there is necessarily brief. The present text seeks to show the material warrant for standard examples of successful inductive inferences at much greater depth. As a result, fewer examples are treated.

that the inferences are warranted by two classes of facts: those specifying when some process will yield the experimental outcome of interest; and those specifying what may confound the experimental outcome. These facts specify when a replication of an experiment is evidentially significant. More important, they specify when the replication is not evidentially significant. The variation in the facts from case to case explains the irregularity of the whole collection.

Arguments from analogy are so varied in their form that, as we shall see in the chapter after, they defy complete characterization even by quite elaborate formulae. The material theory resolves the problem by conceiving analogy in the same manner as do scientists. For them analogy is not an argument form but a fact that asserts the similarity of two systems. This fact warrants inductive analogical inference. The resulting inferences have as varied a form as the facts of analogy themselves. It is this broad range of variation that defeats efforts to find a universal formal characterization.

This pattern persists with the analysis of inductive inferences grounded in notions of simplicity or explanation. Invocations of simplicity in specific cases prove to be abbreviated invocations of background facts. Since these background facts vary from case to case, their summary in an inductively potent principle of parsimony is precluded. Similarly, in specific inferences to the best explanation, explanatory relations contribute nothing to the evidential import. Real examples of this sort of inference in science succeed through the mere adequacy of the favored hypotheses to the evidence and our success in eliminating its competitors by prosaic, non-explanatory means.

Finally, where the probabilistic representation of strengths of inductive support is appropriate, it is because there are specific background facts that warrant them. The examples are many, varied and familiar. Both quantum mechanical and statistical mechanical systems in physics are governed by probabilistic physical laws. These laws provide the warrants for probabilistic inductive inferences over them. In biology, mechanisms of inheritance in population genetics are governed by probabilistic laws. They too warrant probabilistic inferences. An important background probabilistic fact in many areas of the biological and social sciences is the presumption of sampling randomly from a population. This fact is important, for example, in the forensic identification of suspects through DNA analysis. It warrants the probabilistic inferences reported. A related case arises in controlled trials where we randomize subjects into a test and control group. If the randomization is probabilistic, it introduces background probabilistic facts that can warrant probabilistic inferences about whether the effect measured could arise in case the treatment is ineffective.

These examples instantiate a familiar pattern. Whenever a cogent inductive inference appears in a science, it has proven possible to trace the warrant for the inference to background facts.

4. The Mystery of Inductive Inference

The discussion so far has been devoted to the most visible of the problems associated with inductive inference:

(*which?*) Which are the good inductive inferences?

An answer specifies how we distinguish the good from the bad inferences. The material theory of induction says we do it by identifying warranting facts. We do not seek the warrant in universal schemas. That problem is entangled with another problem that is more fundamental, but is largely overlooked in the present literature. How can inductive inference work at all? That is:

(*powers?*) What powers inductive inference?

Once we accept that inductive inference is powered by background facts, it becomes clear why identifying warranting facts has to be the answer to the "*which?*" question.

This second question "*powers?*" needs some elaboration. For it is easy to take for granted that induction lets us do something remarkable. It lets us amplify our knowledge. We pay a small price for this amplification. Our new knowledge is not as certain as the old knowledge from which we proceeded. Sometimes the uncertainty is large. In important cases, the uncertainty is minuscule. Whether it is small or large, we still seem to be getting more than we should. The problem, the big mystery of induction, is to understand how this amplification can happen.

To sharpen the sense of why we need a solution of this second problem, consider an analogous problem. Imagine that we find an oracle. In the darkness, we see the dim outline of the sibyl, wailing and flailing. Her cries reduce and focus into sharp proclamations that time proves to be important and accurate, mostly. And all this for the price of a goat and few drachma in the bronze bowl. Were this to happen, we would not be satisfied merely to note that this oracle has extraordinary predictive powers. We would and should want to know how they are possible. What is it in the order of things that enables this sibyl to open the portal?

The puzzle is the same with induction. It performs a similar miracle, but without the movie-quality special effects. Experience gives us a small part of space for a small span of time. Yet from knowledge of that fragment, we come to be sure that all things began some 14,000,000,000 years ago in an intense conflagration; that tiny smudges of light in the night sky are great galaxies of stars that duplicate our sun many times; and much more, down to the minuscule structure of microbial life. We must ask, what is it in the order of things that allows induction to open this portal? What powers inductive inference?

The dominant trends in the present literature solve neither of these problems well. To solve them, the two problems need to be treated together. We cannot hope to know which are the good inductions without a clear and explicit idea of what powers induction. The literature so far has tried to solve the problems by working with the model of deductive inference. That has

driven us astray, for millennia. It has led us to seek a non-contextual account of what powers induction (*powers?*) and a formal answer to the problem of which are the good inductive inferences (*which?).* Neither works for induction. The central claim of this chapter is that a successful account of induction is contextual and material.

5. The Foundational Argument

The most compact argument for a material theory of induction proceeds by answering the foundational question of what powers induction. It is powered by facts. As indicated in the introduction, the argument has two steps.

Premise 1. Inductive inference is ampliative.

This means that the conclusion of an inductive inference amplifies. It asserts more than the premises. This distinguishes inductive inference from deductive inference. For deductive inferences merely restate what we have already presumed or learned. There is no mystery in what powers the inference and permits the conclusion. We are just restating what we already have in the premises. The warrant lies fully within the premises. If we know all winters are snowy, it follows deductively that some winters are snowy.[17] This derives from the meaning of "all." If something is true of all, it is thereby true of some. The context in which we infer plays no role in powering the deductive inference. The inference succeeds no matter what a winter or snowy might be. The meaning of "all" is enough to empower the conclusion non-contextually. The inference is valid independently of whatever other facts may obtain about weather and climate.

It is quite different with inductive inference. From the premise that all past winters have been snowy in some location, we infer inductively that the next winter will be snowy there. It is entirely possible that this prediction fails. When we conclude in its favor, we assert more that the premises. It is prudent to do so only in certain sorts of worlds. Hospitable ones include those in which climate in the location is stable. An inhospitable world would be one undergoing a warming climate change. We can generalize the crystallographic family of a crystalline substance from one sample to all because our world is hospitable through the background fact of Haüy's principle. But we cannot generalize the size of the one sample to all, for there are no background facts providing for restrictions on possible sample sizes. Correspondingly, we can generalize sizes of living organisms, for different types of organisms are restricted by their physical constitutions to specific scales. Insects cannot grow to human scales because their

[17] For pedants: I follow the informal conversational presumption and tacitly assume that "All winters are snowy." is not true vacuously; that is, its truth requires that there are some winters.

structures would be too weak to support their weight and they could no longer breathe by diffusion. Correspondingly, humans cannot be shrunk to insect scales. Their shrunken brains would have too few neurons for human cognition. At least this is true in our world, which is hospitable to the generalization, but not in a science fiction world in which normal science is suspended.

These examples illustrate the general point: the factual assumption that ours is a hospitable world is the fact that, if true, warrants the inductive inference. It may not always be apparent that this fact warrants the inference. It may appear that the warrant is still provided by some sort of schema. The inference to a future snowy winter, we may think, is still warranted by the schema:

> All past A's have been B
>
> Therefore, the next A will be B.

The supposition incomplete. This schema, if used at all, has a purely intermediate role. It does not have universal applicability. We can use it in the snowy winter case only because the requisite background facts authorize it, when we make the specific substitutions: "winter" for A and "snowy" for B. That is, there is a cascade of warrants that may pass through a schema. The cascade terminates in facts that are the final warrant of the inference.

It is essential here to distinguish two ways that an inductive inference can fail: losing an inductive bet in a hospitable world versus failure of an inductive inference in an inhospitable world. The first case arises because accepting a warranted inductive inference still involves a risk. In a hospitable world of stable climate, it is a warranted inductive inference to infer from a past history of snowy winters that the next winter will be snowy. The next winter, however, may turn out not to be snowy. Such fluctuations are rarer, but quite possible when the climate is stable. Losing an inductive bet like this must be distinguished from the second case in which it was imprudent to take the bet in the first place. If the background facts are of a warming climate in some location, then the background facts do not warrant the inference. If one persists and makes the inference, the conclusion may prove false. The failure reflects the lack of warrant of the inference, not a failure arising from traditional inductive risk.

The material theory of induction arises when we assume that the truth of these background factual presumptions is all that is needed for the inductive inference to be warranted. One might imagine that this might not be so. The facts, we might suppose, play only a partial role in warranting the inductive inference. Might there still be a residual universal formal schema or inductive rule that contributes to the warrant? Such a schema or rule, however, would in turn be subject to the same analysis just given. If it functions to authorize an inductive inference, then it is amplifying what we have already asserted in the premises and all other background facts. It cannot be universal in application for there would be worlds inhospitable to it. We should only

use the rule or schema in worlds hospitable to it. That is, the warrant for its use is the factual supposition that the world is hospitable to it. Once again the inductive warrant has terminated in facts that should be included with the true background facts needed to warrant the inductive inference at issue. That is, the truth of the background factual assumptions, when construed broadly enough, is all that is needed to authorize the inductive inference. With that, we arrive at the major tenet of a material theory of induction.

Hence, inductive inferences are warranted by facts.

What remains open is the precise character of the warranting facts. Aside from the next step, there is little we can say at the general level about the nature of these facts. In particular cases, their character will be straightforward. Our inference to a future of snowy winters is warranted by the assumption that our local climate will persist pretty much as it has, so that winters without snow are possible but unlikely. If the climate warms sufficiently, however, these facts may fail and with it the inductive inference.

In some cases, the background facts may be such that the inductive inference would be deductive if we added explicitly the warranting fact as a premise. Then the inference is revealed to be an enthymeme, a deductive inference with a hidden premise. An example is this version of Curie's inference from the preceding chapter:

> This sample of radium chloride is monoclinic.
> (Weakened Haüy's Principle) *Generally*, each crystalline substance has a single characteristic crystallographic form.
> ---
> Unless exceptions encoded by the "generally" of the principle intervene, all samples of radium chloride are monoclinic.

However, it would also be entirely natural to detach the "Unless…" clause and have the inference:

> This sample of radium chloride is monoclinic.
> (Weakened Haüy's Principle) *Generally*, each crystalline substance has a single characteristic crystallographic form.
> ---
> All samples of radium chloride are monoclinic.

This inference is inductive for we are taking the risk that the exceptions suggested by the *generally* do not arise.

Corresponding complications arise if we infer inductively in the Bayesian framework. If we infer from prior probabilities to posterior probabilities by means of likelihoods using Bayes' theorem, then the inference is deductive. If we broaden the context, this ceases to be so. Propositions asserting evidence and background facts are not provided to us with probability measures. We add them. In doing so, we accept that we can represent their mutual relations of inductive support probabilistically and that their inductive consequences follow from the probability calculus. In doing so, we take an inductive risk that probabilistic analysis correctly represents these relations. If we also proceed as normal people do and accept a proposition with very high posterior probability as established, then we take a second inductive risk in detaching the qualification of high probability.

The second step places a restriction on the character of the warranting facts:

Premise 2. There is no universally applicable warranting fact for inductive inferences.

This premise itself requires support. Part of it is supplied by other arguments in this book that seek to establish that there is no universally applicable logic of induction. For, if there were a universally applicable logic of induction, then by the first step above there would a universally applicable warranting fact.

A more direct grounding for the premise lies in our failure to exhibit such a universally applicable warranting fact. It has been long sought, like the philosopher's stone, and with equal success. The best known attempt at characterizing it is Mill's principle of the uniformity of nature: "The universe, so far as known to us, is so constituted that whatever is true in any one case is true in all cases of a certain description; the only difficulty is, to find what description." (Mill, 1904, Bk III, Ch.III, p. 223) and "Whatever may be the proper mode of expressing it," he wrote, "the proposition that the course of nature is uniform is the fundamental principle, or general axiom of Induction." (p. 224) It is a general fact about the world holding in all domains in which we may seek to infer inductively. It is the one, universal fact that would power all inductive inference.

The trouble with Mill's principle is that, read literally, it is false; but read charitably it is so vague as to be unusable. Take the literal reading. Our world is *not* uniform in all its aspects. Indeed the world fails to be uniform in virtually all its aspects. Otherwise we would live in a largely homogenous environment. At best, the world is uniform in a very few, quite special properties that end up figuring in what we take to be laws of nature. This last statement is the charitable reading. The real challenge now for the principle is to specify just which are those special properties. Yet through the vague generality of its formulations, it provides no such specification. At best, the principle deflates to a weak existential claim: there are uniformly implemented properties in nature, but we do not know precisely which they are. Or, more

generally, Nature is regular and orderly, but in a way that we cannot state or grasp compactly enough to implement as a principle that can be employed practically in a logic of induction.

That the principle needs this shield of ignorance to protect it from scrutiny suggests that there is no real content hidden behind the shield. Certainly it has ceased to have any practical value in our inductive investigations. Salmon (1953, p. 44) long ago wrote the principle's obituary

> …the general result seems to be that every formulation of the principle of the uniformity of nature is either too strong to be true or else too weak to be useful.

This completes the argument for the premise.

If the facts warranting inductive inference are not universal truths, then they must be truths of restricted domains and the inductive inferences they warrant will be restricted to those domains. It may well happen that the inferences warranted in some restricted domain have a regular structure. Then we have an inductive logic applicable to just that domain. For example, Haüy's principle warrants an inductive logic that looks formally like enumerative induction, but is restricted just to generalizations concerning the crystallographic family of samples of crystalline substances. That is, we have the second major characteristic of a material theory of induction.

Hence, all induction is local.

Philosophers are good at finding clever but ineffective loopholes. The following, I have found, is one that many cannot resist. If each domain has its own material facts that warrant inductive inferences in it, why not just form the conjunction of all of them? The resulting conjunction would be a single, *huge* fact that warrants inductive inferences in all domains.

It is correct that this huge conjunction warrants inductive inferences in all domains. However its formation provides no escape from the locality of inductive inferences claimed by the material theory. That locality now reappears in the irreducibility of the huge conjunction to anything more compact. It remains just a single, huge conjunction of this fact and that fact and that other fact and so on, with many, many more conjuncts. To use the huge conjunction in any domain, we have to locate within this immensity the conjunct that applies specifically to that domain, extract it while ignoring all the other conjuncts and apply it. The warranting of inferences in that domain will still be done by facts prevailing just in that domain. The existence of the big conjunction provides no universally applicable scheme beyond the one already central to the material theory of induction: to identify the warrant of an inductive inference, seek facts that prevail in that domain.

The next two sections will supply illustrations of the first and second steps respectively of the argument of this section.

6. The Inductive Inference on 1, 3, 5, 7.[18]

A rapid way to see the importance of background warranting facts is through an inductive inference problem that, by contrivance, is bereft of background facts. The problem is this:

Given the initial sequence of numbers 1, 3, 5, 7,
how does the sequence continue?

It is a trivial mathematical fact that the sequence could continue in any of very many ways. If the only restriction is that these are the first four terms of an infinite series, then there is an uncountable infinity of distinct continuations. The emptiness of the problem specification makes it impossible to favor any one of them, that is, to pick among the deductively authorized possibilities. Without some specification of background facts, to infer inductively about the continuation is impossible.

The possibilities are greatly restricted if we make the natural assumption that the sequence is governed by some simple rule. There are still many possible continuations. The sequence may just be the odd numbers:

1, 3, 5, 7, 9, 11, 13, 15, …

Or it may be the odd primes, including one:

1, 3, 5, 7, 11, 13, 17, …

Or it may be the digits of the decimal expansion of 359/2,645:

1, 3, 5, 7, 2, 7, 7, 8, 8, 2, 8, …

While the possibilities are restricted, the inductive problem is still intractable since the notion of "simple rule" remains under-specified. That makes finding other continuations merely a challenge to our ingenuity in writing laws that look simple in some sense we happen to find congenial.

Another approach embeds the sequence in a context for which we have more information. The numbers may be drawn from a randomizing lottery machine. Then the fact of randomization authorizes a probabilistic analysis. Probabilistic inductive support is distributed uniformly over the remaining, undrawn numbers. Or perhaps the numbers appear in a question on an IQ test or in the interrogation of a psychologist we believe is intent on tricking us. These differing background facts would then authorize different inferences over the continuations, although the complexity of the background will make discerning their precise character troublesome.

[18] This example and a briefer version of the argument of the previous section are given in Norton (2014).

7. The Law of Fall

It is easy to suppose that this inductive problem is merely a philosopher's contrivance, unrelated to real problems of inductive inference in science and thus one that we need not strive to accommodate in our account. That supposition is wrong. The problem turns out be one of the classic problems of inductive inference in science. This particular number sequence happens to figure in one of the great discoveries in the history of science. In his *Two New Sciences* (1638), Galileo Galilei presented his law of fall. In one form, the law asserts that the distances fallen in successive units of time stand in the ratios 1 to 3 to 5 to 7 to…; that is, in the ratio of the odd numbers. Galileo's pathway to this law was long and convoluted. However at least one part of it quite likely involved experimentally measuring the distances bodies fall and the times taken. The *Two New Sciences* (1638, pp. 178-79) describes such an experiment in which a ball is timed rolling down a grooved ramp. The ramp is a surrogate for free fall that slows the motion sufficiently to enable time measurements using Galileo's crude methods. Stillman Drake (1978, p. 89) has identified an early Galileo manuscript that, Drake argues, records the results of just such an experiment.

So let us pose a simplified Galileo-like inductive problem. Given that the incremental distances fallen in successive units of time are in the ratios 1 to 3 to 5 to 7, what will be the distances in subsequent times? Using resources available to Galileo, how might this be solved?

We have a good idea of Galileo's methods. One element was that he presumed that fall is governed by a rule that is expressible simply in the mathematical techniques available to him. The idea is indicated in *Two New Sciences.* Galileo reflects on the gains in speed of falling bodies and asks of them (p. 161)

> …why should I not believe that such increases take place in a manner which is exceedingly simple and rather obvious to everyone?

Galileo's inference is warranted by a fact: the simple nature of this part of the world. This one statement leaves the notion of simplicity at issue underspecified and thus leaves underspecified just which inference is authorized. If we read more broadly in Galileo's writings, we find a stronger statement that identifies the notion of simplicity at issue. He wrote in a famous passage in *The Assayer* (1623, pp. 237-38):

> Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth.

This is mathematical Platonism. It asserts that the world is structured as a copy of the perfect mathematical forms. This factual statement about the world then warrants an inference to a simple mathematical rule as the continuation of the sequence 1, 3, 5, 7, …

This approach may at first be appealing. The world does admit simple mathematical description. Why can we not use this fact to underwrite inductive inferences? The appeal fades rapidly under closer scrutiny. There are three problems.

First, if one is not a Platonist, one judges the warranting fact to be a falsehood and thus the inference an inductive fallacy. The success of mathematical methods in science since Galileo to the present does not, in my view, justify the Platonic view. Rather, as I have argued elsewhere (Norton, 2000, Appendix D), the success merely reflects the post hoc adaptability of mathematics to new scientific discoveries.

Second, attempts to employ the Platonic idea fall prey to the problem that mathematical imagination can conjure up vastly more structures than are implemented in reality. Seek simple laws written in the wrong mathematical language and our investigations will stall and fail. Einstein became a mathematical Platonist as part of his later-life search for a unified field theory. His efforts were stymied by just this problem, since he sought laws that are simple when expressed in the mathematics of tensors and the like on four-dimensional spacetime manifolds. Subsequent theorizing in quantum gravity has branched out in the mathematical structures it uses and typically does not posit a four-dimensional spacetime manifold as a primitive.

Third, when Galileo investigated falling bodies, the mathematics accessible to him was limited to methods drawn from Euclid. They comprise the barest sliver of the mathematics we now employ. It is more than optimistic to expect that the Platonic blueprint of nature is drawn with the mathematics of this tiny sliver.

8. Invariance Under the Change of the Unit of Time

In the face of these mounting difficulties, we may well wonder whether Galileo had sufficient background facts to warrant what still seems like it should be a good inference. Fortunately he did assume another background fact, perfectly tuned to warrant the inference and eliminate all but one of the open possibilities, although this aspect of his work typically gets scant attention in the written history.

Galileo's experimental methods were unable to fix a precise unit of time. At best he could determine that, in one experiment, successive intervals of time were equal. He realized that his experimental result was stable in spite of this variability of the unit of time. In measuring fall, he recovered the same ratios 1 to 3 to 5 to 7 to … no matter what unit of time is used. This

important fact is stated by Galileo when he presents this odd number formulation of his law of fall. He wrote (Third Day, Naturally Accelerated Motion, Thm. II, Prop. II, Cor. I)

> Hence it is clear that *if we take any equal intervals of time whatever*, counting from the beginning of the motion, such as AD, DE, EF, FG, in which the spaces HL, LM, MN, NI are traversed, these spaces will bear to one another the same ratio as the series of odd numbers, 1, 3, 5, 7;…

The invariance of the result is asserted by the italicized text (my emphasis).[19]

With a little arithmetic, we can see how this invariance under the change of unit of time works. In successive units of time, the body falls distances.

$$1, 3, 5, 7, 9, 11, 13, 15, 17, 19, \ldots$$

Now replace the original unit of time with a new one equal to two of the old units. The distances fallen in successive units of time with the new unit are:

$$1+3, 5+7, 9+11, 13+15, 17+19, \ldots$$
$$= 4, 12, 20, 28, 36, \ldots$$
$$= 4\text{x}1, 4\text{x}3, 4\text{x}5, 4\text{x}7, 4\text{x}9, \ldots$$

Galileo's law requires only that these distances be in the ratios 1 to 3 to 5 to 7 to … . Hence we can neglect the factor of 4 and observe that they conform to the law. This invariance, Galileo asserts, obtains no matter which unit of time we select.

The remarkable fact is that there are very few laws of fall that respect this invariance. Using techniques in the calculus and functional analysis not available to Galileo, it is possible to prove that the *only* laws are these. If d(t) is the distance fall in the unit of time (t-1) to t, then[20]

$$d(t) \text{ is proportional to } t^p - (t-1)^p$$

where p is any real number greater than 0. (See Norton, 2014a.) This means that prior to any measurements, the scope of the laws admissible is already reduced to these very few possibilities.

---

[19] Galileo's Latin *quotcunque tempora aequalia* is literally "however so many equal times." Crew and de Salvio render it as "any equal intervals of time whatever." Their looser rendering fits with the overall context in allowing both the number and duration of the intervals to vary. An important part of the context is the earlier statement of the law of fall from which this corollary is derived. The law is first introduced as (p. 161) "…during any equal intervals of time whatever, equal increments of speed are given to it." Galileo's Latin *dum temporibus quibuscunque aequalibus* is correctly rendered by Crew and de Salvio as "during any equal intervals of time whatever," where *quibuscunque* has no restriction to number or duration. These unrestricted, equal time intervals are the ones that reappear in Corollary I.

[20] There is a suppressed proportionality constant in the statement. It is suppressed since Galileo's law concerns ratios of the quantities d(t) and the constant will not affect those ratios.

What now gives the inference great strength is the fact that there is just one free parameter in the law, p. It follows that securing just one ratio of distances experimentally fixes the law uniquely. For example, take the first ratio that Galileo would have measured, d(2)/d(1) = 3. It follows that p must satisfy

$$3 = \frac{2^p - (2-1)^p}{1^p - (1-1)^p} = \frac{2^p - 1^p}{1^p} = 2^p - 1$$

The unique solution is p=2, so that

$$d(t) \text{ is proportional to } t^2 - (t-1)^2 = t^2 - (t^2-2t+1) = 2t-1$$

Hence for successive times t = 1, 2, 3, 4, …, we have d(t) = 1, 3, 5, 7, …, that is, the odd numbers.

This is a remarkable result worth restating: if invariance under changes of the unit of time is to be respected, the only continuation of the two-membered sequence of incremental distances fallen

$$1, 3$$

is the sequence of odd numbers

$$1, 3, 5, 7, 9, 11, 13, \ldots$$

Of course Galileo could not know this result in all generality. However it is quite likely that he was aware of how restrictive the invariance is. One needs only to try out a few alternatives to the odd number sequence arithmetically to realize that all simple alternatives fail. Drake (1969, pp. 349-50) notes a correspondent of Galileo's, Baliani, reported that Galileo had used the invariance as a "probable reason" for the odd number rule.

While Galileo did not elaborate in *Two New Sciences* on this result, Christiaan Huygens soon did. That is, a seventeen-year-old Huygens, prior to his reading of Galileo's *Two New Sciences*, independently found the result.[21] One statement of what he found is given in a letter of October 28, 1646, to Marin Mersenne (Huygens, 1888, pp. 24-28). There Huygens arrived at his result by considering two possibilities: that the incremental distances fallen in subsequent, equal intervals of time grow in an arithmetic progression or in a geometric progression. Only one case gave non-trivial results: an arithmetic progression in the ratios of the odd numbers, 1, 3, 5, 7, … The demonstration is creditable, but less than general since it overlooks the possibility of expressions for the incremental distances d(t) with values of p other than 2 in the formula $t^p - (t-1)^p$. Thus it precludes by supposition many other progressions that would give laws of fall whose ratios remained unchanged under changes of the unit of time. While one might imagine ways that the demonstration could be rendered more general, there seems to be no obvious way

[21] I thank Monica Solomon for drawing my attention to Huygens' work and for sending me a copy of his letter and other supporting materials.

to arrive at the general proof without mathematical techniques stronger than those available then to Galileo and Huygens, such as used in Norton (2014a).[22] This may explain why Galileo did not elaborate on the result in *Two New Sciences.*

Our Galileo-like inductive inference problem admits a ready solution. We take as a premise that the ratios of the incremental distances fallen in successive units of time are 1 to 3 to 5 to 7. There are two warranting facts accessible to Galileo: the rule governing the sequence is expressible simply; and the rule is invariant under a change of the unit of time. Only a small amount of arithmetic exploration will show that this invariance likely rules out all extensions other than the odd numbers. A fuller analysis shows that the second invariance by itself is sufficient to warrant the inference.

9. Can Bayes Help?

One might imagine that the general inductive problem of extending the sequence, 1, 3, 5, 7, is one at which Bayesian methods would excel. Might a Bayesian analysis somehow succeed without the need for specific background facts, contrary to everything that has been said so far? In short, the answer is that it does not provide a successful, universal treatment of the problem. There are two striking failures in the analysis:

- Bayesian analysis fails to offer any inductive learning from the evidence of the initial sequence 1, 3, 5, 7.
- Prior probabilities control the analysis, but the requirement that they normalize prevents them being set in a manner that is universally benign.

To proceed, we will see how a Bayesian analysis might help us decide between two extensions of the sequence 1, 3, 5, 7:

The odd numbers, H_{odd}: 1, 3, 5, 7, 9, 11, 13, 15, …

The odd primes with one, H_{prime*}: 1, 3, 5, 7, 11, 13, 17, …

using the evidence

E: 1, 3, 5, 7

The ratio form of Bayes' theorem asserts:

$$\frac{P(H_{odd} \mid E)}{P(H_{prime*} \mid E)} = \frac{P(E \mid H_{odd})}{P(E \mid H_{prime*})} \cdot \frac{P(H_{odd})}{P(H_{prime*})}$$

[22] One way is to consider not the incremental distances, d(t), but the total distance, s(t), fallen by time t. Then it is easy to show that the invariance is satisfied by setting s(t) proportional to t^p for any p>0. However, showing that these are the *only* laws satisfying the invariance is harder.

Since each of H_{odd} and H_{prime*} deductively entails E, we have $P(E| H_{odd}) = P(E| H_{prime*}) = 1$. Therefore Bayes' theorem reduces to:

$$\frac{P(H_{odd} \mid E)}{P(H_{prime*} \mid E)} = \frac{P(H_{odd})}{P(H_{prime*})}$$

What, according to the theorem, have we learned from the evidence E? The prior probabilities $P(H_{odd})$ and $P(H_{prime*})$ represent our initial uncertainty about the two hypotheses; the posterior probabilities, $P(H_{odd}|E)$ and $P(H_{prime*}|E)$ represent their new values, once we have incorporated evidence E. The reduced form of Bayes' theorem just tells us that conditionalizing on the evidence makes no difference to our comparative uncertainty concerning the two hypotheses. The ratio of the prior probabilities is the same as the ratio of the posterior probabilities. This will be true for any pair of hypothesized sequences that start with 1, 3, 5, 7. In short, we have learned nothing new from the evidence as far as our decision between the two hypotheses are concerned.

Hypotheses logically incompatible with the evidence will be eliminated. Take, for example:

The natural numbers, H_{nat}: 1, 2, 3, 4, 5, 6, …

Since H_{nat} is logically incompatible with E, we have $P(E|H_{nat}) = 0$ and the posterior probability will be $P(H_{nat}|E) = 0$. However this result is not an inductive result. We have simply eliminated all hypotheses deductively incompatible with the evidence. That deductive result is easily gained without the probability calculus or any other inductive manipulations. Where we need help is with the inductive problem. Does the evidence E favor some hypotheses among those deductively compatible with it? Here the Bayesian analysis has failed to provide anything useful. Our inductive preferences are exactly the same before we learn the evidence as they are after we learn it.

This is a quite discouraging start. However, it will be instructive to press on and ask what our posterior probabilities may be with specific prior probabilities. The analysis bifurcates according to whether we are subjective or objective Bayesians. If we are subjective Bayesians, then our prior probabilities are merely expressions of prejudice, constrained only by compatibility with the axioms of the probability calculus. We might decide that those prejudices dictate that the H_{odd} has three times the probability of H_{prime*}. Then we conclude for our posterior probabilities that

$$P(H_{odd}|E) = 3\ P(H_{prime*}|E)$$

Looking at the equation, it may seem we have learned something. But we have not. The threefold difference in posterior probabilities is a direct restatement of our prior prejudices.

If we are objective Bayesians, we will seek prior probabilities that objectively reflect what we know. In this case, by supposition, we know nothing initially, so we have no reason to

prefer one hypothetical sequence over any other. Hence the appropriate prior probability will assign the same, small probability ε to each hypothesis. That is, we have

$$P(H_{odd}) = P(H_{prime*}) = \varepsilon$$

The reduced form of Bayes' theorem now tells us:

$$P(H_{odd}|E) = P(H_{prime*}|E)$$

Once again, we have learned nothing. Our initial assumption was that all hypotheses are equally favored and that remains true for any pair compatible with the evidence.

This last conclusion overlooks a complication that will gravely trouble both subjective and objective Bayesians. The prior probability distribution must normalize; that is, the prior probabilities assigned to all the possible sequences must sum to unity. There is an uncountable infinity of possible sequences.[23] This means that, in a strong sense of most, most sequences must be assigned zero prior probability. Once a sequence has been assigned zero prior probability, its posterior probability on any evidence whatever will also be zero. That means that no evidence, no matter how favorable, will move us to entertain the sequence in the slightest. Hence both subjective and objective Bayesians must make unavoidably damaging decisions, prior to any evidence, as to which few sequences will be learnable.

Of course there are ways we might try to work around the problem. We might try to retain the uniform prior probability distribution simply by dropping the requirement of normalization and using so-called "improper priors." This violation may be excused if it turns out that, after conditionalization, the posterior probability distribution is normalizable. That normalizability is not achieved in this case however. There are infinitely many sequences beginning with 1, 3, 5, 7. After we conditionalize on this evidence, we will be assigning equal non-zero probability to each in this infinity of sequences. Normalization will fail.

More drastically, we might retain a uniform prior probability distribution by the artifice of simply choosing a finite subset of sequences and casting the rest into the darkness of zero probability. If we eschew the uniformity of prior probabilities for variable probabilities, we can expand the set of sequences with non-zero prior probabilities to a countably infinite set. As long as the prior probabilities diminish fast enough as we proceed through the set, the sum of the

---

[23] To see that the set is at least continuum sized, note that a subset of sequences using the digits 1 and 2 only can be mapped one-one onto the reals in the interval [0,1]. The sequence 1, 1, 2, 2, 1, 1, 2, 2, … is mapped to the fraction in binary notation 0.00110011…, etc. To see that the set is no bigger, note that we can map any sequence to a real in [0,1] by replacing the symbol "," by the symbol "0". The sequence 1, 3, 5, 7, 9 , 11, 13, … is mapped to the real .1030507090110130…, etc. The map is not "onto" because some reals, such as 0.100010001 have no corresponding sequence.

probabilities can be unity, as normalization requires. One way of achieving this diminution is to assign these varying non-zero probabilities only to sequences that are arbitrarily long, but always of finite length. If we do this, we need some rule to decide which sequences are more and which less probable. A popular choice is to use a prior probability distribution advocated by Solomonoff (1964). Briefly describable sequences like 1, 2, 1, 2, 1, 2, … have greater prior probability than ones with no simple description. This is implemented by penalizing each sequence's probability by an exponential factor $(1/2)^N$, where N is the length of the shortest description possible for the sequence.[24] Bayesian analysis that employs this prior probability distribution is celebrated with joyous but naïve enthusiasm as providing a "complete theory of inductive inference" (Solomonoff, 1964, p. 7) or "universal induction" (Rathmanner and Hutter, 2011)

The difficulty is that the comparative judgments of this prior probability distribution will never go away. They determine how we might discriminate between H_{odd} and H_{prime*} on learning evidence E = 1,3, 5, 7. Thus the selection of this prior probability distribution is not benign. It must be justified by something solid. Are we to suppose that, as a quite general proposition, our world favors sequences with short Turing machine programs? This favoring might be credible in specific contexts, such as one in which we know that people are thinking up the sequences. But we are to suppose this favoring is true prior to any restriction whatever on where these sequences may appear. It is hard to see any reason for why the world, as a quite universal matter, would prefer to present us with number sequences that are computable and do so in way that exponentially penalizes sequences with longer programs. The literature supporting the Solomonoff approach believes otherwise and matches its joy in its solution of the inductive problem with equally joyous pronouncements grounding the approach. They often resort to appeals to simplicity through "Occam's Razor" (Solomonoff, 1964, p.7; Rathmanner and Hutter, 2011, p. 1101). This reveals an inflated reverence for the insights of a medieval scholastic who wrote six centuries before Turing conceived the notion of a universal Turing machine. For more deflation of simplicity, see Chapter 6 here.

In short, the challenge of accommodating the requirement of normalizability greatly complicates the analysis. More generally, the Bayesian analysis itself creates troubles that multiply and whose intractability deepens the more we try to resolve them. We could continue to wrestle with them. Or we could see that the very fact that we face lingering problems of this gravity is telling us that Bayesian analysis is just the wrong instrument for this inductive problem. Compare that with the simplicity of the material analysis of the problem of extending 1, 3, 5, 7.

[24] N is usually taken to be the length of the shortest Turing machine program that would output the sequence.

Once we locate the appropriate context, as in Galileo's law of fall, we find that the requirement of invariance under a change of the unit of time fixes the extension all but completely.

10. Warranting Facts

What might other warranting facts look like? Once we realize that familiar facts may serve also to warrant inference, we see that we are surrounded by such warranting facts.

Cosmology seeks to discover the structure of the universe on the largest scale. If the universe is infinite in spatial extent, then the finite portion observationally accessible to us is minuscule. What we see is infinitely outweighed by what we cannot see. The essential assumption that allows us to proceed from what we can see to what we cannot is the "cosmological principle." It asserts that the universe is roughly homogenous in its large-scale properties. While this wording is a little vague, standard applications of the principle employ it unambiguously. In our vicinity, matter is distributed roughly uniformly in galaxies in a space of constant, possibly zero, curvature. The cosmological principle authorizes us to infer that this condition obtains everywhere in the whole universe. Much of modern cosmological theory proceeds from that authorization.

Assume we have some isolated system with a given quantity of energy and entropy. The principle of the conservation of energy, the first law of thermodynamics, authorizes us to infer that, however else it changes, this same isolated system will have the same energy at any future time. The second law of thermodynamics authorizes us to infer a similar conclusion about the entropy of the system: it will be the same or greater. A careful statement of the second law allows merely that, with very high probability, the entropy of such systems will be the same or greater. Hence the conclusion is warranted inductively, but with very great certainty.

Assume we have some experiment performed in an isolated laboratory. The principle of relativity authorizes us to infer that a uniformly moving replica of the experiment will yield the same result. A more careful factual statement of the principle allows that it holds only in regions of spacetime that are remote from intense gravitational fields and thus unaffected by the curvature of spacetime revealed through the general theory of relativity. So the factual principle informs us that, *mostly*, the same experimental result will obtain. Thus the inference is inductive.

This series of examples is designed to implement a progression in two aspects. First we progress from the more general to the more specific and local. Second, we progress from examples in which the mediating facts authorize the conclusion deductively to those in which they authorize them inductively. The next and final example extends the progression quite far to a case of greatly narrowed scope and greater inductive risk.

Assume we set up some simple chemical process whose feed includes nitrogen gas. A general fact of chemistry is that nitrogen gas is quite unreactive. Its diatomic molecules are held together by a strong triple bond that is hard to break. This general fact authorizes us to infer, at some relatively high level of inductive certainty, that the simple chemical process will leave the nitrogen gas unaltered. We are not assured of the conclusion with deductive certainty. There are extreme conditions under which nitrogen gas can be compelled to enter into reactions. Finding them was the Nobel Prize winning work of Haber and Bosch a century ago. Their Haber-Bosch process enables the chemical industry to synthesize ammonia from nitrogen and thereby to manufacture both fertilizers and explosives.

This progression gives us factual principles of increasingly narrower scope that warrant inferences inductively. The material theory of induction places no lower limit on the size of the domain over which these factual principles operate.

11. A Non-Contextual, Formal Logic is Exceptional

The scope of successful applications of deductive logics that are non-contextual and formal is enormous. It is one of the great achievements of human thought. Its success makes it easy to think that the right way and the only way to analyze inference is with non-contextual, formal theories. Correspondingly, then, one might think of a materially warranted logic as some kind of failure, perhaps the result of insufficient efforts to find that elusive, universal formal logic of induction. I will argue in this section that the success of non-contextual, formal accounts of deductive logic is exceptional. Hence, we should not use our familiarity with deductive logic to set our expectations for inductive logic. We should not allow it to make us expect that there is a non-contextual, formal logic of induction.

11.1 The Undeserved Success

Which are the good deductive inferences? As long as the problems are kept simple, most people have a pretty good instinctive grasp of which are the deductive consequences of their knowledge and they manage without external guidance. However the limits are readily breached. If each thing has a cause, does it follow deductively that there is one ultimate cause for all things? If for every moment of time there is a later moment of time, does it follow that time endures infinitely? Novices relying on instinct can readily falter in the face of such traps. Can we find an instrument that systematically and reliably separates the good deductions from the bad? The means of discerning the good deductions is so familiar to anyone who has had contact with modern logic that it is easy to underestimate the difficulty of the problem.

This problem was all but solved millennia ago with a simple, profound observation. If you know that

"All electrons have spin half."

then you know that

"Some electrons have spin half."

That deductive inference is assured even if you have no idea of what an electron is and even less of an idea of what "spin half" is all about. You can make the inference merely by attending to the form of the sentences and ignoring the material. You start with "All A's are B." and know that you are then authorized to infer to "Some A's are B." You can ignore all the fussy stuff about electrons and spin. All you need to watch is the form of the sentences.

That deductive inference can proceed in such a simple and efficient manner is a marvel. It is the basis of a formal theory of inference, for we separate out the allowed inferences from the prohibited inferences merely by looking at their form. Specifying the logic then merely amounts to providing a list of schemas, such as

All A's are B.

Therefore, some A's are B.

To use them, we replace A by anything we like and B by anything else we like and—bingo!—there's a valid deductive inference.

One sees in this example that the success of the schema depends upon the non-contextuality of deductive inference. We can transport this schema to any domain, substitute anything for A and B and still be assured that a valid inference results.

This simple schema is just the beginning. Generations of logicians have supplied us a growing repertoire of schemas that embrace many logical operators. We have sentential logics that employ the connectives "not," "or" and "and." One of De Morgan's laws is the schema

not-(A and B)

Therefore, (not-A or not-B)

Predicate logics include individuals and their relational properties and they allow us to quantity over the individuals. If all things "x" gravitate "G(x)," then it is false that something exists that does not gravitate. This is an application of the schema:

For all x, G(x).

Therefore, not-(there exists x, not-G(x))

Modal logics introduce modal operators like "It is possible that…" and "It is necessary that…" Tense logics introduce temporal operators such as "It is always…" and "It is sometimes…"

11.2 Context Dependence of Connectives

In the face of these successes, it may seem that the scope of formal methods in logic in unlimited. However, lingering, recalcitrant anomalies limit the scope of the formal approach. These anomalies manifest in deductive logic when the logical terms used have meanings that are context dependent. Does "some" just mean "at least one"? Or does it mean "more than one but not too many"? The answer varies with the context. Consider the mathematical assertion:

For some x, the quotient 1/x is undefined.

Here "some" can mean "one or more" and the single case of x=0 is the one that makes the sentence true. However take the "some" of:

Some voters disapprove of the governor's decision.

This "some" refers to more than one voter, but probably not a majority. This difference matters to the formal theory, for not all the schemas we may wish to use for "some" will apply everywhere. Consider

Some A's are B.

Therefore, more than one A is B.

It applies to the "some" of the voters but not to the "some" of division by x. The schema is context dependent; it is not universally applicable.

The humble conditional "If A then B." has proven to be a more notorious locus of this sort of trouble. A natural understanding is that this conditional is true when knowing A authorizes you to know B as well. That is, the conditional can be a premise in the argument form "modus ponens":

If A then B.

A.

Therefore, B

That function is served by the "material conditional." According to it, "if A then B" is just the same as "Either A is false or B is true." Thus, if we happen to know that A is true, then we know the first option ("A is false.") fails. So that leaves the second, "B is true." Hence the material conditional has done the job of allowing us to proceed from knowing A to knowing B.

That may seem quite fine until one realizes that, under this understanding, the conditional "if A then B." comes out true whenever B is true, no matter what A says. That is, both

"If pig have wings, then the sky is blue."

"If the grass is green, then the sky is blue."

turn out to be a true, material conditionals, just because the sky is blue. The natural objection is that an "if A then B" statement can only be true if there is something in the antecedent A that makes the consequence B true. That fails in these last examples. Whether pigs have wings or the grass is green is irrelevant to the blueness of the sky. But

"If the sunset is red, then the sky is blue."

can be a true conditional. For the sunset is red because the blue light from a setting sun has been scattered away by the air; and that blue light comprises the blue sky. The blue of the sky is directly relevant to the red of the sunset.

Ingenious systems of relevance logic have sought to formalize the schemas into which "if…then…" properly enters, if understood relevantly. However deciding just what is relevant to what is a delicate issue that may embroil us in significant portions of science. The blueness of the sky results from Rayleigh scattering of blue light by the nitrogen and oxygen atoms of the air that just happen to be the right size for the job. So arcane facts in atomic theory are also relevant, but perhaps they are not as directly relevant as the redness of the sunset. That tells us that relevance is context dependent and may vary in strength. Indeed relevance may prove to be so diffuse that it may not be possible to separate off a small, tight formal logic of relevance as anything other than a crude gloss on a richer relation that is inextricably connected with the factual material of the science.

More generally, the success of universally applicable formal logics of deduction depends on deductive inference being non-contextual. Whenever simple connectives fail to have a non-contextual meaning, as in these examples, the logics in which they appear cease to be universal.

11.3 Sellars' and Brandom's Material Inference

The anomalies for a formal theory of deductive inference above focused narrowly on logical connectives (if…then…) and operators (Some…). They have a context dependent meaning, I argue, that is incompatible with their universal applicability, or least they cannot have it if we fix their meanings once and for all. Wilfrid Sellars and Robert Brandom have developed a broader and more powerful critique of a formal approach to inference in general, not just deductive inference.

Their concerns are not limited to connectives but to all the terms appearing in the inferences. Their core idea is that the meaning of the terms in propositions is what makes good the inferences in which they correctly appear. Brandom (2000, p.52) displays the inference from:

"Pittsburgh is to the west of Princeton"

to

"Princeton is to the east of Pittsburgh."

We recognize this as a good inference, but not for formal reasons. Rather it is good because of the contents of the concepts of east and west. That is, the matter of the inference makes it good.

When I developed the material theory of induction, I was not aware of Sellars' and Brandom's notion of material inference and, in particular, Brandom's use of the term "material

inference." I learned of it through a lovely note written by Ingo Brigandt (2010), which usefully develops and applies the notion of material inference.

The difficulty is that our notions of material inference differ slightly, as far as I can see. That means that it would have been better at the outset if I had chosen another name. For Brandom, the above inference is material since it is made good by the concepts invoked in the premises. In my view, it is material since I locate the warrant for the inference in the background material fact: if something is east of something else, then the second is west of the first. Here I leave open whether this difference is consequential or merely a different entry point into a collection of views that largely agree.

References

Brandom, Robert (2000) *Articulating Reasons: An Introduction to Inferentialism.* Harvard University Press, 2000.

Brigandt. Ingo (2010) "Scientific Reasoning Is Material Inference: Combining Confirmation, Discovery, and Explanation," *International Studies in the Philosophy of Science*, **24**, pp. 31-43.

Drake, Stillman (1969) "Galileo's 1604 Fragment on Falling Bodies (Galileo Gleanings XVIII)," *British Journal for the History of Science*, 4, pp. 340-3583.

Drake, Stillman (1978) *Galileo at Work: His Scientific Biography*. Chicago: University of Chicago Press. Repr. Mineola, NY: Dover, 2003.

Galilei, Galileo (1623) *The Assayer*, pp. 231-280 in Stillman Drake, ed. and trans., *Discoveries and Opinions of Galileo*. New York: Doubleday & Co., 1957.

Galilei, Galileo (1638) *Dialogues Concerning Two New Sciences.* Trans. Henry Crew and Alfonso de Salvio. MacMillan, 1914; repr. New York: Dover, 1954.

Huygens, Christiaan (1888) *Oeuvres Complètes de Christiaan Huygens*. Vol. 1. La Haye: Martinus Nijhoff.

Mill, John Stuart (1904) *A System of Logic: Ratiocinative and Inductive.* New York and London: Harper & Brothers Publishers.

Norton, John D (2000) "'Nature in the Realization of the Simplest Conceivable Mathematical Ideas': Einstein and the Canon of Mathematical Simplicity," *Studies in the History and Philosophy of Modern Physics*, **31**, pp.135-170.

Norton, John D (2003) "A Material Theory of Induction," *Philosophy of Science*, **70**, pp. 647-70.

Norton, John D (2005) “A Little Survey of Induction,” in P. Achinstein, ed., *Scientific Evidence: Philosophical Theories and Applications*. Baltimore: The Johns Hopkins University Press, 1905. pp. 9-34.

Norton, John D. (2014) “A Material Defense of Inductive Inference.” *Synthese*. **191**, pp. 671-690.

Norton, John D. (2014a) “Invariance of Galileo’s Law of Fall under the Change of the Unit of Time.” http://philsci-archive.pitt.edu/id/eprint/10931.

Rathmanner, Samuel and Hutter, Marcus (2011) “A Philosophical Treatise of Universal Induction,” *Entropy* **13**, pp. 1076-1136.

Salmon, Wesley C. (1953) “The Uniformity of Nature,” *Philosophy and Phenomenological Research*, **14**, pp. 39-48.

Solomonoff, Ray (1964). "A Formal Theory of Inductive Inference," *Information and Control* **7** pp. 1–22; pp. 224-254.

Chapter 3

Replicability of Experiment.[25]

1. Introduction

The general idea is simple and instantly compelling. If an experimental result has succeeded in revealing a real process or effect, then that success should be replicated when the experiment is done again, whether it is done by the same experimenter in the same lab ("repeatability") or by others, elsewhere, using equivalent procedures ("reproducibility"). It is, at base, the same idea that leads us to the near universal reaction when a conjurer makes a coin vanish. "Do it again!" we demand. And this time, we will watch more closely.

One readily finds enthusiastic endorsements of the idea in the scientific literature. The opening sentence of a special section in *Science* on "Data Replication and Reproducibility" says (Jasny et al, 2011):

> Replication—the confirmation of results and conclusions from one study obtained independently in another—is considered the scientific gold standard.

An editorial in *Infection and Immunity* on "Reproducible Science" begins its abstract with an unequivocal: "The reproducibility of an experimental result is a fundamental assumption in science." (Casadevall and Fang, 2010, p. 4972) There are few if any doubts about the notion. The principal locus of concern is that replication can be hard to achieve, either because of the difficulty of replicating pertinent conditions or through lack of institutional rewards to the replicating experimenters.

My concern in this chapter is inductive logic. Might replicability provide a universal schema or principle that figures in a formal logic of induction, or at least in that portion of the logic that treats experiments? I will seek to establish in Section 2 that a principle of replicability cannot be given a general formulation that would allow it to serve in a formal logic of induction. I will argue that attempts to find such a general principle collapse under the weight of mounting

[25] A self-contained adaptation of this chapter has been published as "Replicability of Experiment," *Theoria*, **30** (2015), pp. 229-248 under a Creative Commons License: Attribution-Noncommercial-No Derivative Works 4.0 Generic.

complexities arising from the multitude of conditions and outcomes associated with replicability. Rather, successful inductive inferences associated with replicability should be understood as materially warranted. We can identify background facts that authorize the relevant inferences on a case by case basis, without the need for a universal principle. The types of background facts that serve this function are described in Section 3. Once we have identified these facts, the search for a general principle becomes unnecessary, in so far as we are interested in finding the warrants of our inferences. Sections 4 to 7 will develop case studies that show that the import of replication or its failure can be upheld or denied in all possible combinations. This reduces a principle of replicability to one that works except when it does not. We will see at the same time, however, that the successes and failures of these examples are explicable materially. Conclusions are in Section 8.

My goal is *not* to discourage replication of experiments. On the contrary, replication is a powerful way to strengthen the evidential basis of our hypotheses and theories. Rather this analysis is intended only to impugn the idea that replication gains its evidential power from some universal inductive principle of replication.

Before proceeding, we need a brief terminological digression: the terms "repeatability," "reproducibility" and "replicability" are often used loosely and interchangeably. In some contexts, they have been given precise definitions. There, repeatability indicates as exact a replication of all conditions as possible, including the same operators and apparatus; whereas reproducibility calls for changes of these conditions.[26] I will use the terms replication and

---

[26] In the narrower context of standardized measurement, the *International Organization for Standardization* has decreed (ISO 21748:2010(E), p. 3): "Repeatability conditions include: the same measurement procedure or test procedure; the same operator; the same measuring or test equipment used under the same condition; the same location; repetition over a short period of time. Reproducibility requires only that the measurement must reappear under changed conditions. That is, (ISO 21748:2010(E), p. 3): "reproducibility conditions[:] observation conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in different test or measurement facilities with different operators using different equipment[.]"
Source: "Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimates," Publication ISO 21748: 2010(E).
Similar definitions are found in *National Institute of Standards and Technology*, NIST Technical Note 1297 (1994), Definitions D.1.1.2 and D.1.1.3 and in the *International Union of Pure and Applied Chemistry*'s "Gold Book": *Compendium of Chemical Terminology*, 2nd ed.. Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997).

replicability to cover both notions. Most of the general analysis below applies equally to repeatability and reproducibility.

2. Failure of Formal Analysis

What kind of an inductive notion is replicability? If we wish to pursue a formal analysis, is it possible to state it as a general principle? A good start is this:

> Successful replication of an experiment is a good indicator of a veridical experimental outcome; failure of replication is a good indicator of a spurious experimental outcome.

This is far from a self-contained principle. Each term needs further explication. The more straightforward are the notions of veridical and spurious experimental outcomes:

> A veridical experimental outcome is one that properly demonstrates the process or effect sought by the experimental design.
>
> A spurious or artefactual experimental outcome fails to do so; it arises from an unintended disruption to the experimental design.

This is a rich enough characterization for us to proceed, even though many details are left open.

How close have we come to a universal inductive principle? Do we have an inductive analog of the universal, formal principles of deductive logic? In asking, we should bear in mind what the latter are like. One such universal deductive principle is the law of the excluded middle. It asserts:

> For any proposition P, either P is true or P is false.

This deductive principle is a schema: we can insert any proposition we like for "P" and recover a truth, the application of the principle to that proposition. It is self-contained. There are no tacit conditions limiting just which propositions can be substituted for "P"; and there is no ambiguity in what is meant by the truth or falsity attributed to the proposition (Or at least there are none beyond the usual evasions made by philosophers when they have to use these terms.)

It is quite different with the above characterization of replicability of experiment. The first difficulty is that the characterization includes many notions that require elaboration if the characterization is to rise to the level of precision of the law of the excluded middle. Just what is "a process or effect sought by the experimental design"? Just when is a second experiment replicating an earlier experiment as opposed to being a different experiment that looks similar to it? Elaborating these and similar questions is likely to be tedious and unlikely ever to yield a formulation that can stand without the need of further elucidation.

The second difficulty is more serious. The characterization employs inductive notions whose explication is unlikely to be achievable by formal means. It speaks of "good indicators."

This is an inherently vague notion. In the case of a single successful or failed replication, the strength of the indication can vary over a wide range. Presumably there is some idea that multiple, successful replications are better than just one. How much better are they? Is there a point of diminishing returns? When there are some successes of replication and some failures, how do we trade them off to come to our final assessment? Somehow the formal analysis will need to specify in general, abstract terms how all this accountancy is to be effected.

Finally the most serious problem facing a formal analysis of replicability is that the principle appears to be defeasible in every way possible. That is, there are cases of *successful* replication in which the replication is judged to be a strong indicator of a veridical outcome; and cases in which the success is judged epistemically inert. In the reverse direction, there are cases of *failure* of replication that are judged to be a strong indicator of a spurious outcome; and cases in which the failure is judged epistemically inert. Thus a full statement of the principle must provide independent criteria for when it applies or when it does not. Without such independent criteria, it becomes the sad specter of the principle that applies except when it does not.

Looking ahead, most of this chapter will be devoted to displaying examples in which all these combinations of success and failure are realized. The examples to be developed are listed in the table:

	Import of replicability upheld	Import of replicability discarded
Successful replication	H. Pylori Stomach Ulcers (result accepted as veridical)	Intercessionary prayer (result rejected as spurious)
Failed replication	Cold fusion (result rejected as spurious; and skeptics discount cases of successful replication)	Miller experiment contradicts relativity theory (relativity theory upheld)

Table 1. Illustrations of all combinations of success and failure of replicability.

The "import of replicability" refers to the standard reading: successful replication indicates a veridical outcome; failure of replication indicates a spurious outcome. In the cases in the first column, the import of replicability is upheld as expected; in those of the second, it is discarded.

These three difficulties present formidable challenges to formulating a precise principle of replicability: it must be complete enough not to need further explication of its central terms; it must replace the vague inductive term "good indicator" with something that allows precise accountancy for multiple successes and failure; and it must define independent conditions of

applicability flexibly enough accommodate the full range of cases in which replication or its failure is taken to be epistemically significant or epistemically inert.

3. A Material Analysis

While a formal account of replicability faces formidable obstacles, a material analysis will prove to have little trouble passing these same obstacles. The hard question of whether successful replication or its failure is epistemically significant or inert is answered on a case by case basis. The inductive import of each outcome is determined by the particular facts obtaining in the background of each case. They warrant the inductive arguments that proceed from those outcomes.

Ultimately, each case is unique and requires its own detailed analysis. However, at a more superficial level, it is possible to identify two general classes of background facts that serve to license the different inferences associated with replicability in each case. These facts are not narrowly associated just with replicability. Rather they are facts that warrant the inference from the observed experimental outcome to the process or effect sought by the experimental design. Or, if they take an inhospitable form, they may warrant an inference from the observed outcome to the conclusion that it is spurious. These facts are:

A. Experimental conditions: these background facts specify conditions under which the effect or process of interest will manifest in a veridical experimental outcome.[27]

B. Confounding conditions: these background facts specify the conditions conducive to spurious experimental outcomes. These conditions simulate a veridical experimental outcome, when the sought effect or process is not present; or they may interfere sufficiently to produce an unsuccessful outcome, when the effect or process is present.

A familiar illustration of facts of type A and B arises in randomized controlled trials. We wish to determine if some treatment—a new drug, for example—is efficacious. We randomly assign subjects to a test and a control group, both blinded. The test group is given the treatment and the control group is given a placebo. If the outcome is a statistically significant, beneficial difference between the test and control group, we infer from it to the efficacy of the treatment.

The inductive inference to this conclusion is warranted by appropriate facts in class A and B. In class A is the key fact is that test subjects but not control subjects are given the treatment, so a beneficial difference between them can be due to the treatment. Implicit in this fact is another that is not commonly made explicit: that there is at least some possibility that treatment can bring about the effect. While this sort of fact is not one that we commonly call into

[27] This is sometimes called "construct validity."

question, it can be crucial. Critics of homeopathy (such as me) will refuse to accept that a controlled trial of a homeopathic remedy can demonstrate the remedy's efficacy, for the remedy contains no active ingredients by its formulation. Similarly, we shall see below that skeptics of the healing efficacy of prayer find just the corresponding sort of fact to be missing.

In class B, we require the facts that preclude a spurious outcome. Randomization is important here, for it assures us that the only systematic difference between the test and control group is the administering of the treatment, so that any ensuing difference between them can only be due to the treatment. Blinding is also important, so that the subjects and the result collecting experimenters do not know who is in the test or the control group. For otherwise, a statistically significant difference between the two groups might result from this knowledge itself, through the placebo effect or through the expectations of the experimenters recording results.

In short, the facts in class A warrant the inference to the conclusion that the efficacy of the treatment *can* be responsible for a positive outcome. The fact in class B warrant the inference to the conclusion that another factor *cannot* be responsible for a positive outcome. We combine the two to conclude that the efficacy of the treatment *is* responsible for a positive outcome.

Now let us return to replicability. With any experiment, we can be uncertain whether appropriate facts in classes A and B prevail. Successful replication does not test all of them. Rather it tests whether certain unfavorable confounding conditions of class B are present. If we get the same positive outcome when a different operator performs the experiment, then we know that the first positive outcome was not due (solely) to some infelicity associated with the first operator. By systematically replicating the experiment with different operators, different standards, different materials, different laboratories, and so on, we eliminate the possibility of confounding conditions associated with each of the factors listed. If we test for repeatability in the technical sense—that is we replicate the experiment with all these factors unchanged—we are testing whether some random error in the execution of one experiment might be responsible for a spurious outcome.

This seems so straightforward, how is it that we find prominent cases in which the normal import of replicability is denied? The reason is that this import involves the complete inference from the observed outcome to the sought effect or process. That requires facts in both classes A and B to support the inference. In some of the disputed cases discussed below, however, we find that the denial of the import of replicability results from a presumption of failure of facts in class A, which are not directly tested by replication. In one, however, we will find disagreement over whether confounding conditions of class B have been appropriately arranged.

In the following sections, we will see the four cases of Table 1 elaborated. In the case of intercessionary prayer, we shall see successful replication of experiments judged by skeptics to be insufficient to establish the process sought. Their reason is that they do not find the requisite

facts of class A do not obtain. In the case of cold fusion, we shall see that establishment skeptics and dissident supporters of cold fusion differ on the import of the mixed record of successful and failed replication. Their differences are traceable to differences of opinion on which facts in class A obtain. In the Miller relativity experiments, however, failure to reproduce an earlier experiment is judged not to impugn the earlier result, since its supporters became convinced that Miller had not eliminated confounding effects covered by facts in class B.

4. H. Pylori Stomach Ulcers: Successful Replication

In 2005, Barry Marshall and Robin Warren won the Nobel Prize in Physiology or Medicine with the citation reading "for their discovery of the bacterium Helicobacter pylori and its role in gastritis and peptic ulcer disease." Prior to their work, it had been assumed that stomach ulcers were caused by stress and spicy food. The idea that a bacterium may be involved was discounted. The stomach is highly acidic and bacteria do not tolerate such environments well.

By taking biopsies from 100 participant patients, as reported in their initial letter (Marshall and Warren, 1983), they were able to demonstrate an association between the presence of the bacterium Helicobacter pylori and gastritis and ulcers, with 100% association for duodenal ulcers. The importance of replication even at this early stage became clear when they sought to publish a more complete account. Warren (2005, pp. 301-302) recounts the decisive moment.

> We sent our definitive paper to the *Lancet* in 1984 ([Marshall and Warren, 1984]). Although the editors wanted to publish, they were unable to find any reviewers who believed our findings. Our contact with Skirrow became crucial here. We told him of our trouble, and he had our work repeated in his laboratory, with similar results. He informed the *Lancet* and shortly afterwards they published our paper, unaltered.

Contrary to a persistent myth, the new work was assimilated and rapidly repeated. As part of an account debunking this myth, Atwood (2004) reported:

> Within a couple of years of the original report, numerous groups searched for, and most found, the same organism. Bacteriologists were giddy over the discovery of a new species. By 1987—virtually overnight, on the timescale of medical science—reports from all over the world, including Africa, the Soviet Union, China, Peru, and elsewhere, had confirmed the finding of this bacterium in association with gastritis and, to a lesser extent, ulcers.

One replication was more of a media stunt than controlled science. To prove the association, Marshall drank a beaker of Helicobacter pylori and subsequently succumbed to gastritis.

This is a "text book" case of the proper functioning of replication and there is little in it to distinguish formal and material approaches. The earlier reluctance to accept Marshall and Warren's work is readily explained materially. As long as it was taken as a background fact that bacteria do not live well in the highly acid environment of the stomach, there are insufficient facts in the background to support for the facts in class A. Detection of bacteria can only be through some coincidental contamination. The successful inference from the presence of the H. Pylori bacteria to the conclusion that they cause gastritis and ulcers required acceptance of a new fact in class A: that bacteria with the capacity to cause gastritis and ulcers can survive in the stomach. The rapid replication of the outcome in many laboratories affirmed the requisite fact of class B: that their presence is not due to some confounding effect peculiar to Marshall and Warren's laboratory.

5. Cold Fusion: Failed Replication

The episode of controlled fusion is traditionally presented as one in which an avenue of research was closed because of failure of replication. At the most superficial level, that may be a correct description. However a closer look at the episode reveals something more complicated than the application of some principle of reproducibility. There certainly were many failed attempts at replication reported. However there were also many successful replications reported. This has lead to a bifurcation in the community into those who discard the idea of cold fusion (the establishment view) and those who continue to pursue it (a dissident minority). No simple inductive principle concerning replicability of experiment can capture the inductive reasoning associated with this bifurcation. It derives essentially from differences in the background assumptions of the groups. Talk of replication is really a gloss on more complicated inferences, as the material theory of induction indicates.

Traditional nuclear power generation derives from the fission—the splitting apart—of radioactive Uranium or Plutonium atoms. This fission is distinct from the nuclear reactions that power stars like our sun. They are driven by fusion—the joining together—of atoms of hydrogen and other light elements to form heavier elements. In the process, prodigious quantities of energy are released. It has long been a goal of the nuclear power industry to adapt fusion reactions to power generation. Their present terrestrial use has been limited to the uncontrolled fusion in hydrogen bombs. The difficulty is that enormously high temperatures are needed to smash the hydrogen atoms together sufficiently energetically to ignite a fusion reaction. Materials at these high temperatures are difficult to control in a power station and practical, fusion-based nuclear power generation remains a distant dream.

In March 1989, chemists Martin Fleischmann and B. Stanley Pons announced in a press release from the University of Utah that they had found a way of carrying out fusion reactions on a laboratory bench at ordinary temperatures. Their experiments did not use hydrogen but a heavier isotope of hydrogen, deuterium, in the form of deuterium oxide, also known as "heavy water." They electrolyzed the heavy water using palladium electrodes. Over a lengthy electrolysis, one of the palladium electrodes, the cathode, would become saturated with deuterium and, as a result, the individual deuterium atoms would be driven closely enough together to ignite a nuclear fusion reaction. At least, that is what they claimed had happened, on the basis of the large quantities of heat produced. These quantities were greater than could be recovered from chemical changes, they asserted. In one burst, the released heat had melted and vaporized part of the electrode, destroying some of the equipment. Then, Steven Jones, working at nearby Brigham Young University, revealed that he had been working largely independently on a similar cold fusion project and had experimental results involving not the generation of heat, but neutrons, a familiar signature of nuclear reactions.

Whether the researchers succeeded in igniting fusion reactions remains debated. However there is no doubt that they ignited a scientific and popular frenzy. The principal trigger was the possibility of a new process that would revolutionize the power generation industry. There was a scramble to replicate the cold fusion experiments in the US and internationally. The resulting episode was complex and fascinating on many levels. Cold fusion, if affirmed, would be a scientific discovery of the highest order. That lofty pinnacle was overshadowed by the possibility of new technology for a major industry and its lucrative patent rights. These financial motivations lent an uncommon urgency in what was otherwise the realm of arcane specialists. There were other tensions, such as the professional rivalry of physicists and chemists. Here were physicists failing to tame nuclear fusion with enormous, expensive devices. Now some chemists succeed with a project plotted in one of their kitchens and funded personally. Then there was a soap-opera quality to the rivalry between the Fleischmann/Pons and Jones projects. They had planned to coordinate their communications, but the arrangements had misfired and Fleischmann and Pons took the unusual course of announcing their discovery through a press release without Jones' knowledge.

Let us set all these complications aside and focus on the inductive inferences. While there was initially considerable confusion over the inductive import of the experiments, that confusion resolved within a year into two views and it has largely remained so bifurcated. The establishment response was that the experiments failed to demonstrate fusion on the lab bench and that only modest resources should be assigned to further research. The minority, dissident view was that a great discovery had been made and all efforts should be put into developing it.

We find a clear statement of establishment view in the November 1989 report of the Energy Research Advisory Board to the US Department of Energy (ERAB, 1989). It concluded in its Executive Summary that

> The Panel concludes that the experimental results on excess heat from calorimetric cells reported to date do not present convincing evidence that useful sources of energy will result from the phenomena attributed to cold fusion. In addition, the Panel concludes that experiments reported to date do not present convincing evidence to associate the reported anomalous heat with a nuclear process.

The Board was reserved in its recommendation for action:

> The Panel recommends against the establishment of special programs or research centers to develop cold fusion. However, there remain unresolved issues which may have interesting implications. The Panel is, therefore, sympathetic toward modest support for carefully focused and cooperative experiments within the present funding system.

The dissident community continued its research and, in 2004, was successful in pressing the US Department of Energy to reopen its evaluation. The community supplied a document, "New Physical Effects in Metal Deuterides," that was subjected to peer review and discussion. It was found (DOE, 2004) that "…the conclusions reached by the reviewers today are similar to those found in the 1989 review." The bifurcation remained unbreached.

Both sides deferred to reproducibility as a guiding standard. The 1989 Advisory Board report (ERAB, 1989) commences its preamble by noting the failure of reliable replication:

> Ordinarily, new scientific discoveries are claimed to be consistent and reproducible; as a result, if the experiments are not complicated, the discovery can usually be confirmed or disproved in a few months. The claims of cold fusion, however, are unusual in that even the strongest proponents of cold fusion assert that the experiments, for unknown reasons, are not consistent and reproducible at the present time.

However mere problems of reproducibility cannot be the principal basis for the solidly negative conclusions reached by the Advisory Board. For their report documents both successful and failed replications of the various types of experiments aimed at testing cold fusion. For example, in relation to experiments yielding excess heat, the report's Table 2.1 lists five experiments that found excess heat and thirteen that did not. While the ratio of five to thirteen certainly favors the no-heat result, it is hardly sufficient to dismiss the effect, especially when its reality, if demonstrated, would be of great utility.

The deeper grounding for the negative report is laid out early in the report (pp. 6-8) when answers are offered to the rhetorical question "Then why the skepticism?" The first reason is developed only in a few sentences: many researchers have been unable to replicate the excess heat effect; and these calorimetric measurements are technically rather difficult. The two remaining reasons are developed in some detail and amount to conflicts between the particulars of the positive experiments and the accepted science of nuclear reactions.

The second reason was summarized as:

> the discrepancy between the claims of heat production and the failure to observe commensurate levels of fusion products, which should be by far the most sensitive signatures of fusion.

The nuclear reactions proposed for cold fusion involve fusion of two deuterium atoms to produce other atoms. Various reactions were possible and they would yield tritium, isotopes of helium or other products. The quantities of these fusion products detected did not match the quantities of heat reported. It was as if one burns wood in a fire. From the heat generated, one can determine how much wood ash must fall through the grate. The positive experiments were not finding the right amounts of ash.

The most important discrepancy was in neutron production. The most likely fusion reactions would produce neutrons and in large quantities. The report noted:

> The initial announcement by Pons and Fleischmann in March 1989 exhibited the discrepancy between heat and fusion products in sharp terms. Namely, the level of neutrons they claimed to observe was 10^9 times less than that required if their stated heat output were due to fusion.

This discrepancy was noted very early by critics and, by itself, was deemed sufficient for instant dismissal of the claims of cold fusion. Here is how one popular narrative from 1989 reported the problem (Peat, 1989, p. 82)

> According to Robert L. McCrory of the University of Rochester's Laboratory of Laser Energetics, for example, if nuclear fusion was really taking place, then the only way to make sense of all that heat was to have a trillion neutrons being emitted each second—enough to kill everyone in the room.
>
> By now the following joke had begun to circulate around the world's laboratories:
>
> FIRST SCIENTIST: Have you heard about the dead-graduate-student problem?
>
> SECOND SCIENTIST: No, what's that.
>
> FIRST SCIENTIST: There are no dead graduate students.

The third reason was summarized as "cold fusion should not be possible based on established theory." Deuterium does not undergo fusion reactions under normal conditions

because the electrostatic repulsion of the nuclei prevent its atoms approaching closer than about 0.1 nanometers, which is too great a separation for a nuclear reaction to start. The hope of the cold fusion researchers was that a palladium electrode could be so densely laden with deuterium that sufficiently close approaches would occur. The report, however, disputed these hopes. The closest approach of deuterium atoms in palladium is just 0.17 nanometers. That is over twice the distance (0.074 nanometers) separating two deuterium atoms in molecular deuterium, D_2. The cold fusion researchers would be bringing the deuterium atoms closer if they merely left them in the form of free molecular deuterium.

Supporters of cold fusion also defer to the idea of reproducibility. Sturms (2007, p. 49) initiated the discussion of the challenges to cold fusion with the resounding affirmation:

> Replication is the gold standard of reality. If enough people are able to make an effect work, the consensus of science and the general public accept the effect as being real and not error or figment of imagination.

He affirmed that replication has been successful:

> A Myth has formed about cold fusion not being duplicated, being based on error, and being an example of "pathological science", […] i.e. wishful thinking. None of this description is correct. The basic claims have been duplicated hundreds of times and continue to be duplicated by laboratories all over the world, although success is difficult to achieve.

However he also allowed that the replication has not been uniformly successful (p. 117):

> Replication occurs when other people observe the same effects using essentially the same conditions. Unfortunately, in the case of cold fusion, the required conditions are not known. Occasionally, when a lucky combination of conditions has been created, the effects are observed. These effects have been seen many times, as the results listed throughout the book demonstrate, but not always on command. This failure of the effects to occur every time they are sought has become a major issue for the field and needs to be examined in detail because some confusion exists about what replication actually means.

The record of successful replication was reinforced with massive tables listing many successes. The table listing experiments that report successful "anomalous power" production spans nearly ten pages (pp. 52-61).

Sturms came to very different conclusions than the Advisory Board concerning cold fusion. He regarded cold fusion as established fact to be announced with text-book like certainty (p. 190):

> The phenomenon of cold fusion or low energy nuclear reaction occurs in an unusual solid or even within complex organic molecules. A variety of nuclear reactions are initiated, depending on the atoms present. Some of these reactions occur at a rate sufficient to make measurable heat. The most active reaction produces ^{4}He when deuterium is present. Other reactions occur at lesser rates, but rapidly enough to accumulate detectable nuclear products.

Where the Advisory Board report found the existing theory of nuclear fusion secure and unfavorable to cold fusion, Sturms inverted the relation and impugned the theory for its failure to accommodate experiment.

His treatment of neutron emissions illustrates this inversion. Standard nuclear physics allows for deuterium to fuse in several ways. The most probable reactions yield high neutron and proton emissions. The reaction favored by cold fusion supporters was the fusing of two deuterium atoms to yield a ^{4}He atom, for that reaction involved only gamma ray emission, but no neutrons. The difficulty is that the neutron free reaction is weaker by a ratio of 10^7 in cross-section than the other reactions. Somehow the novel environment of the cold fusion experiment would need to bring about a great enhancement of this reaction. This, the Advisory Board, found to be a fatal problem (ERAB, 1989, Sect. B.2):

> We know of no way whereby the atomic or chemical environment can effect such an enhancement, as this ratio is set by nuclear phenomena and is on a length scale some 10^4 times smaller than the atomic scale.

The point is mildly stated, but the idea is powerful. Fusion reactions involving deuterium had been well researched and well understood. Proponents of cold fusion had to argue that this established theory fails for some as yet unknown reason when the fusion reaction occurs within a palladium electrode. Effects of this type were otherwise unknown and implausible because fusion requires a closeness of approach of the deuterium atoms at which scales the palladium atoms are distant spectators.

Sturms (2007, p. 13) took a different view:[28]

> If theory and observation are in conflict, theory wins [in the skeptics view]. In this case, the absence of neutrons proved that the effect does not occur even when tritium and extra heat are measured, because theory requires neutrons be produced. In their minds, the extra heat must be a measurement error and the

[28] I have not found an establishment response to this argument, but it is not too hard to imagine its content: the establishment view is not rejecting evidence, but considering a larger class that includes the experiments and observations in other arenas that support the standard theory of fusion reactions.

> tritium must be contamination. Evidence to the contrary was simply ignored. This is how faith-based science operates, but not the kind of science we are taught to respect. On the other hand, reality-based science acknowledges what nature reveals and then attempts to find an explanation. Rejection occurs only if a satisfactory explanation cannot be demonstrated. This demonstration is still in progress for cold fusion.

In sum, the real basis of the varying appraisals of cold fusion lay in inductive inferences grounded by background facts of class A. These facts specified the conditions under which cold fusion would manifest experimentally. In the establishment view, these facts called for rates of neutron and other fusion production not reported in the experiments; and, in addition, these facts denied that deuterium saturated electrodes could bring the deuterium atoms close enough to ignite fusion in the first place. Hence these facts warranted the inference to the conclusion that the experiments had failed. The dissidents, however, were willing to conjecture looser background theories, including some undeveloped or even unknown theories that would warrant the inference from the experimental results to cold fusion. Both deferred to the idea of reproducibility. Yet, with the same record of experiment, they came to different conclusions.

My proposal is that they are not calling upon a universal principle of reproducibility that resides within some abstracted, logic of induction. Rather, the idea of reproducibility is merely a gloss on inferences that are quite specific to the case at hand and dependent essentially on background assumptions. It is exactly because the two groups differ in their background assumptions that they can come to judge different inferences warranted.[29]

[29] According to the material theory, that does not mean that both inferences are sound. The situation is little different from the corresponding case of deductive logic. If two scientists employ the same premises but different deductive schema to arrive at contradictory conclusions, at least one of the schema is a fallacy. Correspondingly, if two scientists arrive at differing conclusions by inductive inference, at least one has a false warranting fact presumed.

6. The Miller Experiment: Failed Replication with no Inductive Import[30]

How are we to deal with a case in which there are multiple successful replications of an experiment, but a prominent, well-executed failure? Understood as a formal principle, reproducibility gives us no real guidance. It cannot authorize us simply to dismiss the one failure of replication as inductively inert. Or at least it cannot do so without extensive elaboration on just what conditions distinguish those cases in which the failure carries import and those in which it does not. Such elaborations are not at hand and not likely to be forthcoming.

A material analysis of cases like this, however, faces no such general problems. For approached materially, there is no universal principle implemented. There are only particular cases, each of which is ultimately to be analyzed individually.

Here is a celebrated example. Nineteenth century electrodynamics had given center stage to the ether, the medium that carries light and electric and magnetic fields. It surrounds the earth and the earth's motion through it creates currents that blow past us, much as a car's motion creates a headwind. Famously, the Michelson-Morley experiment of 1887 had failed to detect this ether wind. The experiment employed an extremely sensitive interferometer that split a light beam into two folded pathways and then recombined the beams. The results were read from changes in the interference patterns formed by the recombined beams as the interferometer was slowly rotated. While its importance in Einstein's pathway to special relativity remains debated (see Norton, 2014), the null result of the experiment is foundational for special relativity. Had this experiment detected an ether wind or ether drift, it would have detected the absolute motion of the earth, in contradiction with the principle of relativity.

On December 29, 1925, Dayton C. Miller (1926), addressed the American Physical Society in Kansas City. He recounted his own efforts to replicate the Michelson-Morley experiment, and reported the results of his latest efforts of 1925, when his apparatus was set up on Mount Wilson near the Observatory in California. He had found a positive result of 10 km/sec for the ether drift. It was less than the 30 km/sec or so that might otherwise be expected from the motion of the earth. Yet it was not a null result. This replication of the Michelson-Morley experiment had failed.

---

[30] This chapter was written prior to the publication of Volume 15 of the *Collected Papers of Albert Einstein* (Buchwald, 2018), whose documents relate to Einstein's appraisal of the Miller experiment. The editorial introduction (pp. lx–lxvii) provides considerable further details of Einstein's appraisal and those of his contemporaries.

This was not a failure to be taken lightly. Now, over a hundred years after the discovery of special relativity, we classify experiments challenging special relativity with circle squaring and perpetual motion machines. That dismissal was not so easy in 1926, especially in light of who Dayton C. Miller was. He was then the President of the American Physical Society; and he was employed by the Case School of Science, in Cleveland, the site of the famous Michelson-Morley experiment of 1887. His experiments had a venerable lineage. In 1902 to 1904, he had collaborated on ether drift experiments with Michelson's original collaborator, Edward Morley. They had reused parts of the apparatus of the original 1887 experiment. These parts included the iron trough that held the mercury in which the interferometer floated and the original circular wooden float. These parts, Miller (1933, p. 209), noted with some pride of ownership in his later review, "have been continued in use by the writer to the present time."

While there were other ether drift experiments from the time, Miller's used one of the longest folded pathways for light, which would give his one of the greatest sensitivities.[31] The experiments of 1926 built on the experience with Miller's earlier collaboration with Morley and successive refinements of the apparatus and experimental design through multiple experiments in a new series starting in 1921. It was feared, for example, that a basement in Cleveland, a mere 300 feet above the level of Lake Erie, may be too shielded from the ether current. For this reason, the entire apparatus was relocated to a mountainside next to the Mount Wilson Observatory, at an elevation of about six thousand feet. Miller's (1926, 1933) recounts the elaborate cautions undertaken to avoid and control all imaginable sources of error.

The report of Miller's positive result produced great interest in both scientific and popular circles. Miller was even awarded a $1000 prize by the American Association for the Advancement for Science for a related article. Einstein soon succumbed to popular pressure to respond. He wrote a short note for the popular press, published January 26, 1926, in the *Vossische Zeitung*, a well-known liberal newspaper in Berlin.[32] His remarks included:

> There is, however, in my opinion *practically no likelihood* that Mr. Miller is right. [Einstein's emphasis]. His results are irregular and point rather to an undiscovered source of error than to a systematic effect. Furthermore, Miller's results are in and of themselves hardly credible, because they assume a strong dependence of the velocity of light upon the height above sea level. Finally a

[31] For a compendium of other ether drift experiments from that time, see Miller, 1933, pp. 239-40 and Shankland et al., 1955, p. 168.

[32] This article was found by Klaus Hetschel (1992) from whose paper the translation of the text is drawn. See Hetschel (1992) for more details of the scientific and popular reaction to Miller's experiments.

> German physicist (Tomaschek) recently performed an electrical experiment also at a considerable height above the sea (the Trouton-Noble experiment), the result of which speaks against Miller's results insofar as it supports the absence of an "ether wind" at great altitudes.

From our perspective, what is notable about Einstein's response is that it invokes no matters of general inductive principle. Had Miller's claims somehow contravened an identifiable, universal inductive principle, it would have been easy for Einstein merely to point that out, much as one might identify a deductive fallacy. Rather, Einstein proceeds precisely as one would expect from the material theory. He gets the sharpest image of the inductive import of Miller's work by looking most narrowly at it.

Einstein's critique draws on facts in classes A and B above. For example, he complained that Miller's results are "irregular." Einstein did not elaborate, but, presumably, his concerns are similar to those expressed by Hans Thirring later in a June 1926 communication to *Nature*. In explaining his complete disagreement with Miller's interpretation of the experimental results, Thirring (1926) noted several irregularities within Miller's data. Since the ether wind will come from one direction in space, the direction detected by the interferometer should rotate through all points of the compass in the course of a day, as the daily rotation of the earth rotates the apparatus once per day in space. Yet, Thirring (p. 82) found:

> …an effect pointing towards the north-west quadrant of the compass in about ninety-five per cent. of all observations. This fact seems to be fatal to the assumption of an ether drift of constant direction towards a certain point of the heavens…

The facts at issue here are those in class A, which specify the conditions under which the process of interest manifests an experimental outcome. Under the supposition of an ether theory, the process of interest, the earth's motion through the ether, would manifest as an ether wind of a definite direction in space. That was not found, so that these background facts could not license the inference from the experimental outcome to the ether current.

Einstein then conjectured an "an undiscovered source of error." He did not specify what this source might be. However Einstein was quite direct in his private notes to correspondents. He wrote to his friend and confidant, Michele Besso, on December 25, 1926:[33] "I think that the Miller experiments rest on an error in temperature. I have not taken them seriously for a minute." He pressed this concern in a subsequent correspondence with Miller later in 1926, with Miller

[33] As quoted in Holton (1969, pp. 185-86).

dismissing it by describing the elaborate corrections put it place to control temperature effects.[34] Einstein's doubts may have had a firmer foundation than the brevity of his *Vossische Zeitung* remarks suggest, for he had long taken a keen interest in Miller's experiment. During Einstein's 1921 visit to the US, he had taken the trouble to visit Miller and, on Miller's report, had spent over an hour and a half discussing the ether drift experiments.[35] Einstein's suspicions were affirmed when Shankland et al. (1955) later performed a painstaking re-analysis of Miller's results, finding that positive results were associated with temperature variations in apparatus.

This second set of inferences drew on facts in class B. Einstein and Shankland and his colleagues had a sense of the processes that could produce a confounding result and, as Shankland and his colleagues affirmed, the pattern of results, in conjunction with these facts supported the conclusion of the thermal original of Miller's results.

7. Intercessionary Prayer: Successful Replication with No Inductive Import

The converse case is also possible: the successful replication of experiments, yet those successes are nonetheless regarded as inductively inert. Once again no formal account of reproducibility of experiment can accommodate this unless it specifies the conditions under which successful replication does and does not have inductive import. Approached materially, each case is treated individually and we face no insurmountable problems of general principle.

In intercessionary prayer, one entreats a deity or supernatural power to intervene in mundane affairs. The entreaty is most commonly for well-being and health and especially the speedy recovery of the sick. In the nineteenth century, two leading scientists, John Tyndall and Francis Galton, proposed that the efficacy of prayer could be assessed by objective tests of the type routinely employed in science.[36] If the sick do indeed fare better when they are prayed for, that good effect ought to be discernible through simple statistical analysis. They were skeptical. Galton had been collecting data for what amounted to a rather fragile retrospective study. He displayed a table of the mean lifetimes of males who survived past 30 years. Recalling that

[34] For details, see Hentschel (1992, p. 608). Einstein noted that temperature changes of as little as 1/10th degree in the air of the light path would be sufficient to generate results of the magnitude of Miller's.

[35] As affirmed by a letter of Miller's quoted in Holton (1969, p. 186).

[36] For a brief history, see Brush (1974).

sovereigns in every state are the subjects of public prayer, such as "Grant her in health long to live," he observed of his table (Galton, 1872):

> The sovereigns are literally the shortest lived of all who have the advantage of affluence. The prayer has therefore no efficacy, unless the very questionable hypothesis be raised, that the conditions of royal life may naturally be yet more fatal, and that their influence is partly, though incompletely, neutralized by the effects of public prayers.

The proposal, as one might expect, evoked derision from theological circles. James M'Cosh (1872, pp. 777-778) retorted

> We laugh at Rousseau's method of settling the question of the existence of God: he was to pray and then throw a stone at a tree, and decide in the affirmative or negative, according as it did or did not strike the object. The experiment projected by Professor Tyndall's friend is scarcely less irrational.

The mood had changed by the later twentieth century. Controlled studies of intecessionary prayer were conducted and continue to be conducted. Randolph Byrd (1988), for example, reported a prospective randomized double-blind trial of the effects of intercessionary prayer on the recovery of patients in a coronary care unit. He reported statistically significant improvements in recovery among those in the test group receiving prayer. Harris et al. (1999) performed a similar study on cardiac patients, again finding prayer to be associated with improvements in recovery. While not all experimental tests of intercessionary prayer have produced positive results, there are sufficient for meta-level surveys to be written. Astin et al. (2000) report the two studies above as the only ones producing positive results among the five surveyed. However, in the broader category of "distant healing," 57% of the studies reported positive results, which supported the final conclusion that the field "merits further study." A later review (Roberts et al., 2009)[37] was less optimistic. They found the results among the ten trials surveyed to be equivocal and recommended against further investigation.

Most of these reports are of little use in our efforts to understand what grounds inductive inference in relation to the reproducibility of experiment. Both surveys grapple awkwardly with the problem of some successful and some failed replication and, from them, arrive abruptly at a synoptic judgment. We are given little insight into how the analysts balanced the competing inductive import of the successes and failure to arise.

---

[37] Curiously, this report included positive results from the spoof Leibovici (2001) study. It also noted a later critic who pointed out their error, but nonetheless did not disavow the study, concluding: "The Leibovici 2001 was not in jest. It is a rather serious paper, intended as a challenge." (pp. 56-57).

There is a subgroup, however, whose members make clear that they regard successful replication of the intercessionary prayer experiments as inductively inert, for they do not believe that these studies have any inductive powers at all. Their analysis conforms with the material approach to reproducibility. For successful replication requires the facts in classes A and B above to be hospitable. This skeptical group does not find facts in class A supporting an inference from the experimental outcome to the supernatural intervention proposed. Hence replication adds nothing to an outcome that was already inductively inert.

Needless to say, this group includes atheist polemicists like Richard Dawkins. He remarks in his *God Delusion* (p.86) that "the very idea of doing such experiments is open to a generous measure of ridicule…" Theists also have traditionally been skeptical of such experiments. Their analyses can be more measured and thus prove more illuminating. The three authors of Chibnall et al. (2001), a Catholic, a Protestant and a Jew, describe how they set out to perform an experimental test of distant prayer. They "became convinced that the very idea of testing distant prayer scientifically was fundamentally unsound." In a telling, detailed analysis, they argue powerfully that, in effect, the requisite facts of the class A do not obtain: we have no good reason to expect the effect or process of interest (supernatural intervention) to be manifested in the experimental outcome (statistics of recovery rates among patients). They ask:

> If prayer is a metaphysical concept linked to a supernatural being or force, why would its efficacy vary according to parameters such as frequency, duration, type, or form? The very concept of prayer exists only in the context of human intercourse with the transcendent, not in nature. The epistemology that governs prayer (and all matters of faith) is separate from that which governs nature. Why, then, attempt to explicate it as if it were a controllable, natural phenomenon? … there is no reasonable theoretical construct to which to link prayer because of, we would argue, its very nature. No model guides our understanding of intercessory prayer as a treatment in the way we know that drug pharmacokinetics, type, dose, schedule, interactions, and treatment length are critical to an antibiotic as a treatment. In fact, we believe no scientific model can guide it.

Perhaps one of the most revealing of all the intercessionary prayer studies was reported in the December 2001 issue of the *British Medical Journal*. Leibovici (2001) collected all reports of patients who were detected with blood infections in a university hospital in Israel (Rabin Medical Center, Beilinson Campus) in 1990-1996. In 2000, he randomized the cases and arranged for prayer for a test group. The results show no improvement in mortality among the test group but a statistically significant shortening of both hospital stay and fever duration. The results were "retrospective" in the sense that these outcomes had already happened at the time the prayers

were administered. It was suggested that we should not assume that "God is limited by a linear time, as we are."[38]

This peculiar report produced the uproar one might expect. Letters to the editor in the April 27, 2002, issue of the *British Medical Journal* covered a wide range of complaints; and it was at times hard to tell if they were written in the same spirit at the original article. They included a defense of the laws of physics against breakage and protests over the ethics of experimenting on subjects whose consent could no longer be secured at the time of the experiment. The letters were concluded with an "Author's Reply," in which Leibovici admitted that the paper was really a spoof, but with a deeper purpose:[39]

> The purpose of the article was to ask the following question: Would you believe in a study that looks methodologically correct but tests something that is completely out of people's frame (or model) of the physical world—for example, retroactive intervention or badly distilled water for asthma?

Of three possible answers, Leibovici endorsed the third:

> To deny from the beginning that empirical methods can be applied to questions that are completely outside the scientific model of the world. Or in a more formal way, if the pre-trial probability is infinitesimally low, the results of the trial will not really change it, and the trial should not be performed. This, to my mind, turns the article into a non-study, although the details provided in the publication (randomization done only once, statement of a wish, analysis, etc) are correct.

Leibovici's assessment expresses in miniature why a formal account of controlled trials fails, where a material account succeeds. He notes that one can have a trial that meets all the requisite formal conditions. That was how he set up the study his article. Nonetheless the study has no inductive import. This situation is inexplicable if one adheres to a general, formal account of the reproducibility of experiment. The material approach faces no such problems. In it, the trial can

---

[38] I learned of this bizarre paper from a talk by John Worrall.

[39] Fact can be stranger that fiction. Over a year after the scam was admitted, Olshanky and Dossey (2003) published a note in the same journal that dismissed Leibovici's disavowal. In a narrative laden with pleas for open-minds, Einstein, Stephen Hawking, quantum mechanics, string theory and consciousness, they urged that we should subject these non-local, anomalous effects to serious study. This paper gives me great confidence in humanities' ability to turn every stone, for clearly no idea, no matter how absurd, lacks proponents.

have inductive import only if requisite background facts are hospitable. This, Leibovici asserts, is not the case here.

8. Conclusion

What is the inductive import of a successful or failed replication of an experiment? Mostly, successful replication is favorable to the result sought; and failures to replicate are unfavorable. However this is true "mostly" only. This broad similarity across many cases supports the illusion that there is some general inductive principle concerning reproducibility at work. However, efforts to specify that general principle precisely lead to mounting difficulties and failure.

Rather, as the material theory of induction requires, the question is ultimately answered differently in different cases according to the background facts obtaining. The more we narrow the types of experiments considered, the more precise the answers become. This is what we would expect from a material approach to induction, since with this narrowing the variability in background facts is reduced. What appeared as a universal principle is really only a resemblance among many distinct inductive inferences that vary in details according to their domains. No universal principle of inductive logic provides a warrant for these individual inferences. They are warranted by the particular facts prevailing in each domain.

The situation is quite like the case of enumerative induction. In many domains, we find the background facts warranting an inference from some individuals bearing a property to all individuals in that class bearing the property. I argued in an earlier chapter that these cases must be treated individually. The different background facts obtaining in each case will specify which individuals and properties in the domain are subject to the generalization. Nonetheless, as a looser gloss, the warranted inferences will look something like a progression from "Some As are B." to "All As are B." They can be glossed loosely as enumerative induction, but all efforts to find a single inductive schema implemented in all the cases fails. The unity is superficial.

References

Astin, John A. et al. (2000), "The Efficacy of 'Distant Healing': A Systematic Review of Randomized Trials," *Annals of Internal Medicine*, 132, pp. 903-910.

Atwood, Kimball C. (2004), "Bacteria, Ulcers, and Ostracism? H. Pylori and the Making of a Myth," *Skeptical Inquirer*, 28.6 (November/December 20004).

Brush, Stephen G. (1974), "The Prayer Test," American Scientist, **5**, pp. 561-563; **63** (1975), pp. 6-7.

Buchwald, Diana Kormos et al., eds. (2018) *The Collected Papers of Albert Einstein. Volume 15. The Berlin Years: Writing & Correspondence, June 1925-May 1927.* Princeton: Princeton University Press.

Byrd, Randolph C. (1988), "Positive Therapeutic Effects of Intercessory Prayer in a Coronary Care Unit Population," *Southern Medical Journal*, **81**, pp. 826-29.

Casadevall, Arturo and Fang, Ferric C. (2010) "Editorial: Reproducible Science" *Infection and Immunity*, **78**, pp. 4972–4975.

Chibnall, John T. et al. (2001) "Experiments on Distant Intercessory Prayer: God, Science, and the Lesson of Massah," *Archives of Internal Medicine*, **161**, pp. 2529-36.

Dawkins, Richard (2008), *The God Delusion.* Boston: Mariner.

DOE (2004), *Report of the Review of Low Energy Nuclear Reactions.* Downloaded as http://newenergytimes.com/v2/government/DOE2004/DOE-CF-Final-120104.pdf

ERAB (1989), Cold Fusion Research, November 1989: A Report of the Energy Research Advisory Board to the United States Department of Energy. Washington, DC. DOE/S-0073 DE90 005611

Galton, Francis (1872), "Statistical Inquiries into the Efficacy of Prayer" *Fortnightly Review*, **12**, pp. 125-35.

Harris, William S. et al. (1999), "A Randomized, Controlled Trial of the Effects of Remote, Intercessory Prayer on Outcomes in Patients Admitted to the Coronary Care Unit," *Archives of Internal Medicine*, **159**, pp. 2273-78.

Jasny, Barabara R., et al. (2011) "Again, and Again, and Again …," *Science*, **334**, p. 1225.

Hentschel, Klaus (1992), "Einstein's Attitude Towards Experiments: Testing Relativity Theory 1907-1927," *Studies in History and Philosophy of Science,* **23**, pp. 593-624.

Holton, Gerald (1969) "Einstein, Michelson, and the "Crucial" Experiment," *Isis*, **60**, pp. 132-197.

Leibovici, Leonard (2001), "Effects Of Remote, Retroactive Intercessory Prayer On Outcomes In Patients With Bloodstream Infection: Randomised Controlled Trial," *British Medical Journal*, **323**, No. 7327 (Dec. 22 - 29, 2001), pp. 1450-1451.

M'Cosh, James (1872), "On Prayer. III" *Contemporary Review*, **20**, pp. 777-782

Marshall, Barry and Warren, J. Robin (1983), "Unidentified curved bacilli on gastric epithelium in active chronic gastritis" *Lancet* 1(8336) (June 4), pp. 1273—1275.

Marshall, Barry and Warren, J. Robin (1984), "Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration," *Lancet* 1 (8390)pp. 1311–15

Miller, Dayton C. (1926) "Significance of the Ether-Drift Experiments of 1925 at Mount Wilson," *Science*, **63**, pp. 433-443.

Miller, Dayton C. (1933), "The Ether-Drift Experiment and the Determination of the Absolute Motion of the Earth," *Reviews of Modern Physics*, **5**, pp. 203-42.

Norton, John D. (2014) "Einstein's Special Theory of Relativity and the Problems in the Electrodynamics of Moving Bodies that Led him to it," pp. 72-102 in *Cambridge Companion to Einstein*, M. Janssen and C. Lehner, eds., Cambridge University Press.

Olshansky, Brian and Dossey, Larry (2003) "Retroactive Prayer: A Preposterous Hypothesis," *British Medical Journal*, **327**, pp. 1465-1468.

Peat, F. David (1989), *Cold Fusion: The Making of a Scientific Controversy.* Chicago: Contemporary Books.

Roberts, Leanne, et al. (2009), "Intercessory prayer for the alleviation of ill health (Review)," *The Cochrane Collaboration* in *The Cochrane Library*, 2009, Issue 3, John Wiley & Sons

Soddy, Frederick (1907) "Radioactivity," pp. 311-343 in *Annual Reports of the Progress in Chemistry. 1906."* Vol. III. London: Guerney and Jackson.

Shankland, R. S. et al. (1955), "New Analysis of the Interferometer Observations of Dayton C. Miller," *Reviews of Modern Physics*, **27**, pp. 167-78.

Sturms, Edmund (2007), *The Science of Low Energy Nuclear Reaction: A Comprehensive Compilation of Evidence and Explanations about Cold Fusion.* Singapore: World Scientific Publishing.

Thirring, Hans (1926), "Prof. Miller's Ether Drift Experiments," *Nature*, **118**(No. 2595), pp. 81-82.

Warren, J. Robin (2005) "Helicobacter—The Ease and Difficulty of a New Discovery," Nobel Lecture, December 8, 2005. http://nobelprize.org/nobel_prizes/medicine/laureates/2005/warren-lecture.pdf

Chapter 4

Analogy

1. Introduction

Reasoning by analogy is a venerable form of inductive inference and was recognized already millennia ago by Aristotle. Over these millennia it has been the subject of persistent analysis from the perspective of formal approaches to inductive inference. The goal has been to find the formal criteria that distinguish good from bad analogical inference. These efforts have met with mixed success, at best.

As we shall see below, the difficulties these efforts have faced are similar to those facing the formal explication of other sorts of inductive inference. If analogical reasoning is required to conform only to a simple formal schema, the restriction is too permissive. Inferences are authorized that clearly should not pass muster. This familiar problem is illustrated below in the case of a generic account of analogical inference, drawn from the older literature and described in Section 2. This is Joyce's (1936, p. 260) account, which I label "bare analogy" to reflect its simplicity. It has long been recognized that bare analogy authorizes too many inferences. This failure and its long-standing recognition is recounted in Section 3.

The natural response has been to develop more elaborate formal templates that are able to discriminate more finely since they capture more details of various test cases. Two elaborations are recounted here. Section 4 reviews Hesse's two-dimensional account, which is in turn derived from an analysis by Keynes. Section 5 reviews Bartha's articulation model. It was designed to remedy the shortcomings of Hesse's account by still further elaborations. Section 6 describes how these elaborations cannot escape the inevitable difficulty. Their embellished schema are never quite embellished enough. There is always some part of the analysis that must be handled intuitively without guidance from strict formal rules.

Section 7 turns to the material approach. According to it, the continuing expansion of the schema of the formal approach is inevitable since, according to the material approach, there is no single formal schema that can embrace all cases. As one tries to find schema that fit a growing body of cases better, the schema must introduce further distinctions and elaborations; and it must

do so without end. For there are always new instances to be accommodated and a need for schema that fit more closely.

That the material approach is a better way to understand analogies and analogical inference in science is indicated by a curious divergence between the philosophical literature and the scientific literature. The philosophical literature categorizes analogy as a *form of inference* to be analyzed using some version of the formal methods of logical theory. The scientific literature approaches analogies as *factual* matters to be explored empirically; or at least it does so for the important analogies that figure centrally in the sciences. For the scientists, there are many inferences associated with the analogy. But the analogy itself is a factual matter.

This gap between the philosopher and the scientist is hard to close if we approach inductive inference formally. If, however, we take a material approach to inductive inference, the gap closes automatically and the difficulties faced by the formal approach evaporate. We no longer need to display some universal schema that separates the good from the bad analogical inferences. Rather an analogical inference is good just in so far as there is a warranting fact to authorize it. Each warranting fact can be identified on a case by case basis without the need for it to conform with some elaborate template. That warranting fact is the factual analogy that scientists pursue empirically.

Sections 8, 9 and 10 illustrate the material approach with three cases of analogies in science: Galileo's discovery of mountains on the moon, the Reynolds analogy in fluid flow and the liquid drop model of the atomic nucleus. Section 11 presents general conclusions. An appendix provides technical details of the Reynolds analogy and a little of its history.

2. Bare Analogy

Argument by analogy has long been a standard in the inventory of topics of logic texts in the older tradition. It is specified formally in terms drawn ultimately from syllogistic logic. Joyce (1936, p. 260) states it as:

> S_1 is P.
>
> S_2 resembles S_1 in being M.
>
> [therefore] S_2 is P.

Mill (1904), Book III, Ch. XX, §2) gives an equivalent characterization in words:

> Two things resemble each other in one or more respects; a certain proposition is true of the one, therefore it is true of the other.

This simple argument form has proven quite fertile in the history of science. Galileo observed shadows on the moon that resembled the shadows of mountains on the Earth in both their shape

and motion. He pursued the resemblance to posit that there are mountains on the moon and to determine their height. Darwin's celebrated argument in the early chapters of *Origin of Species* exploits an analogy between domestic selection by breeders and the selective processes arising in nature. Gravity and electricity resemble one another in being forces that act between bodies or charges, diminishing in strength with distance. So in the eighteenth century, it was natural to expect that the analytic methods Newton developed for gravity might apply to electricity as well, even issuing in an inverse square law. Two more fertile analogies will be developed in more detail below: analogies among transport phenomena, notably the Reynolds analogy; and the analogy between an atomic nucleus and a liquid drop.

3. Its Failure

In spite of this record of success, descriptions of the argument form also routinely concede its inadequacy. Joyce (1936, p. 260) insists that the scheme he had just described has further hidden conditions.

> The value of the inference here depends altogether on the supposition that there is a causal connexion between M and P. If this be the case, the inference is legitimate. If they are not causally related, it is fallacious; for the mere fact that S_2 is M, would then give us no reason for supposing that was also P.

This amounts to a gentle concession that the formal scheme laid out is not able to separate the good from the bad analogical inferences. The addition, the fact of a causal connection, lies well outside the vocabulary of syllogistic logic in which this argument form is defined. That vocabulary is limited to individuals and properties and assertions about them using "not," "Some…" and "All…" For example: "Some As are not B."

Recalling classic examples of the failure of analogical reasoning shows us that this pessimistic appraisal is still too optimistic. The depressions Galileo found in the moon's surface resemble terrestrial seas. But there are no water filled seas on the moon's surface. Lines on the surface of Mars resemble terrestrial canals. But there are no such canals on Mars. Fish and whales resemble one another in many of their features. But one extends the resemblance at one's peril. Whales are mammals, not fish, and do not breathe through gills or lay eggs. In the eighteenth and early nineteenth century, heat was found to flow like a fluid from regions of higher heat density (that is, higher temperature) to those of lower heat density. Pursuit of the resemblance leads one to conclude that heat is a conserved substance. That heat is not conserved, but is convertible with work, was shown by the mid 19th century by Joule and others. Studies by Clausius, Maxwell and Boltzmann showed that heat is not even a substance in its own right. It is really a disorganized distribution of energy over the very many components of other substances.

In the nineteenth century, the wave character of light was reaffirmed. In this aspect it resembles the wave motions of sound or water waves. Since both these waves are carried by a medium, the air or water, analogical reasoning leads to the positing of a corresponding medium for light, the ether. The positing of this medium fared poorly after Einstein introduced relativity theory.

We see through these examples that formally correct analogical inferences frequently yield false conclusions. Joyce's added requirement of a causal connection is not sufficient to reveal the problems of the analogical failures just listed. Water on the moon or Mars would be causally connected with seas and canals. The property of surviving underwater is causally connected with having gills. The passage of heat from regions of higher to lower temperature is causally connected with the heat as a substance and temperature measuring its concentration. The wave motion of light is causally connected with the supposed medium that carries the waves.

We may want to discount these sorts of failure as a familiar artifact of inductive inference in general. When one infers inductively one always takes an inductive risk and inevitably, sometimes, we lose the gamble. The frequency with which we lose the gamble has supported a more pessimistic conclusion on analogical inference in science (Thouless, 1953, Ch. 12):

> Even the most successful analogies in the history of science break down at some point. Analogies are a valuable guide as to what facts we may expect, but are never final evidence as to what we shall discover. A guide whose reliability is certain to give out at some point must obviously be accepted with caution. We can never feel certain of a conclusion which rests only on analogy, and we must always look for more direct proof. Also we must examine all our methods of thought carefully, because thinking by analogy is much more extensive than many of us are inclined to suppose.

This unreliability of analogical reasoning is a fixture of handbooks of logic. They commonly have sections warning sagely of the fallacy of "false analogy." The reader is entertained with numerous examples of conclusions mistakenly supported by analogies too weak to carry their weight. The difficulty with these accounts is that the falsity of the analogy is only apparent to us because we have an independent understanding of the case at hand. There is little beyond banal truism to guide us away from false analogies when the difficulty was not already obvious at the outset.[40] Merely being warned to watch for weak analogies is unlikely to have helped an early

[40] Bartha (2010, p. 19) has performed the useful service of collecting a list of eight "commonsense guidelines." They include: "(CS1) The more similarities (between the two domains), the stronger the analogy." "(CS3) The greater the extent of our ignorance about the two domains, the weaker the analogy." "(CS5) Analogies involving causal relations are more plausible than those not involving causal relations."

nineteenth century scientist who infers that light waves must be carried by a medium, as are other waves; or that heat is a fluid since it resembles one in so many features. Until further empirically discovered facts are considered, these analogies seem quite strong.

After reviewing many examples of successful and unsuccessful analogies, Jevons (1879, p. 110) comes to a sober and cautious conclusion:

> There is no way in which we can really assure ourselves that we are arguing safely by analogy. The only rule that can be given is this, that the more closely two things resemble each other, the more likely it is that they are the same in other respects, especially in points closely connected with those observed . … In order to be clear about our conclusions, we ought in fact never to rest satisfied with mere analogy, but ought to try to discover the general laws governing the case.

Once one has been steeped in the literature on analogical reasoning and has sensed both its power and resistance to simple systematization, it is easy to feel that Jevons' rule is not such a bad outcome, in spite of its vagueness. It is a good tonic, therefore, to recall what successful rules look like in deductive logic. Modus ponens[41] is a valid inference, always. Affirming the consequent[42] is a deductive fallacy, always. We should take this as a warning. That our rules need to be protected by vagueness and ambiguity may be an alert that there is no precise rule to be found.

4. Two-Dimensional Analogy: Hesse's Account

If a formal account of analogical inference is to succeed, it will need to be significantly richer than the schema of bare analogy just discussed. There have been important efforts in this direction. The most successful and the most promising of these richer accounts is due to Mary Hesse and, more recently, Paul Bartha. First I will sketch the central, common idea of the account and then give a few more details of Hesse's and Bartha's versions.

An analogical inference passes from one system to another. Following Bartha (2010, p. 15), I will call the first the "source" and the second the "target." A successful analogical inference, in this richer account, does not just pass a property from the source to the target. It passes a relation over the properties of the source to the analogous relation over the properties of the target. The source may carry properties P and Q where P and Q stand in some causal, explanatory or other relationship. If the target carries a property P* analogous to P, the analogical inference authorizes us to carry over the relation to the target system, where we now

[41] If A then B; A; therefore B.

[42] If A then B; B; therefore A.

infer to a property Q* that stands in the same causal or explanatory relation to P*. This is the crucial enhancement. This relation makes it reasonable to expect that, if the target system carries P*, then it also carries Q*. I call this approach "two dimensional" because we have relations extending in two dimensions: there are relations contained within each of the source and the target systems; and there are the relations of similarity between the two systems.

Hesse's (1966) study of models and analogies in science provided a fertile tabular picture in which the two dimensions are arrayed vertically and horizontally. Hesse gave tables illustrating particular examples. Bartha (2010, p.15) extracts the general schema as

Source	Target	
P	P*	(positive analogy)
A	~A*	(negative
~B	B*	analogy)
Q	Q* (plausibly)	

The first column indicates the properties carried by the source and the second indicates those carried by the target. Properties corresponding under the analogy are indicated by adding an asterisk. The property P* in the target corresponds to P in the source.

The table indicates the introduction of the terms "positive analogy" and "negative analogy," drawn originally from Keynes (1921, Ch. XIX). The positive analogy is the properties on which the source and target agree; the negative analogy is the properties on which they disagree. Establishing possession of the as yet unaffirmed property Q* by the target is the goal of the analogical inference. The table does not indicate the relations obtaining in the two dimensions, the vertical and the horizontal. They are specified by Hesse (1966, p. 59) as: "…horizontal relations will be concerned with identity and difference… or in general with *similarity* and vertical relations will, in most cases, be *causal*."

The general sense is that the strength of support for this conclusion depends on a trade-off between the positive and negative analogy. The stronger the positive analogy, the more the conclusion is favored; but the stronger the negative analogy, the more the conclusion is disfavored. However I have found no simple formula or simple synoptic statement in Hesse's text for how this balance is to be effected. In discussing a particular example, however, Hesse (1966, pp. 58-59) gives guidelines for a particular case. These guidelines can be generalized by the simple expedient of suppressing the particulars of the case by ellipses and the substitution of symbols in order to simulate a general schema.[43] We recover:

[43] The unedited quote reads:

> The validity of such an argument will depend, first, on the extent of the positive analogy compared with the negative … and, second, on the relation between the new property and the properties already known to be parts of the positive or negative analogy, respectively. If we have reason to think that the properties in the positive analogy are causally related, in a favorable sense, to [Q], the argument will be strong. If, on the other hand, the properties of the [target] which are parts of the negative analogy tend causally to prevent [Q*] the argument will be weak or invalid.

If any general schema is intended by Hesse, it must be this or something close to it. There is considerably more discussion in Hesse's text, but I find it mostly inconclusive. The chapter "Logic of Analogy" (p.101) is devoted to the question of whether the presence of an analogy makes it reasonable to infer to some new property of the target system. "Reasonable" is given a weak reading only; it amounts only to the comparative notion of one hypothesis being more reasonable than another. Grounding for the comparative judgment is then sought in several then extant approaches to evidence, with largely negative results.

5. Bartha's Articulation Model

Bartha (2010, pp. 40-46) mounts a careful critical dissection of Hesse's theory that reveals its problems and short-comings. Bartha's own theory is the best-developed account of analogy I have found in the philosophical literature. It sets out to resolve the problems of Hesse's account and is based on an extension of Hesse's two-dimensional approach (p. 35). The goal of Bartha's (2010, Ch. 4) "articulation model" is to enable a judgment of the plausibility of an analogical inference. The term "plausibility" is itself employed as a term of art and is given two

---

> Under what circumstances can we argue from, for example, the presence of human beings on the earth to their presence on the moon? The validity of such an argument will depend, first, on the extent of the positive analogy compared with the negative (for example, it is stronger for Venus than for the moon, since Venus is more similar to the earth) and, second, on the relation between the new property and the properties already known to be parts of the positive or negative analogy, respectively. If we have reason to think that the proper- ties in the positive analogy are causally related, in a favorable sense, to the presence of humans on the earth, the argument will be strong. If, on the other hand, the properties of the moon which are parts of the negative analogy tend causally to prevent the presence of humans on the moon the argument will be weak or invalid.

explications, probabilistic and modal (pp. 15-19). The articulation model proceeds with the vertical and horizontal relations of Hesse's two-dimensional model. However the bulk of Bartha's analysis is devoted to the vertical relations and it greatly extend those of Hesse. Instead of merely requiring that the properties of the source system are causally related, Bartha allows four different sorts of vertical relations among these properties: they may be predictive, explanatory, functional or correlative. The first two come in deductive and inductive forms. The final two come only in inductive forms. Analogical inference carries these relations from the source to the target system.

The conditions for a successful analogical inference in the articulation model are elaborate. There are two general principles (p. 25): "prior association," which requires the existence of an explicit vertical relation that is to be extended by the analogical inference; and "potential for generalization," which requires "no compelling reason" that precludes extension of the prior associations to the target system. The formal specification of the model then approaches the judgment of plausibility in two stages. The first, "prima facie plausibility," requires the positive analogy to be relevant to the prior association and the absence of critically relevant factors in the negative analogy. The second stage assesses qualitative plausibility on the basis of three criteria: strength of prior association, extent of positive analogy and presence of multiple analogies.

The implementation of these two stages seems to differ according to the type of prior association. Further conditions become more clearly articulated, as the implementation proceeds. For example, in the discussion of "predictive/probabilistic analogies," (pp. 120-21) it turns out that there are five important determinants of plausibility: strength of prior association, extent of correspondence, the existence of multiple favorable analogs, only non-defeating completing analogs and only non-defeating counteracting causes. Perhaps the most difficult case is that of multiple analogies. Its treatment requires a formal extension of the original theory. A ranking relation "is superior than" is introduced as a partial ordering on the set of analogical arguments at issue. There is much more to explore in Bartha's richly elaborated account. However, sufficient of both accounts has been developed here for me to indicate why I think a different approach is preferable.

6. Problems of the Two-Dimensional Approach

Hesse's and especially Bartha's analyses of analogy are impressive for their care and detail; they significantly enrich the original formal notion of bare analogy. In particular, Bartha is surely correct to refocus attention on the vertical relations within each of the source and target, as opposed to the horizontal similarity relations between them. For these vertical relations matter

more—or so I shall argue below. If a formal analysis of analogical inference can succeed, this is likely the right direction. However, my view is that they are proceeding in the wrong direction. What was wrong with the bare notion of analogy was precisely that it tried to treat some inductive inferences formally rather than materially, and the resulting simple schema fitted poorly. The two-dimensional approach seeks to tighten the poor fit by including more formal apparatus. Yet each new formal notion brings with it further problems, compounding the difficulties and threatening an unending regress. Here are some of the problems.

Hesse strains to explicate in general terms even the simple notion of similarity that constitutes the horizontal relations. She does not favor "formal analogy," which refers to "the one-to-one correspondence between different interpretations of the same formal theory." (1966, p.68) The simple example is the analogy of a father to the state. The scientific example (whose details are not elaborated) is "the formal analogy between elliptic membranes and the acrobat's equilibrium, both of which are described by Mathieu's Equation." She continues: "This analogy is useless for prediction precisely because there is no similarity between corresponding terms." (p. 69) Instead she favors "material analogy," which are "pretheoretic analogies between observables." (1966, p.68) Examples of the favored material analogy are the analogy of the pitch of sound with the color of light; and the sphericity of the Earth with the sphericity of the Moon. These material analogies reduce the similarity relation to sameness of properties. The Earth and Moon are analogous in their sphericity since they carry the same property, sphericity.

While one can see the appeal of a limit to more secure material analogies, it is clearly overly restrictive. It disparages the fertile analogy between Newtonian gravity and Coulomb electrostatics, for example. It is a formal analogy in that it connects gravitational and electrostatic fields by virtue of their both satisfying the same field law (up to signs in the source term). There are other problems. A formal test that checks whether an analogy is material requires clear guides for when some term is "pretheoretic" and an "observable." There are many traps here. The analogy between pitch and color can be implemented only if we have numerical measures of pitch and color. Since these measures depend on a wave theory for both, are they still pretheoretic? Since they are inferred from measurements, are they observables?

Hesse's vertical relation is causality and it is similarly troubled. If we are to recover a serviceable, formal account of analogy, we must in turn have access to a serviceable formal account of causation. We must be able to confront each instance of a vertical relation with some formal criterion that tells us whether the relation is causal. Hesse's (1966, p. 87) summary is vague on just what is meant by causal relations. The vertical relations are "causal relations in some acceptable scientific sense…," which seems to suggest that discerning them is unproblematic. In this regard, Hesse seems unfazed by the plethora of candidate explications of causation that she lists. They include (1966, p.79) a Humean relative frequency account in which

causation is co-occurrence; a hypothetico-deductive account, in which causal relations are delivered by some higher level law; a modal account in which causes are necessities; and an ontological account in which causes are productive. We can hardly expect each of these theories to agree in every application. We have to know which is the right theory and then how to apply it in a formal account. The length of Hesse's list already indicates the difficulty in clarifying causation. Some half century after her list was formulated, we are now even farther from the goal of a general, formal account of causation. For my own quite pessimistic appraisal, see Norton (2003).

Bartha's articulation model is designed to free Hesse's more limited model from arbitrary restrictions. However, if an account this complicated is what is needed for a successful formal treatment of analogy, we surely have reason to wonder if a formal analysis is the right approach. Our starting point was a simple and familiar idea. If systems share some properties, they may share others. This idea has been used repeatedly to good effect in science. As we pass through the various efforts to explicate the idea formally, we have arrived at a multi-stage procedure with many specializing components and trade-offs. Yet we are still not in possession of a fully elaborated formal schema. The trading off of many of the competing factors still seems to be effected at crucial moments by our inspection and intuitive judgment.

Rather than examining these problems in detail, I want to indicate one aspect of the articulation model that is directly relevant to the decision between a formal and a material approach to analogical inference. The vertical relations of the articulation model are characterized in inferential terms. When P and Q are related predictively, P entails Q. When P and Q are related through explanation, Q entails P so that P explains Q. The third and fourth functional and correlative relations are explicated similarly as inductive relations. Hence, in this model, an analogical inference passes a property, expressed in inferential terms, from the source to the target. That means the analysis is meta-logical, since the analogical inferences are performed at a higher, that is a "meta," level on lower level structures that are in turn characterized by inferential properties. This meta-logical character places a rather extraordinary burden on the articulation model. If it is to give a formal schema for analogical inference, it must provide a schema for the analogical parts of the inference at the meta-level, and also schemas for each of the lower level forms of inductive inference. In short, it must solve the formal problems of analogical inference and also every other form of inference it invokes.

The simple solution to the last problem is to approach inductive inferences materially. Then to note that one may infer inductively from P to Q requires that there is some factual relation between P and Q that authorizes the inference. That is all it requires, for there is no supposition of a universal schema. This factual relation is what is passed by the analogical

inference, so that the amended model would lose its meta-logical character. Rather than pursuing this hybrid material/formal model, let us return to the full material approach.

7. Analogy in the Material Theory of Induction

In the material theory of induction, that there is an analogy between two systems is captured in a fact that may be merely conjectured or, better, may be explored empirically. This fact of an analogy then warrants an analogical inference, which is the passing of particular properties of the source system to the target. The precise character of the fact of analogy and precisely which properties may be passed will vary from case to case. There will be at best a loose similarity only between different analogical inferences in that, in all of them, we are authorized to pass properties from one system to another. There is no universal schema that can specify just which properties can be passed in which circumstances.

Hence, we should expect efforts to find a formal schema to face precisely the difficulties sketched in the last three sections. A simple formal schema will at best fit a range of cases imperfectly. Efforts to narrow the gap between the schema and the cases will require the proposal of more elaborate, more fragmented schemas. In an effort to capture a diversity not governed by any formal rule, they will need to divide the cases into a growing number of categories and subcategories. These refinements will allow a better fit, but the fit will never succeed perfectly for every case. We may eventually arrive at a formal system as elaborate as the articulation model, which, I have argued above, still falls short of the final, fully elaborated formal schema. No matter how complicated the successive proposals become, they will still never be adequate to all the cases. Gaps will remain.

There are two notions in the material analysis. The first is the fact of an analogy or just *fact of analogy*. This is a factual state of affairs that arises when two systems' properties are similar, with the exact mode of correspondence expressed as part of the fact. The fact is a local matter, differing from case to case. There is no universal, factual "principle of the uniformity of nature" that powers all inductive inference. Correspondingly, there is no universal, factual "principle of similarity" that powers analogical inference by asserting that things that share some properties must share others.[44] The fact of an analogy will require no general, abstract theory of

[44] If one is tempted by a principle of similarity, note that every failure of an analogy is a counterexample to a simple statement of the principle. The real principle would separate the projectable similarities from the unprojectable, even if only statistically. Formulating such a principle amounts to the same problem as finding a formal theory of analogy, which, this chapter urges, is an insoluble problem.

similarity. The fact of analogy will simply be some fact that embraces both systems. There is no general template to which the fact must conform.

The second notion is an *analogical inference warranted by a fact of analogy*. Such an inference may arise if we know the properties of one system but not the other. We may then conjecture that there is a fact of analogy obtaining between the first system and the other system. This conjectured fact then becomes the fact that warrants the inference. If the conjectured fact is unequivocal and held unconditionally, the analogical inference from one system to another may simply be deductive, with all the inductive risk associated with the acceptance of the fact of analogy. In other cases, there will be some uncertainty or vagueness in the conjectured fact of analogy. The analogy is asserted as likely; or even merely possible; or that the particular way the analogy is set up might not be correct, but something like it might be. These hesitations confer an inductive character onto the inference warranted by the fact of analogy.

The fact of analogy must be able to power this inference. Since there is no "principle of similarity," the fact of analogy cannot merely assert some similarity between the two systems. It must assert a factual property of the second system that is sufficient to warrant the inference to its properties. For this reason, it will turn out that similarities between the two systems will be less important in the material analysis. Rather the similarities will appear more as conveniences of expression. It is cumbersome to specify how dark shapes on the moon appear as shadows of tall prominences when they obstruct linearly propagating sunlight. It is easy for Galileo to say that they are just like the shadows of mountains on the earth.

The material approach reorients our focus in two ways:
First, the focus will be on the fact of an analogy, for that controls the inferential connection between source and target systems. Moreover, it will turn out in the examples below that the fact of an analogy will tend less to express a brute similarity between source and target systems. It will tend to express a property that they share. The fact of possession of this property by the target system will drive the resulting inference, rather than similarity with the source.
Second, there will be no general formal principles sought to assess the strength or weakness of an analogical inference. Its strength will be assessed by examining the fact of analogy that warrants the inference. If we doubt the strength of the inference and wish to refine our assessment, we would not seek to refine and elaborate formal principles. We would not, for example, seek better guides on just how, as a matter of general principle, we should balance the competition of positive and negative analogies. We would instead engage in empirical investigations of the fact of analogy. Knowing more, the material theory asserts, enables us to infer better.

In the following, I will show these ideas are implemented in three cases of analogy. The first is Galileo's discovery of the mountains of the moon. The second and third are analogies that

have played an important role in recent science: the Reynolds analogy for fluid flow and the liquid drop model of the atomic nucleus.

8. Galileo and the Mountains of the Moon

Galileo's (1610) *Siderius Nuncius*—the Starry Messenger—is an extraordinary document. In it Galileo reports the discoveries he made when he turned his telescope onto the heavens and observed systematically. One of the most striking was that the surface of the moon has mountains and valleys analogous to those on earth. The announcement of that discovery provided strong support to a major shift in scientific thinking then underway. The heavens, it was coming to be realized, were not the realm of immutable perfection but rather more like the earth. Here was observational evidence that the moon was not a perfect heavenly sphere after all, but resembled the craggy, pockmarked earth.

Galileo did not directly *see* mountains on the moon. Their presence was inferred from what he saw. He tracked the advancing division between light and dark on a waxing moon. His telescope showed him that its edge was not a smooth curve but an "uneven, rough and very wavy line." More important was the way it changed over time. As it slowly advanced, bright points of light would appear ahead of it. They would grow and soon join up with the advancing edge. Galileo finds the analogy to the illumination of mountains on earth irresistible. He exclaims (1610, p. 33):

> And on the earth, before the rising of the sun, are not the highest peaks of the mountains illuminated by the sun's rays while the plains below remain in shadow? Does not the light go on spreading while the larger central parts of these mountains are becoming illuminated? And when the sun has finally risen, does not the illumination of plains and hills finally become one?

Galileo is careful to exempt certain darker areas on the moon whose shading does not change with time. In so doing, he provides a positive summary of his conclusion concerning the shadows of the mountains (pp. 37-38):

> They [these other markings] cannot be attributed merely to irregularity of shape, wherein shadows move in consequence of varied illuminations from the sun, as indeed is the case with the other, smaller spots which occupy the brighter part of the moon and which change, grow, shrink, or disappear from one day to the next, as owing their origin only to shadows of prominences.

There is a similar analysis that identifies the depressions in the moon's surface that we now know as "seas."

Once secure in the conclusion that the moving dark shapes seen on the surface of the moon are shadows of mountains and valleys, Galileo proceeds to the most striking result (pp. 40-41). The higher the mountain, the farther ahead of the advancing edge that its peak will be illuminated. In some cases, Galileo noted, the peaks first appeared sometimes at more than one twentieth of the moon's diameter. This illumination, Galileo presumed, came from a ray of sunlight grazing tangent to the moon's surface at the edge of light and dark and then proceeding in a straight line to the mountain peak. These presumptions reduced computing the height of the mountain to the simple geometry of triangles. The result was a height of four miles for the largest mountain, which fares well against modern assessments.

Galileo's presentation of the analogy of earth and moon is compelling. However, from the perspective of the logic, the arguments are presented in fragments only and the reader is left to fill in the details. No doubt, once we undertake this exercise, different reconstructions of the logic will emerge. Here is one way of reconstructing it from the material perspective.

The controlling fact of the analogy is just this:

> The mode of creation of shadows on earth and of the moving dark patterns on the moon is the same: they are shadows formed by straight rays of sunlight.

This fact then authorizes two inferences. They both start with the same premise:

> There are points of light in the dark that grow (as Galileo described) ahead of the advancing bright edge on the moon.

They proceed to two conclusions:

> The bright points are high, opaque prominences.
>
> The higher ones are as much as 4 miles high.

Both inferences proceed deductively if the fact of analogy is as stated. The details are tedious, so I will not rehearse them. It is simply a matter of inferring from a shadow to the shape that produced it. For example, the moment a bright spot first appears ahead of the advancing edge, we know that the bright spot lies on a straight line, tangent to the moon at the edge of the advancing brightness. It now follows that that bright spot is elevated above the spherical surface of the moon, and by an amount recoverable by simple geometric analysis of triangles.

It is worth noting two features of the inferences. First, the analysis looks initially like a textbook instance of a simple analogical inference. Loosely, the earth and moon are similar in their shadows; the earth has mountains causing them; therefore the moon does too. However closer inspection shows that notions of analogy and similarity play a small role. The earth functions as a convenient surrogate for any uneven body turning under unidirectional light. Galileo could equally have called to mind a person's head turning in a room lit by a lantern. As the person's face turns to the light, the tip of the nose would first be lit, before the full nose. What matters is the posit that the moon and its changing pattern of light and dark result from

shadows cast. The inference is not driven as much by analogy as by subsumption of the moon into a larger class of illuminated bodies.

Second, the above reconstruction contains deductive arguments only. Galileo's full analysis is inductive. The inductive elements have been confined above by the selection of the fact of analogy. It comes after the inductive part of the analysis is complete. In that inductive part Galileo infers that the moving dark patches are shadows formed by straight rays of sunlight. The basis for his conclusion is the way the bright and dark spots change; they move just like shadows so cast. However that does not entail deductively that they are shadows. The inference is inductive, albeit a fairly safe one. To see that it is inductive, we need only recall that the inference requires also the assumption that no other mechanism could produce patterns of light and dark that move as Galileo observed.

Galileo is taking the inductive risk of accepting this assumption. Other mechanisms are possible and further analysis would be needed to rule them out conclusively. One lies close at hand. In the middle of his discussion, Galileo seeks to assure us that the mountains and valleys need not be visible to us in the periphery of the moon, where we are aligned to see them in elevation. As an addendum to his discussion, he conjectures that the moon's surface may be covered by a layer of "some substance denser than the rest of the ether." (p. 39) This substance may obstruct our view of the lunar terrain at the moon's periphery, for then our gaze passes through a great thickness of the material. Noting that the illuminated portion of the moon appears larger, Galileo conjectures that some interaction between this material and sunlight may be deflecting our gaze outward. Finally, puzzled that "the larger spots are nowhere seen to reach the very edge," Galileo conjectures: "Possibly they are invisible by being hidden under a thicker and more luminous mass of vapours." (p. 40)

The illumination of the mountain tops ahead of the advancing edge employs light that grazes the moon's surface and thus passes through a great thickness of this optically active, denser material. Galileo needs to assume that this optical activity is insufficient to create illuminated mountain tops as something like mirages, that is, by the bending of light towards us by this denser medium.

9. Reynolds Analogy

The explicit identification of analogies has played a prominent role in the analysis of transport phenomena. These are processes in fluids in which momentum, heat and matter are transported. Analogies within these processes form a standard chapter or more in the textbooks. The earliest of these analogies is the "Reynolds analogy," named for Osborne Reynolds, the nineteenth century scientist-engineer who founded the field. Its central idea is of an identity of

the processes that transport momentum and heat. Hot gases flowing through a tube, for example, are slowed by friction with the tube's walls. This friction transfers momentum out of the gas and that loss is manifested as a need to maintain a pressure difference to keep the gas flowing. The gas will also transfer heat to the cool tube walls. In the analogy, the two processes operate with identical mechanisms. For more discussion see the account of the Reynolds analogy below in Appendix A.

This textbook attention to an analogy is quite revealing, since it shows directly how a particular science conceives an analogy. It conceives the analogy as an empirical fact. The fact has two modes of expression, as reported in the Appendix. In the looser mode, the analogy asserts that the mechanisms or laws governing momentum and heat transfer are the same. That version is somewhat ambiguous. Since heat and momentum are different quantities with different properties, just how can the mechanisms or laws be the same? If we construe the sameness to mean that the rates of momentum and heat transfer are numerically proportional under the same conditions, then there is a simple quantitative expression of this sameness in terms of two dimensionless numbers. The friction factor f measures the frictional losses of momentum from a moving fluid; the Stanton number St measures the rate of heat transfer. This second, more precise form of the analogy sets these two numbers equal, up to a constant factor: $f/8 = St$.

In material terms, this literature is equating the analogy with the fact of analogy. The associated analogical inferences are present, but draw only subsidiary attention. The most common is to use the analogy to authorize an inference from momentum transfer to heat transfer. That is, if we know the friction factor f for some system, we use the fact of analogy to infer to the Stanton number St. From the Stanton number we can infer rates of heat transfer. This inference has great practical utility. Friction factors are relatively easy to determine from pressure differences. The corresponding rates of heat transfer are a great deal harder to measure.

This practical utility of the Reynolds analogy means that there is some premium on determining just how good an analogy it is. When faced with this problem, the literature does not seek guidance from a formal theory of analogical reasoning. It does not ask for rules on how to trade off the competition of positive and negative analogy. The refinement of the analogy is regarded as an empirical question to be settled by measurement. The equation to be tested is just that $f/8 = St$. It was evident already quite early that the analogy obtains only in special cases. It fails for fluids in laminar flow and even liquids in turbulent flow, but succeeds as a relatively poor approximation for gases in turbulent flow. Since the fundamental analysis of fluids in turbulent flow is difficult, the exploration of the analogy and refined versions that replace it, has remained largely a matter of brute-force empirical measurement.

10. Liquid Drop Model

In the 1930s, after the discovery of the neutron, the new field of the nuclear physics was born. The nucleus of an atom was recognized as consisting of many particles. The most common isotope of Uranium, U^{238}, consists of 92 protons and 146 neutrons, which sums to an overall nucleon number of 238. The nucleus was found to exhibit energetically excited states, somewhat like the excitations of an electron in a hydrogen atom. However the single particle methods that had worked so well for electrons in atoms were inapplicable to the many-body problem posed by the atomic nucleus. The many particles of the nucleus, all clustered together, seemed something like the many molecules clustered together in a liquid drop. The liquid drop model of the nucleus was based on this analogy. The hope was that the physics of drops might also coincide with at least some of the physics of nuclei.

The liquid drop model was already an established element of nuclear theory[45] in the 1930s, before it found its most popular application. In 1939, Lise Meitner and Otto Frisch (1939) sent their celebrated letter to Nature in which they proposed that certain processes were dividing the nuclei of Uranium atoms. This "fission" process, they suggested, could be understood using the liquid drop model. The capture of neutrons by Uranium nuclei may be sufficient stimulus to break them apart, much as an unstable liquid drop is easily broken up by a slight tap. The idea was taken up by Bohr and Wheeler (1939), who extended the liquid drop model quantitatively to encompass fission.

A liquid drop is held together because its constituent molecules are attracted to each other. For molecules deep within the drop, these attractions do not pull markedly in any direction and thus, by themselves, do not contribute to the drop's cohesion. Molecules near the surface, however, are attracted towards the drop by those deeper in the drop. A drop may have many shapes. However the larger the surface area, the more it has molecules on its surface seeking to move towards the center. Hence the drop naturally adopts a shape with the smallest surface area, a sphere, as its lowest energy state. This tendency to spherical form is commonly described as arising from a tension in the surface driving the drop to its smallest area. The general theory assigns a surface tension energy to the drop, proportional to its surface area. If the drop is energized by tapping, for example, it oscillates, somewhat like the ringing of a struck bell. As the drop deforms and increases its surface, it excites to higher energy states and absorbs the added

[45] For an early review before fission, see Bethe (1937, §53). For a history of the origin of the liquid drop model, see Stuewer (1992). I thank Michel Janssen for drawing Roger Stuewer's history of the liquid drop model to my attention.

energy of the tap. Finding the spectrum of these oscillations was an already solved problem of classical physics.

The motivation for the liquid drop model of the nucleus is the idea that the stability of the nucleus arises is some analogous way. It leads to the assumption that there is a nuclear energy corresponding to the surface tension energy of the drop. The volume of a nucleus is proportional to A, the number of nucleons. Volume varies with radius^3 and surface area with radius^2. Therefore the surface area of the nucleus varies as $A^{2/3}$ and the liquid drop model posits an energy proportional to $A^{2/3}$. Further, the various excitation modes of the nucleus were assumed to correspond to those of a liquid drop with suitably adjusted parameters.

Finally, the instability of a nucleus that results in fission could be analyzed quantitatively. The surface tension effect tends to hold the nucleus together. However a nucleus is positively charged, carrying Z protons. This positive charge creates forces that drive the nucleus apart. They come to be favored as the nucleus grows larger. The point at which they overcome surface tension is computed in the model by finding the state in which the slightest energizing of the nucleus will lead to such violent oscillations that the nucleus must split. The computation yields a stability condition expressed in terms of the number of protons Z and the number of nucleons A. The ratio Z^2/A must be less than 42.2 (as quoted by Blatt and Weisskopf, 1979, p. 304). U^{238} is perilously close to this figure, so it is expected to be prone to fissioning. For it, $Z^2/A = 92^2/238 = 35.5$. This result is traditionally quoted as a great success for the model.

The model appears, initially, to be a textbook case of analogical inference. In their synoptic treatise on nuclear physics, Blatt and Weisskopf (1979, p. 300) give what amounts to an inventory of the positive and negative analogies. “We find the following points of analogy,” they remark and then proceed to list three elements of the positive analogy. They can be stated in simplified form, writing “A” for both the number of molecules in the drop and the number of nucleons in the nucleus. They are:

- The volume of a liquid drop and the volume of a nucleus are both approximately proportional to A.
- The energy to evaporate a drop and the binding energy of nucleus are both approximately proportional to A, subject to correction by a surface tension term.
- Surface tension corrects this energy for a liquid drop by an additive term in $A^{2/3}$; and a semi-empirical formula for the binding energy of a nucleus also has an additive term in $A^{2/3}$.

However Blatt and Weisskopf harbor considerable doubts about the analogy. “It is probable that this analogy is only very superficial,” they continued. What followed amounted to an inventory of the negative analogy, consisting of:

- The stability of a liquid drop derives from repulsive forces that preclude molecules approaching one another by less than a minimum distance of the order of the size of electron orbits. There is no similar limit known for the approach of nucleons.
- Molecules in a drop follow the classical dynamics of localized particles. Nucleons have de Broglie wavelengths of the order of inter-nucleon distances and are governed by quantum mechanics.

At this point in the narrative, what is needed is some assessment of how good the analogy is. What Blatt and Weisskopf do *not* do is to try to assess the competition between these rivaling factors by appeal to general rules such as one might expect from a formal approach to analogical inference. Rather, they derive the formula for the energy levels of a nucleus as indicated by the model and subject it to experimental test. They decide (p. 305) that the energy levels fit observation poorly. " The liquid drop model of the nucleus is not very successful in describing the actual excited states. It gives too large level distances." However the liquid drop model works better when it comes to fission: "The limit for stability against fission is well reproduced…"

This mode of assessment is just what the material theory calls for. The fact of analogy, as revealed through this assessment, is a rather bare one. It is:

> The energy of a nucleus has an additive surface term proportional to $A^{2/3}$; and the nucleus' oscillatory modes match those of a liquid drop with corresponding parameters.

This fact is sufficient to support the inferences made under the model; and this fact is what Blatt and Weisskopf are actually putting to test.[46]

We also see once again that the similarity of the source and target is a subsidiary matter. What matters to the analogy is what is expressed in the fact of analogy, that the liquid drop and nucleus share just the properties listed.

11. Conclusion

The material theory of induction succeeds in simplifying our understanding of analogical reasoning in its acceptance of the dual role of facts: they may be premises in arguments and they may also serve as warrants of inference. Crucially, the material theory allows that displaying such facts provides the justification of the analogical inference and is the endpoint of analysis that seeks to determine the validity of the analogical inference. While there will be similarities

[46] For a more recent assessment with similar import see Wagemans (1991), pp. 8-12.

among different analogical inferences, there will be no overarching similarity of sufficient power to allow the separation of good and bad inductive inference by purely formal means.

A formal approach faces a more elaborate challenge. It can allow that a fact of analogy can somehow play a role in justifying an analogical inference. But this recognition cannot terminate a successful formal analysis. The validity of an analogical inference must be established ultimately by displaying conformity with a universal schema. The enduring difficulty is that, no matter how elaborate these schemas may have become, none proves to be final and complete. That this difficulty is irremediable is predicted by the material theory of induction.

Appendix A. Reynolds Analogy

The General Idea

In the dynamic analysis of systems with moving fluids, analogies have been found between three of the most important types of processes. The three processes are often called the "unit operations" of chemical engineering. They are momentum transfer, heat transfer and mass transfer.

The simplest and most studied instance is a fluid (gas or liquid) flowing in a cylindrical tube. As the fluid flows through the tube, its passage is resisted by friction with the wall of the tube. At the center of the tube, the fluid moves with the greatest velocity and therefore has the highest momentum density. At the wall of the tube, friction has brought the fluid to a halt, so that the outermost layer of fluid has no momentum. This frictional slowing is understood as a momentum transfer process. Momentum from the inner part of the fluid is passed to its outer surface, where it is lost to friction. This loss of momentum must be compensated by an applied force if the fluid is to continue flowing. That applied force creates a pressure difference along the length of the tube.

Heat transfer can arise in same system. The tubes might be in the boiler of a steam engine. Hot flue gases from the fire pass through a bundle of tubes that are surrounded by a jacket of boiling water. Heat is transferred from the gases in the tubes, through the tube walls into the water. To illustrate mass transfer, we might imagine that the gases contain some contaminant that is to be scrubbed out. The inner surface of the tube carries some absorbing solution. In the mass transfer operation, the contaminant passes from the gas into the solution.

The analogies arise from the idea that the mechanisms of three processes are the same, so that they are governed by the same quantitative laws. That simple idea has proven to be difficult to verify in all generality. The earliest proposals for implementing the analogies proved to work only under very restrictive conditions. In spite of the early failures, the idea of the analogy has

proven appealing and has generated a literature of many different and more complicated implementations.

Our interest is the underlying logic used with these analogies. We can recover that well enough merely by looking at the first and best known analogy, the "Reynolds analogy." It is the proposition that the mechanisms of momentum and heat transfer are the same. Texts differ in their statements. Here are a few selected at random:

> Reynolds postulated that the mechanism for transfer of momentum and heat are identical. (Foust et al., 1960, p. 173.)
>
> …Reynolds suggested that momentum and heat in a fluid are transferred in the same way. He concluded that in geometrically similar systems, a simple proportionality relation must exist between fluid friction and heat transfer. (Kakaç and Yener, 1995. p. 203)
>
> Reynolds proposed that the laws governing momentum and heat transfer were the same. (Glascow, p. 156)

These statements are strong and it is not entirely clear how they are grounded.

The Original Reynolds Analogy

Reynolds' authority is routinely invoked. Reynolds' (1874) original note certainly proposes some connection between the rate of heat transfer and internal motions in a fluid. However it is unclear that he intends a complete identity of both mechanism and law as asserted above. His analysis was not conducted in the context of the modern theory of transport phenomena and his paper does not give the quantitative expression now attached to the analogy. There are none of the dimensionless numbers we shall see shortly: no friction factors or Stanton numbers. Reynolds' own celebrated analysis of fluid flow in pipes was published nine years later. Reynolds' synopsis of his 1874 paper from his later collected papers reads:

> The heat carried off by a fluid from a surface proportional to the internal diffusion of the fluid near the surface—the two causes natural diffusion of the fluid at rest, and the mixing due to the eddies caused by visible motion—the combined effect expressed by: $H = At + B\rho vt$—this affording an explanation of results attained in Locomotive Boilers—experimental verification. (Reynolds, 1900, p.xi)[47]

For later reference, this equation is numbered by Reynolds as (I):

$$H = At + B\rho vt \qquad \text{(I)}$$

[47] H is the time rate of heat passed per unit surface area, t is the temperature difference between the surface and fluid, ρ is the fluid density, v its velocity and A and B are constants.

The closest Reynolds comes to a direct assertion of analogy arises in connection with a second equation he numbers as (II)

$$R = A'v + B'\rho v^2 \qquad \text{(II)}$$

where R is the resistance to fluid flow in the pipe. The essential quantitative assumption of Reynolds' (1874, p. 83) analysis was:

> And various considerations lead to the supposition that A and B in (I) are proportional to A' and B' in (II).

This analogy asserts less than the sameness of laws. In drawing an analogy between momentum and heat transfer, the temperature difference t is analogous to the velocity v, for each magnitude drives the transport. Heat transport arises from a temperature difference and momentum transport arises from the velocity differences of a velocity gradient. Under this association, the "B" term of equation (I) would need to be $B\rho t^2$, which it is not.

There is a way that the equations (I) and (II) can be fully analogous, however Reynolds does not make these details explicit, so we cannot know if he intended them. We assign dual roles to the velocity v. In its first role, it measures the fluid flow, so that the term ρv measures fluid flux. In its second, it drives momentum transport and is analogous to temperature difference t. We would then suppose that the first appearance of v in the v^2 term of (II) represents fluid flux and the second v in the v^2 term of (II) represents driving force. Then both B terms of (I) and (II) would have the analogous form "B (fluid flux) (driving force)."

Reynolds explicit use of a more limited analogy that determines how large the velocity v needs to be for the "B" term of (I) to dominate. The proportionality of the constants enabled Reynolds to argue that this arose under the same conditions for which the B' term of equation (II) dominates. That, he reported, arose for "very small" v.[48]

There was an immediate practical application of the dominance of the B term for commonly arising velocities. When it dominates, the temperature of the discharged fluid is

[48] Reacting to Reynolds' name, modern readers will likely find it irresistible to associate the conditions in which the A and B term dominate as regimes of laminar and turbulent flow respectively. However, Reynolds' (1883) celebrated study of laminar and turbulent flow was published nine years later and supports different relations. In it, Reynolds (p. 975) reports that previous experiments had adhered to laws $i = v^2$ or $i = Av + Bv^2$, where i is a pressure term. He now corrects these laws by setting the pressure term proportional v in the laminar regime and to $v^{1.723}$ in the turbulent regime.

independent of the velocity v.[49] That means that a locomotive boiler operating with larger flue velocities would be equally efficient at withdrawing heat from the flue gases no matter how great the flow of flue gases. This result, Reynolds could report with obvious satisfaction, explained an otherwise surprising fact about boilers: they are "as economical when working with a high blast as with a low." (p. 84)

The Modern Reynolds Analogy

If we cannot ground the analogy of the modern textbooks in Reynolds' original work, there are informal justifications available. There are two regimes for fluid flowing in tubes. If the flow is slow or the fluid very viscous, then the flow is laminar. It has the perfectly regular streamlines of slowly flowing honey. When the velocity is high, however, these perfect lines are disturbed by tumultuous eddies, readily visible if smoke or a tracing dye is injected into the fluid. These eddies mix the fluid quite efficiently. They will carry the fluid in bulk from the center of the tube to the wall and back. In this process, they transport both the momentum and heat of the fluid, making it plausible that the same law governs both transports. It is at best a weak grounding, for we proceed with little more than a caricature of turbulence and ignore a laminar region in the fluid that will be at the tube's inner surface. Since the plausibility argument can be given at best for turbulent flow, some authors limit assertion of the Reynolds analogy to turbulent flow. This is so with Kay and Nedderman (1974, pp. 143-44), who also sketch the above grounding.

Whether well-grounded or not, the goal is to generate a quantitative relation from the analogy. To do that, we need to find quantitative measures of both momentum and heat transfer. In the case of fluid flow in tubes, the pressure difference, ΔP, is an easy-to-measure manifestation of the momentum transfer process within the tube. This pressure difference will depend upon many other factors. It will change with many variables: the average speed of the fluid v, the length of the tube L, its diameter D, as well as the physical properties of the fluid, such as its density ρ. If we seek general regularities that govern this pressure difference, it turns out that we can accommodate many of these variables by considering a dimensionless number

[49] When the B term dominates, it follows from (I) that the heat withdrawn H is proportional to the mass flux ρv. So doubling the mass flux will just double the heat withdrawn, which entails that there is no change in the temperature reduction of each unit of mass of the flue gases passing through the boiler.

formed from these variables. The most commonly used is a dimensionless number, the friction factor[50]

$$f = (D/L)\Delta P/(\rho v^2/2)$$

We need not linger over why this particular combination of variables is introduced. It will be sufficient for our purposes to treat f as generalized measure of pressure difference and thus a measure of momentum transport.

In the case of heat transport, we are interested in the time rate q that heat is transmitted to the tube walls. The total rate will vary with the area of the walls A and the temperature difference ΔT between the tube wall and the fluid mean temperature that is driving the transport. To accommodate these variables, the goal of analysis is usually a heat transfer coefficient h, where

$$h = q/A\Delta T$$

Since the heat capacity at constant pressure C_P, mean velocity v and fluid density ρ can also affect the process, it turns out to be most convenient to embed the heat transfer coefficient in the dimensionless Stanton number

$$St = h/C_P\rho v$$

Once again, we need not linger now over just why the number is assembled as it is. We need only treat it as a generalized measure of the rate of heat transport.

Determining just how much momentum and just how much heat are transported out of the tube under nominated conditions is not easy. If the flow is turbulent, it cannot be done from first principles. However if we assume with the modern Reynolds analogy that the same process transports both, then, whatever the amounts may be, they are closely connected. A fairly straightforward if tedious computation (given in the next section) finds that connection to be expressed as an equality between the two dimensionless numbers that measure momentum transport and heat transport:

$$f/8 = St$$

This is the quantitative statement of the Reynolds analogy. It is an empirical claim that can be tested quite readily. It turns out only to hold under quite limited conditions. It holds as a relatively poor approximation for gases in turbulent flow, but fails for liquids and fluids in laminar flow. See Glasgow (2010, pp. 156-57) for a brief historical sketch of the discovery of limits to the analogy and of efforts to improve it.

---

[50] The definitions of these dimensionless numbers can sometimes differ in constant factors. I follow the conventions of Foust et al. (1960).

Generating the Quantitative Relation

Now we will linger over why the two numbers St and f are chosen to be as they are. Following Foust et al, 1960, p. 173, we may generate the quantitative expression for the Reynolds analogy, f/8=St, as follows. The context is a fluid of density ρ flowing with mean velocity v in a tube of diameter D and length L. Momentum, heat and, in general, other quantities are transferred to the tube wall. It is assumed that this transport of an unspecified quantity is governed by the relation

$$\text{flux at wall} = -K\,(\text{concentration at wall} - \text{mean concentration})$$

The "flux at wall," is the time rate of transport of the quantity per unit wall area. The two concentrations are just the amount per unit volume of the quantity, respectively at the wall and averaged over the whole fluid. The real point of the equation is to define the general transport coefficient K, whose values will vary with any change in the physical properties of the fluid and the geometry of the tube.

The supposition is that this equation holds for both heat and momentum transport, so that we can define a coefficient K_{heat} and K_{mom} for each. The quantitative expression of the Reynolds analogy arises from setting the two coefficients equal.

For the case of heat, the "flux at wall" is q/A, where q is the total rate of heat transport from the fluid and A is the tube wall area. The concentration of heat is just $\rho C_P T$. Hence we can write

$$q/A = -K_{heat}\,(\rho C_P T_{wall} - \rho C_P T_{mean}) = -K_{heat}\,\rho C_P\,(T_{wall} - T_{mean})$$

The second equality obtains if both ρ and C_P vary negligibly over the system. In general this assumption fails. However, for common engineering applications, it holds quite well in a wide range of cases. If we compare this last equation with the definition of the heat transfer coefficient h

$$q/A = h\Delta T = -h\,(T_{wall} - T_{mean})$$

we can then identify

$$K_{heat} = h/\rho C_P = (h/\rho C_P v)\,v = St\,v$$

where $St = h/\rho C_P v$ is the Stanton number defined earlier.

For the case of momentum, we proceed as follows. The total pressure force acting on the fluid is (pressure drop) x (flow area) = $\Delta P\,\pi D^2/4$. By Newton's second law, this quantity is the total rate of loss of momentum from the fluid. All this momentum is lost through transport to the tube wall, since friction from the wall surface is the only other force acting on the fluid. The tube wall has area $L\pi D$. Hence

$$\text{momentum flux at wall} = (\Delta P \pi D^2/4) / (L\pi D) = (\Delta P/4)(D/L)$$

The momentum concentration is (mass density) x velocity. At the wall, the velocity is zero, since the fluid is halted by friction with the tube wall. Thus the momentum density at the wall is zero. The mean momentum density is just ρv. Combining and substituting into the general transport equation used to define K we recover

$$(\Delta P/4)(D/L) = -K_{mom}\,(0 - \rho v)$$

so that

$$K_{mom} = (D/L)\,(\Delta P/4\rho v) = (1/8)\,v\,(D/L)\,\Delta P/(\rho v^2/2) = v\,f/8$$

where $f = (D/L)\Delta P/(\rho v^2/2)$ is the friction factor defined earlier.

We now express the Reynolds analogy in the setting equal of the two coefficients[51]

$$K_{heat} = St\,v = v\,f/8 = K_{mom}$$

from which we recover the quantitative expression for the Reynolds analogy

$$St = f/8$$

References

Bartha, Paul (2010) *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments.* Oxford: Oxford University Press.

Bethe, Hans (1937) "Nuclear Physics: B. Nuclear Dynamics, Theoretical," *Reviews of Modern Physics,* **9**, pp. 69-249.

Blatt, John M. and Weisskopf, Victor F. (1979) *Theoretical Nuclear Physics.* New York: Springer Verlag; repr. Mineola, NY: Dover, 1991.

Bohr, Niels and Wheeler, John (1939) "The Mechanism of Nuclear Fission," *Physical Review*, **56**, pp. 426-50.

[51] It may seem odd at first to set K_{heat} and K_{mom} equal, rather than merely proportional. For they pertain to the transport of different quantities, heat and momentum, where each is measured by its own system of units. Just this reason would preclude us setting *rates* of heat and momentum transport equal, for the equality would fracture if we merely changed our units for measuring heat from calories to BTU. However this will not affect the coefficients K. For they are insensitive to unit changes in the quantity transported. If we change the numerical value of the heat flux by moving our units from calories to BTU, there will be a corresponding change in the heat concentrations, so that the value of K_{heat} remains unchanged.

Foust, A. S, Wenzel, L. A, Clump, C. W, Maus, L. and Andersen, L. B. (1960) *Principles of Unit Operations.* New York Wiley.

Galilei, Galileo (1610), "The Starry Messenger" pp. 27-58 in *Discoveries and Opinions of Galileo.* Trans., Stillman Drake. Garden City, New York: Doubleday Anchor, 1957.

Glasgow, Larry A. (2010) *Transport Phenomena: An Introduction to Advanced Topics.* Hoboken, New Jersey: John Wiley and Sons.

Hesse, Mary B. (1966) *Models and Analogies in Science.* Notre Dame, IN: University of Notre Dame Press.

Jevons, W. Stanley (1879) *Logic.* New York: D. Appleton & Co.

Joyce, George Hayward (1936) *Principles of Logic.* 3rd ed. London: Longmans, Green & Co.

Kakaç, Sadik and Yener, Yaman (1995) *Convective Heat Transfer.* Boca Raton, Florida: CRC Press.

Kay, John Menzies and Nedderman, R. M. (1974) *An Introduction to Fluid Mechanics and Heat Transfer.* 3rd ed. Cambridge: Cambridge University Press.

Keynes, John M. (1921) *A Treatise on Probability.* London: Macmillan and Co.

Meitner, Lise and Frisch, Otto (1939), "Distinegration of Uranium by Neutrons: A New Type of Nuclear Reaction," *Nature*, **143**, pp. 239-40.

Mill, John Stuart (1904) A System of Logic, Ratiocinative and Inductive. 8th ed. New York: Harper & Bros.

Norton, John D. (2003) "Causation as Folk Science," *Philosophers' Imprint*, 3(4), <www.philosophersimprint.org/003004/>; reprinted in H. Price and R. Corry (eds), *Causation and the Constitution of Reality: Russell's Republic Revisited*, Oxford: Clarendon Press, pp. 11–44.

Reynolds, Osborne (1874) "On the Extent and Action of the Heating Surface of Steam Boilers," *Proceedings of the Literary and Philosophical Society of Manchester,* **14**, 1874-75; pp. 81-84 in Reynolds (1900).

Reynolds, Osborne (1883), "An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels". *Philosophical Transactions of the Royal Society*, **174**, pp. 935–982.

Reynolds, Osborne (1900) *Papers on Mechanical and Physical Subjects.* Vol. 1. Cambridge: Cambridge University Press.

Stuewer, Roger H. (1994), "The Origin of the Liquid-Drop Model and the Interpretation of Nuclear Fission," *Perspective on Science*, **2**, pp.76-129.

Thouless, Robert H. (1953) *Straight and Crooked Thinking.* London: Pan.

Wagemans, Cyriel, ed. (1991) *The Nuclear Fission Process.* Boca Raton, FL: CRC Press.

Rev: November 13, 2012; August 1, December 2, 2014.
Minor edits August 7, 2017; January 31, June 16, 2018.

Chapter 5

Epistemic Virtues and Epistemic Values: A Skeptical Critique[52]

1. Introduction

Epistemic virtues or epistemic values, we are told, play a major role in our assessments of the bearing of evidence in science. There is something quite right about this notion; and there is something quite wrong about it. My goal in the chapter is to explain each.

In brief, what is right about the notion of epistemic virtue or value is that criteria such as simplicity and explanatory power do indeed figure overtly in the evidential assessments made by scientists. Any comprehensive account of inductive inference must have a place for them. A material theory of induction accommodates them by treating them as surrogates for further background facts that ultimately do the epistemic work.

What is wrong about the notion is the words used to express it. The problem is simple enough to be described here fully at the outset. The terms "virtue" and "value" have prior meanings and rich connotations. These prior meanings conflict with the idea that the criteria they label are successful epistemically, that is, that they do guide us closer to the truth. Unless we erase these prior meanings and connotations, we tacitly adopt a form of skeptical relativism about inductive inference. More specifically, when we use the terms in this context, we place the criteria on the wrong side of two distinctions, that is, on the sides that indicate that the criteria do not serve their epistemic purpose.

The first is the distinction between means and ends. In the non-skeptical view, the goal of inductive inference in science is to get closer to the truth. The criteria that guide us are *means* to this end. Values and virtues are commonly understood to be things that we esteem in their own right. They are *ends*. If we now label the criteria as ends, we are tacitly discounting their function as means. We are, in effect, indicating that scientists prize simplicity for simplicity itself, thereby overlooking that simplicity is sought in the epistemic context as an intermediate that, we hope, brings us closer to the truth.

[52] I thank Heather Douglas for helpful discussion.

The second is the distinction between things that are *imposed* by outside conditions on a community versus those that the community freely *chooses* for itself. Criteria that guide a community toward true theories cannot be freely chosen, or at least they cannot be freely chosen if they are to be successful guides. The world constrains powerfully which criteria succeed. Choose ones that breach these constraints and we are guided poorly. We should not rely on the reading of entrails or astrological signs as guides to the truth, for our world is not such that they succeed. Choose ones that are better adapted to the world and we enjoy the success of modern science. If one holds that these criteria can be freely chosen, one forfeits the difficult and delicate adjustment of the criteria to the world that is needed if they are to be successful guides to truth. This is the view of a skeptic, much as skeptics about astrology believe that astrologers can freely choose the predictive significance of each star sign, for these skeptics hold that no choice leads to successful prediction.

Facts are traditionally distinguished from values. We may not know what the facts of the matter are in any particular case. However a factual claim is either true or false, but not both, and, if two people disagree on a factual claim, at least one of them is wrong. It is not so with values (and the values that underwrite our judgment of what is virtuous). The same two people can legitimately hold contradicting values. There is no corresponding necessity that at least one of them is wrong. They choose their values as they please and, while each may try to argue for the superiority of his or her values, ultimately they can legitimately agree to differ.

When we label criteria for theory choice "values" or "virtues," the choice of language connotes that they are freely chosen. That is incompatible with the idea that the criteria are successful, for whether a criterion is successful is not a matter of our choice. It is imposed by the world and the successful criteria are to be discovered or inferred from suitable analysis, not stipulated as conventional choices. In this second way, the terms "value" and "virtue" for the criteria conveys the skeptical view.

In the following, Section 2 reviews a standard and celebrated instance of the use of epistemic values: the supplanting of geocentric by heliocentric astronomy. Section 3 describes how the material theory of induction can accommodate inductive inferences in which epistemic values or virtues are invoked. These values, the theory asserts, are convenient surrogates for more complicated background facts that provide the warrant for the inferences. A common way that epistemic virtues enter into scientific discourse is reviewed in Section 4. Bare hypothetico-deductive confirmation is too permissive in how it accords evidential support. Demanding in addition the presence of certain epistemic virtues provides a way of restricting it permissive scope.

Section 5 turns to an early instance of the present confusion over values in philosophy of science. In 1953, Richard Rudner urged in the title of his paper "The scientist *qua* scientist makes

value judgments." I respond that Rudner's paper establishes no such thing. It shows only something few doubt: scientists *qua* members of society make ethical value judgments. Finally, Section 6 turns to Thomas Kuhn's highly influential 1973 Matchette Lecture, "Objectivity, Value Judgment, and Theory Choice." In it Kuhn laments that his critics have misread his writings as espousing a radical skepticism about the rational grounding of science. While he promises to set the record straight, Kuhn proceeds with an account that invites the same criticism. Kuhn's paper introduces characteristics used in theory choice and soon redescribes them misleadingly as values. The narrative focuses on such questions as how different scientists may weight certain values differently when they compete. Whether and how these values might be truth conducive in theory choice is never addressed.

2. The Classic Example: Ptolemy versus Copernicus

A celebrated example has long figured prominently in the epistemic virtues literature. In the sixteenth and early seventeenth century, astronomers were weighing competing celestial systems. Should they follow the traditional geocentric system of Ptolemy? In it, the sun, moon and planets orbit the earth in motions that were compounds of several circular motions. Or should they follow the heliocentric system of Copernicus? In it, the earth with its orbiting moon joined the planets and all orbit the sun.

Both were quite successful at the routine task of astronomy of predicting just when each celestial body would appear in each place in the sky. This purely descriptive task is known as "saving the appearances" or "saving the phenomena." Since the Copernican account was constructed from more recent observations, it fared a little better at this task. However it was well within the reach of Ptolemaic methods to equal it, if only some Ptolemaic astronomer was willing to put the effort into tinkering with the system.

The decision between the systems was made on other grounds. There were competing considerations. The difficulty with the Copernican hypothesis was making physical sense of an earth that was supposed to be careening through the heavens. The great appeal of the Copernican system was that it qualitatively simplified Ptolemy's system. The Copernican system acknowledged that our view of the planets came from a moving platform that takes one year to return to the same spot. This motion of our vantage point impresses the appearance of further circular motions on the planets and these impressed motions were coordinated since they derived from the same source, the earth's motion. Crudely put, the planets appear to wobble in synchrony because we view them from a wobbling platform. With this insight, Copernicans could then identify certain correlated motions within the Ptolemaic system as being just these projections. They could be separated from the true motions of the planets themselves. This gave the

Copernicans a powerful advantage, for they could explain the coordination among these motions as necessities of a heliocentric system, whereas Ptolemaic astronomers could only ascribe them to arbitrary coincidences within the geocentric system.

This greater simplicity and harmony of the Copernican system carried the day. That victory depended upon a strong appeal to aesthetic sensibilities. This is reflected in Copernicus' own dim assessment of the geocentric system in his Preface to *On the Revolutions of the Heavenly Spheres* (1543: 1992, p.4):

> [the geocentric astronomers'] experience was just like some one taking from various places hands, feet, a head, and other pieces, very well depicted, it may be, but not for the representation of a single person; since these fragments would not belong to one another at all, a monster rather than a man would be put together from them.

A little later in the Preface, Copernicus (1543; 1992, p.9) exults over the harmony of his system, listing how coincidences of the Ptolemaic system are explained by his system:[53]

> In this arrangement, therefore, we discover a marvelous symmetry of the universe, and an established harmonious linkage between the motion of the spheres and their size, such as can be found in no other way. For this permits a not inattentive student to perceive why the forward and backward arcs appear greater in Jupiter than in Saturn and smaller than in Mars, and on the other hand greater in Venus than in Mercury. This reversal in direction appears more frequently in Saturn than in Jupiter, and also more rarely in Mars and Venus than in Mercury. Moreover, when Saturn, Jupiter, and Mars rise at sunset, they are nearer to the earth than when they set in the evening or appear at a later hour. But Mars in particular, when it shines all night, seems to equal Jupiter in size, being distinguished only by its reddish color. Yet in the other configurations it is found barely among the stars of the second magnitude, being recognized by those who track it with assiduous observations. All these phenomena proceed from the same cause, which is in the earth's motion.

We are to be repulsed by the monstrous Ptolemaic system and captivated by the harmony of its heliocentric competitor. While each can in principle perform equally at saving the appearances, these aesthetic considerations, Copernicus urges, should lead us to favor his system.

In so far as we characterize these factors as aesthetic, they are vague. Beauty, as the popular saying goes, is in the eye of the beholder. There are many ways we might specify

---

[53] For an account of just how Copernicus understood the notions of harmony and symmetry in this context, see Goldstein and Hon (2008, Ch. 5).

precisely how the Copernican system implements this aesthetically described superiority. It may merely be that it is simpler in requiring fewer independent hypotheses. Or we may judge the heliocentric system more harmonious in locating the centers of more of the gross motions in the sun. Here we understand harmony as appealing to some sense of beauty, perhaps captured in some aesthetic of parsimony or perfection of balancing parts. Or we may judge the superiority to lie in the way the systems relate to the evidence supplied by the celestial appearances. While both systems save the appearances, the Copernican system does a better job of explaining them. It attributes certain coordinated motions in the appearances of all planetary motions to one single cause of our earth's motion. Or we may judge the Copernican system to be better tested by the appearances. For the apparent motion of one planet will enable us to fix our earth's motion. We must then find that motion reflected in the apparent motions of the other planets, on pain of refuting the Copernican hypothesis.

Whichever account of the superiority of the Copernican system we choose, that superiority is expressed in the same general way. The Copernican system in its relation to the evidence of the appearances is more virtuous than the Ptolemaic. The virtue is of a special type. It is epistemically potent. The system that possesses it is better supported by the evidence. These are epistemic virtues.

3. Epistemic Virtues and the Material Theory of Induction

How can the possession of these properties be epistemically potent and strengthen the inductive support provided by evidence? That is the principal question to be addressed here. Are we to seek some general principle of inductive logic that affirms greater inductive support to simpler hypotheses, more harmonious hypotheses, to hypotheses that explain better or enter into relations of overdetermination?

The material theory takes a quite different approach. It allows that we may find that some principle of this type that works more or less well in some domain. However any such principle will always have a limited scope and eventually we shall pass beyond its domain of applicability to examples where it fails. The material theory dictates that there can be one answer to the question of the origin of its epistemic power. Ultimately the properties that are commonly called epistemic virtues must be surrogates for background facts or assumptions. They provide the warrant for the inductive inference.

Below, I will try to locate a little more precisely how these properties can enter into accounts of inductive inference. In the next chapter, I will give a more detailed analysis of one of the best known, simplicity, and display how its inductive power—in so far as it has any—derives from its role as a surrogate for background facts or assumptions.

4. Repairing Hypothetico-Deductive Confirmation

There are no universal rules for inductive inference. Correspondingly, there are no universal rules governing the nature of the properties often called epistemic virtues and how they enter into evidential relations. However there is broad and common circumstance in which these properties play a reasonably well-defined role. They arise as a part of efforts to repair an excessively permissive account of inductive inference, hypothetico-deductive confirmation.

In this account of confirmation, we have cases of hypotheses, hypotheses with auxiliary assumptions or theories that deductively entail certain evidential statements. The truth of these evidential statements is then taken to support the hypotheses that entailed them. The idea is familiar and examples abound. Big bang cosmology predicts a 3º Kelvin cosmic background radiation as a residual of the inferno of the early universe. Starting with celebrated measurements by Penzias and Wilson in 1965, the existence of this thermal background radiation was confirmed and eventually judged to provide strong evidence for big bang cosmology.

This bare account has had a troubled history. Both geocentric and heliocentric systems can do a good job of entailing the observed motions of celestial objects. That means that they "save the phenomena." Whether this provided evidence of their respective systems' truth was the divisive issue of the sixteenth and early seventeenth century. In the most famous, known forgery in science, Copernicus' publisher Osiander introduced a spurious preface to Copernicus' celebrated work on 1543. He urged there that Copernicus' hypotheses "need not be true nor even probable." They "merely provide a reliable basis for computation," which means that they should be regarded as nothing more than a reliable means for astronomers to predict the observable motions of the celestial objects. He provided a quite powerful argument against reading truth into the hypotheses that saved the phenomena. It was an elementary fact of the astronomy of his time that two different constructions could yield the same observable motions. He gave the widely known example of the equivalence of an eccentric circle and a suitable designed deferent-epicycle. Successfully saving the phenomena would favor each equally, so that pragmatic considerations directed the choice of construction: "the astronomer will take as his first choice that hypothesis which is the easiest to grasp."

The difficulties for this bare notion of hypothetico-deductive confirmation remain today. They are seen most easily through the following consideration. Let A and B be two propositions whose truths are quite independent of one another. One gets a good approximation of this condition by drawing the propositions from widely different domains. Proposition A may be drawn from astronomy, for example, and B may be some proposition in economics. We can form the deductive inference:

Hypothesis: A and B

Evidence: A

The hypothesis deductively entails the evidence. Yet does the truth of the evidence now supply inductive support to the hypothesis, as the hypothetico-deductive scheme indicates? Clearly the hypothesis (A and B) gets no inductive support from the evidence A beyond the simple fact that A is itself a logical part of the hypothesis. For the hypothesis to gain inductive support from the truth of the evidence in the sense intended by the hypothetico-deductive scheme, the support of the evidence A for itself as a logical part of the hypothesis would somehow have to carry over to the other logical part of the hypothesis B. There is no connection that carries the support from A to B since the two are, by supposition, independent.

In cases of this type, the hypothetic-deductive scheme fails completely. What distinguishes the cases in which it does work? They will be distinguished by the obtaining of further conditions that provide a bridge between A and B over which the inductive support can pass. The display of properties often called epistemic virtues provides a way of showing these further conditions hold. Mere saving the phenomena, merely entailing true observations is not enough. It must be done the right way. We have already seen in the example of Copernican astronomy that there are many ways of characterizing just what that right way may be. We may look to special properties of the hypotheses themselves. They may be simple or harmonious. More realistically, we may compare properties. Of two hypotheses equally able to save the phenomena, we accord more support to the simpler or more harmonious. Alternatively, we may identify a property of the relation between the hypothesis and the evidence. An explanatory relation is highly prized and the better the evidence is explained, the more support accrues to the explainer.

Conversely, we may find some relations defective. Such is the case with ad hoc hypotheses. These are hypotheses specifically contrived to conform to the evidence. That fact means that they get no inductive support from it. In early 1916, Einstein had completed his general theory of relativity. In a review article on his new theory, Einstein accused his Newtonian predecessor of just such adhocery. Newtonian theory distinguishes inertial motions from non-inertial motions. Yet, Einstein complained, it provides no causal account of the difference. Rather the distinction is merely posited by declaring a preferred "Galilean space" in which an inertially moving body is at rest. He declared (1916, p.771) "The preferred Galilean space … is however a merely ad hoc cause and not an observable thing." Einstein promised that his new theory would provide the observable cause. The distribution of observable masses would fix which were the inertial, Galilean spaces.

5. Non-Epistemic Values

So far, I have identified how the properties often called epistemic values and virtues can have a role in inductive inference. That is the part that the epistemic values literature gets right. I now pass to the part it gets wrong. I have already outlined the troubles in the introductory paragraphs of this chapter: the terms "virtue" and "value" introduce a covert skepticism about inductive inference through the prior meanings and connotations of the terms. I will shortly identify the work of Thomas Kuhn as most responsible for the present misidentification of epistemic criteria. He was aided in establishing the misidentification, I believe, by an earlier tradition in philosophy of science. That earlier tradition challenged the standard notion that scientific practice was free of value judgments, where the values at issue were of the more familiar ethical type, such as the valuing of human life.

In 1953, Richard Rudner (1953), later to be editor-in-chief of the journal *Philosophy of Science*, published an article in the journal whose title and main claim was that "The scientist *qua* scientist makes value judgments." Rudner's argument maintained a distinction between the strength of evidential support for some hypothesis and the decision by some scientist to accept it. Values did not enter into the determination of the strength of support; they entered into the decision to accept a hypothesis. He wrote (p.2; Rudner's emphasis):

> …in accepting a hypothesis the scientist must make the decision that the evidence is *sufficiently* strong, or the probability is *sufficiently* high to warrant the acceptance of the hypothesis. Obviously our decision regarding the evidence and respecting how strong is "strong enough", is going to be a function of the *importance*, in the typical ethical sense, of making a mistake in accepting or rejecting hypothesis…*How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be.*

While Rudner did not explicitly delineate the sort of values he had in mind, his two examples clarify them. He suggests that our values may slow our acceptance of the hypothesis that a drug is free of a lethal contaminant, since an error will have fatal consequences. He wondered correspondingly how high a probability the scientists of the Manhattan project needed to accept that their detonation of the first atom bomb would not trigger a planet destroying chain reaction.

Rudner's analysis is at best exaggerated and at worst dependent on an equivocation.[54] There are two problems. First and less seriously, the type of ethical value judgments Rudner describes are rarely made in scientific practice. Overwhelmingly, the types of hypotheses assessed by scientists are mundane and bereft of dire apparent human import. Decisions over

[54] For a more extensive analysis of the weaknesses of Rudner's argument, see Levi (1960).

lethal contaminants in drugs and, especially, planet destroying chain reactions are uncommon. In the latter case especially, the hypothesis of a dire chain reaction could only arise after scientists over many preceding decades had accepted a plethora of hypotheses in quantum theory, chemistry and engineering, all remote from the ethically fraught hypothesis. In these and many other cases, the scientist could not anticipate the long-term consequences of their discoveries. When Niels Bohr accepted the hypotheses of his 1913 model of the atom that played a foundational role in the development of modern quantum theory, was he to anticipate that this theory would ground the development of nuclear fission bombs two decades later and, as a result, alter his threshold of acceptance?

To claim that the "scientist *qua* scientist" makes value judgments admits no gradation. It makes no distinction between the scientist, for whom the fraught ethical value judgments are rare and challenging moments, and the judge in a court of law whose day-to-day work requires ethical value judgments routinely. Rudner establishes at best that, on rarer occasions, scientists make ethical value judgments in their work.

The second problem is more serious. It pertains to this last conclusion. Rudner's argument equivocates on the term "scientist." There is a narrower and a broader sense. In the narrower sense, a scientist is merely someone who investigates nature, reporting what bearing the evidence has, with indifference to the broader human ramifications. Virtually all the work of scientists proceeds in this mode. They find strong support in the evidence for the hypothesis that electrons are spin half particles. In agreement with Rudner's supposition, ethical value judgments do not enter into the assessment of how strongly evidence supports the hypothesis. The hypothesis is accepted and it is done without any consideration of the human import of the hypothesis, for none is apparent. This work is the province of the scientist in this narrower sense. It requires no ethical value judgments to be made.

This narrowness continues when scientists evaluate hypotheses that may have human import, such as Rudner's examples that a particular preparation procedure produces a contaminant free drug or that an atom bomb will not trigger a planet destroying chain reaction. Mere acceptance of hypotheses like these will not have any human import. That import only arises when the acceptance of the hypothesis will lead to consequences in the larger society. The scientist may need to decide whether to endorse the procedure in a published manual of procedures for drug preparation. Or the scientist may need to advise the principals of the Manhattan Project on the dangers of their planned Alamogordo atom bomb test.

That is, the human import only arises when the scientist has ceased to act as a scientist in the narrower sense. The scientist is now acting in the broader sense of someone who practices science and monitors the import of his or her work within the wider human society. When operating in that broader sense, scientists should be aware of the human consequences of their

actions and they should moderate their actions accordingly. In this broader sense, scientists make ethical value judgments in many ways that pertain to their engagement with the larger society. Who do they hire to work in their lab? Who do they fire? Are the safety precautions and procedures in the lab adequate to protect the lab staff? Should they purchase cheap, possibly stolen materials? Should the discharge from their lab be allowed to contaminate a nearby stream?

That ethical quandaries arise for scientists is a direct result of this broader role taken by scientists. It is not specifically a result of their doing scientific work. It is a result of their doing something, whether science or not, that impinges on the broader society.

Hence, Rudner simply got it wrong. Scientists *qua* scientists do not make ethical value judgments. Scientists *qua* members of society make ethical value judgments.

6. Kuhn's Obfuscation

While Rudner may have equivocated on the term "scientist," he is not responsible for the conflation of epistemic criteria with values. That distinction belongs to Thomas Kuhn. His 1973 Matchette lecture, "Objectivity, Value Judgment, and Theory Choice," launched the present popularity of the broadened scope of values talk in philosophy of science.

The origins of the lecture lie in Kuhn's earlier, wildly successful *Structure of Scientific Revolutions.* That work brought us the notion that revolutions in science are akin to religious conversions and that they carry us between paradigms that are incommensurable, defying rational comparison. The attempts to compare paradigms rationally become circular since the means of rational evaluation, Kuhn (1970, p. 94) assures us, resides within one or other paradigm. As a result, we are assured that: "paradigm choice can never be unequivocally settled by logic and experiment alone." And: "As in political revolutions, so in paradigm choice—there is no standard higher than the assent of the relevant community."

These are strong claims sure to raise the hackles of anyone who sees science as aspiring to rationally grounded discoveries about the world. The world does not adopt some state merely because some community agrees it has it. Yet Kuhn has just declared communal assent to be a highest standard, which means it cannot be overruled by logic and experiment. Curiously Kuhn (1973, p. 321) professed to be dismayed by critics whom he quoted as accusing him of making theory choice "a matter of mob psychology." This last description is at worst merely a colorful overstatement of the view Kuhn expresses in *Structure* in the academically muted "no standard higher than the assent of the relevant community." Now Kuhn (1973, p.321) responds in the Matchette lecture that these assessments of his views "manifest total misunderstanding." He will set the record straight.

This is a reassuring start. His celebrated *Structure*, it now appears, did not state clearly what Kuhn really thought about theory choice. Since many of its skeptical assertions are unequivocal, we must assume that he did not mean literally what he said. Or perhaps he expressed his views in a misleading way that invited misinterpretation. We can now learn what he really meant. Perhaps he merely meant that communal assent follows when one paradigm is favored over another according to some epistemically sound criteria. The superiority consists in conformity to these rationally grounded criteria and not in communal assent. Rather, we are to suppose that the relevant community is sufficiently astute to recognize this conformity, so that we outsiders can use their assent as a reliable indicator of the superior choice. This is one possible clarification that would escape the charge of relativism. We are ready for some such clarification.

What follows in the Matchette lecture, however, is simply a repeat of whatever had gone wrong in *Structure*. Someone expecting an account of the rational basis of theory choice in science finds nothing of the sort.

The account begins with a non-exhaustive list of the characteristics that (p.322) "provide *the* [Kuhn's emphasis] shared basis for theory choice." This list comprises accuracy, consistency, scope, simplicity and fruitfulness. It is not hard to give an account of how these characteristics can be rationally grounded. Consistency is the easiest. If a theory fails to have it, that is, if it is an inconsistent theory, then at least some of its propositions must be false. If we seek truth, we should avoid inconsistency. Accuracy refers to agreement between the consequences of the theory and the results of observation and experiment. This characteristic shows conformity of theory with known facts and, clearly, the better that conformity the better the facts weigh in the theory's favor. The remaining characteristics are not so straightforward but certainly within the compass of further analysis. The following chapter, for example, treats simplicity from the perspective of a material theory of induction.

Simple affirmations of this type would preclude the impending misunderstanding that Kuhn holds these characteristics to be merely the preferences of some particular group of people at some time in history. Yet no such affirmations are made. Rather the text moves as rapidly as possible to the question of how scientists weigh the force of the different criteria when they conflict and, eventually, how they change over time. We are only five pages into the article when we find a lengthy treatment of how individual differences between scientists have to be considered to explain why different scientists may weigh the criteria differently. It is a curious development in an account that is supposed to display that Kuhn does not hold the skeptical relativism of which he is accused. A simple answer to the accusation is to explain why he thinks these criteria are good guides to the truth after all. Instead, the focus has become the flaws and weaknesses of the criteria and how other, extra-rational factors are needed.

A charitable reader may still imagine that Kuhn's criteria form the basis of a rationally grounded system and not merely the predilections of some group. Perhaps Kuhn finds the point too obvious to mention. This charity is hard to maintain. Some ten pages into the article (p. 330), what was initially labeled "characteristics" or "criteria" are relabeled "values" or "norms." The transformation is not benign. It is justified by the specious claim that (p. 331):

> the criteria of choice with which I [Kuhn] began function not as rules, which determine choice, but values, which influence it.

The term criteria is quite properly used to label factors that merely influence a choice and it is a better term to use in so far as it is free of the tendentious connotations of "value." As I noted earlier, the connotations of the terms "value" and "norm" contradict the idea that Kuhn's criteria are the basis of rationally grounded account of theory choice

First there is the distinction between means and ends. A characteristic can readily be understood as an intermediate in a fuller account. Selecting for it can be a means to some other end, such as getting closer to the truth. The term value has different connotations. It is normally understood to designate something valued in its own right. It is itself an end or a goal. When theory choice is described as a "value judgment," as in the paper's title, the normal understanding is that the choice is made to realize the values in question as an end. In effect we are told that we seek consistent or simple theories because we value consistency and simplicity as an end and not because we regard them as an intermediate means for getting closer to the truth.

Second, there is the distinction between that which is imposed on the community by the outside world and that which is chosen freely by the community. In calling the criteria "values," Kuhn indicates that they are of the second type. For we are not forced by reason alone to the values we adopt. We choose them and enjoy considerable freedom in the selection. In foreign policy, we may debate whether to go to war. The debate becomes irresolvable when we find that the debating parties proceed from different values. The pacifists, we find, base their view on the value judgment that killing is wrong in all circumstances. The militarists make a value judgment that some killing is warranted to preserve sovereignty. We can debate the facts and expect agreement from reasonable people. But if we differ in our values, we have arrived at an irresolvable end. Analogously, if our theories are guided by values that we can choose freely, then debates over the correct choice is correspondingly futile. There is no right choice. That contradicts the idea that these criteria are epistemically successful, for the successful criteria must be discovered. They cannot be chosen as communal conventions.

When Kuhn relabels the characteristics or criteria as "values" and, less commonly, "norms," he is inviting the simple confusion that he thinks they are free choices of a community and sought as worthy ends in themselves, much as these communities may choose to value life, liberty, self-sacrifice, compassion or the ability to play football well. Kuhn's examples of values

do nothing to dispel the confusion. He writes: "Improving the quality of life is a value…" (p. 330) "Or again, freedom of speech is a value, but so is preservation of life and property."[55] (p. 330) Each of these is readily identifiable as an end that may be freely chosen. A dour religious sect that values deprivation and suffering will not value the improvement of quality of life; and they may also be indifferent to the preservation of both life and property. For, they believe, better awaits in the world to come. A highly authoritarian society may not value freedom of speech, since they regard it as contravening their values of obedience and respect of authority. Lest Kuhn leave any doubt that others may choose different values, the paragraph ends with the remark that most of us have "…an acute consciousness that there are societies with other values and that these value differences result in other ways of life, other decisions about what may and may not be done." (p.331)

This freedom of choice in our values conforms with the troublesome assertion of *Structure*: "As in political revolutions, so in paradigm choice—there is no standard higher than the assent of the relevant community." The language mirrors Rudner's tendentious claim of the role of social values in theory acceptance. In both cases, "values" determine what the scientists accept. The supposed misunderstanding of *Structure* is invited again.

Is it too much to ask for Kuhn to answer the accusation of skeptical relativism by giving the rational grounding of his criteria? Kuhn suggests that it is too great a demand. He dismisses (p. 326) the search for an "algorithm" that could determine theory choice as "a not quite attainable ideal." What of the extraordinary power of science to (p. 332)[56]

> …repeatedly produc[e] powerful new techniques for prediction and control. To that question, I have no answer at all, but that is only another way of saying that I make no claim to have solved the problem of induction.

Here Kuhn seeks to escape the burden of displaying an account of the rationality of theory choice that shows how its choices guide us closer to the truth. He seeks to escape it with a dilemma: either give an algorithm for theory choice and solve the problem of induction or give nothing at all. It is a false dilemma. There is a path between its horns. One can seek to show that the criteria he lists are conducive to the truth at least in some cases. That can be done without providing an algorithm for theory choice or without solving the problem of induction. The criterion of consistency, as I remarked above, is easy. Lose consistency and we know we are farther from the

---

[55] Kuhn offers these examples as part of a discussion of how values may conflict.

[56] Also Kuhn writes: "Though the experience of scientists provides no philosophical justification for the values they deploy (such justification would solve the problem of induction), those values are in part learned from that experience and they evolve with it." (p. 335)

truth. I will argue in the next chapter that the criterion of simplicity is really a surrogate for specific facts that do guide us well, locally.

In sum, what are we to make of Kuhn's Matchette lecture? As far as I can see, it is a muddled paper by a well-meaning but confused scholar. He has failed to see that his notion of rationality is a radically skeptical one and he is irked and baffled when his critics point it out to him. If that were all that is at issue, the paper would be best left and forgotten. However that is not all. This paper has now become the *locus classicus* of a new literature on values in science. It has legitimated the mislabeling of the criteria for theory choice as "epistemic values" or "epistemic virtues." There is a banal fact that scientists use criteria in choosing among theories. That banality is now redescribed in language whose connotations convey a skepticism about the rational grounding of those choices. There is no treatment of how these criteria might bring us closer to the truth or even mention that they do so. Rather theories are chosen because scientists value consistency and simplicity, much as a religious body might value piety.

The effect is to group together use of these benign criteria with Rudner's tendentious claim that scientists *qua* scientists make ethical value judgments. The blurring of the distinction between criteria and values invites a fallacy. Scientists do use criteria like consistency and simplicity in theory choice. Misdescribe this banality as scientists choosing theories by value judgments and we appear to have established that values permeate the apparently value-neutral content of scientific theories. This rhetorical subterfuge, whether intentional or inadvertent, is avoided simply by reverting to the neutral language of "criterion" and "characteristic."

The confusions and conflations of Kuhn's Matchette lecture have exercised considerable influence. They were endorsed by an otherwise astute President of the Philosophy of Science Association, Ernan McMullin, in his Presidential Address.[57] McMullin urges that the epistemic criteria at issue really are values. He bases this extraordinary conclusion on the same fragile grounds as Kuhn: they influence but do not determine the outcome. McMullin (1982, p. 16) writes:

> …these criteria clearly operate as *values* do, so that the theory choice is basically a matter of value-judgment. Kuhn puts it this way:
>
> > The criteria of [theory] choice function not as rules, which determine choice, but as values which influence. Two men deeply committed to the same values may nevertheless, in particular situations, make different choices, as in fact they do. [reference]

[57] McMullin was President, 1981-82. Kuhn was himself later President, 1989-90.

While criteria may be like rules in so far as they influence but do not determine outcomes, they are unlike values in the two senses I have outlined: criteria are means, not ends; criteria are imposed, not chosen. Their relabeling as values is unsupportable.

McMullin persists, designating "epistemic values" as those "which are presumed to promote the truth-like character of science." (p. 18) They are distinguished from non-epistemic values, such as the political, moral, social and religious. It is encouraging that the distinction appears to be maintained cleanly. Epistemic values are distinguished as those whose choice is "likely to improve the *epistemic* status of the theory, that is, the conformity between theory and world." (p.19 McMullin's emphasis) That is a serviceable standard for delineating epistemic criteria, however they are named. Yet such caution is ineffective when the distinction is ridden over, rough shod, by such claims as "Value judgment permeates the work of science as a whole." (p. 18)[58]

Finally, one may object that the issue is merely one of connotation and that, after Kuhn, the terms "value" and "virtue" have lost the connotations that trouble me. If that is so, why not revert to the neutral language? That reversion would, no doubt, be resisted. For it would break the connection between the provocative but mistaken role for values in science supposed by Rudner and the benign but common role for criteria like consistency in theory choice. The literature in "science and values" would become the heterogeneous literature in "science, criteria for theory choice and ethical values" and Kuhn's paper, "Objectivity, Value Judgment, and Theory Choice," would become "Objectivity, Criteria-Based Judgment and Theory Choice." The misleading connotations do persist and do matter.

References

Copernicus, Nicholas. (1543; 1992) *On the Revolutions.* Trans. Edward Rosen. Baltimore: The Johns Hopkins University Press, 1992.

Einstein, Albert (1916), "Die Grundlage der allgemeinen Relativitätstheorie," *Annalen der Physik*, 49, pp.769-822.

Kuhn, Thomas (1970) *The Structure of Scientific Revolutions.* 2nd ed. Chicago: University of Chicago Press.

[58] For completeness, I note that the concluding Section 6 of McMullin's paper is devoted to arguing that the objectivity of science can be defended from the relativism suggested by its permeation with values. Would the section have been needed had he merely retained the neutral term "epistemic criteria" thereby erasing the epistemically deleterious connotations of the term "value"?

Kuhn, Thomas (1973), "Objectivity, Value Judgment, and Theory Choice," in *The Essential Tension: Selected Studies in Scientific Tradition and Change.* Chicago and London: University of Chicago Press, 1977.

Goldstein, Bernard R. and Hon, Giora (2008), *From Summetria to Symmetry: The Making of a Revolutionary Scientific Concept*. Springer.

Levi, Isaac (196) "Must the Scientist Make Value Judgments," *Journal of Philosophy*, **57**, pp. 345-57.

McMullin, Ernan (1982) "Values in Science," *PSA Proceedings of the Biennial Meeting of the Philosophy of Science Association,* Vol. Two: Symposia and Invited Papers, pp. 3-28,

Rudner, Richard (1953) "The Scientist *Qua* Scientist Makes Value Judgments," *Philosophy of Science*, **20**, pp. 1-6.

Rev: November 13, 2012
December 14, 2014
Minor typos corrected, June17, 2018

Chapter 6

Simplicity as a Surrogate[59]

1. Introduction

The idea is found almost everywhere, from the most prosaic to the most abstruse settings. Choosing the simpler option speeds you to the truth. In ordinary life, when the lights go out, we choose the simpler hypothesis that the electrical power has failed. We discard the more complicated hypothesis that all the bulbs malfunctioned at the same moment and, worse, that each malfunctioned for a different reason. In cosmology, we choose the simpler hypothesis that the same physical laws obtain here as in distant places and epochs, even though we cannot rule out that they may differ in parts quite remote from us.

Do these judgments implement a universal principle of inductive inference that says:

> If two hypotheses are each adequate to the phenomena, the simpler is more likely true.

My goal in this chapter is to deny the efficacy of any such universal principle of inductive inference. For the material theory of induction entails that no such rules are efficacious. To explain the popularity of appeals to simplicity, I will urge that good invocations of simplicity are really veiled references to background facts or assumptions whose content functions to license the relevant inductive inference. The apparently singular appeal to simplicity actually masks an appeal to such a diversity of context dependent facts that no univocal meaning can be attached to it.

This is the sense in which simplicity is a surrogate. In so far as it is epistemically efficacious, the short and snappy invocation of simplicity is really a surrogate for background facts or assumptions. These background facts do the real epistemic work and, commonly, are much harder to capture in a comparably short slogan. There will be cases in which these backgrounds resemble one another so that a common idea of simplicity appears to be invoked.

---

[59] My thanks to Fellows in the Center for Philosophy of Science, Fall 2012, for discussion of an earlier draft of this chapter.

However the extent of these cases will always be limited. As we move farther afield, we will encounter cases in which the backgrounds differ sufficiently for the similarity to fail. In general, there is no well-specifiable, universally applicable, epistemically efficacious principle of simplicity in inductive inference.

This analysis is a deflationary analysis of simplicity that runs counter to the celebration of simplicity in the scientific literature. It does have a small pedigree in the philosophical literature. It is the view of simplicity long defended by Elliott Sober. His Sober (1988) uses emphasized text to summarize his view as:

> Whenever a scientist appeals to parsimony to justify the conclusion that one hypothesis is more reasonable than another in the light of observational data, substantive assumptions about the world must be involved. In practice, parsimony cannot be "purely methodological." (p.40)

and then more compactly:

> Appeal to simplicity is a surrogate for stating an empirical background theory. (p.64)

The following section provides a brief illustration of how apparently epistemically efficacious invocations of simplicity are really indirect appeals to background facts. Section 3 brackets off two cases of lesser interest in which simplicity offers only pragmatic gains. They are the cases in which simplicity is urged as an efficient search heuristic and in which simplicity is demanded merely to give a compact summary of past experiences.

The two sections that follow develop and deflate two primary senses of simplicity. The first principle, discussed in Section 4, expresses simplicity in a count of entities or causes. The classic statement is Ockham's razor: "Entities must not be multiplied beyond necessity." It fails as a principle of parsimony, I will argue, since there is no clear way to count the numbers of things to be minimized. The principle is reinterpreted as truism of evidence, that one should not infer to more entities than the evidence warrants, where this evidential warrant is understood materially. The second principle of parsimony, discussed in Section 5., requires us to infer to hypotheses whose description is simple. This principle fails as an independent principle since modes of description vary. These variations greatly affect the descriptive simplicity of hypotheses. This form of the principle can only guide us if we fix the mode of description and the guidance will be good only if that mode is properly adapted to the prevailing facts.

Section 6 will examine in more detail the most popular illustration in the philosophical literature of the use of simplicity, curve fitting. The invocation of simplicity in standard curve fitting, I argue, is a surrogate for specific background facts. They are: the obtaining of a particular model of how error in data confounds some true curve; that the parameterization used is suitably adapted to the backgrounds facts; and that, in the strongest cases of this adaptation,

the hierarchy of functional forms fitted corresponds to background assumptions on the presence, likelihood and strength of certain processes. Ascending the hierarchy is not authorized by some abstract principle that tells us to proceed from the simpler to the more complex. Rather it is a successive accommodation of the curves fitted to the most likely or strongest processes and then to those less so. The concluding two sections 7 and 8 illustrate this last adaptation of the hierarchy in the examples of fitting orbits to observed positions in astronomy and the harmonic analysis of tides.

2. How it Works: The Birds

Just how can simplicity serve as a surrogate for background facts? Here is an easy illustration. Imagine that you are walking on the beach over sand washed smooth by the ocean waves. As you walk over a clear expanse of smooth sand, you notice a track left by a bird (Figure 1).

Figure 1. Bird Tracks

The prints are clear and unmistakable. You can just see how the bird waddled over the sand—left, right, left, right—leaving the prints. But why assume that it was just one bird? Perhaps the left foot prints were made by a one-legged bird that hopped awkwardly over the sand. Then a second one-legged bird, this time having only the right leg, pursued it, leaving the right footprints in just the right place to simulate the waddle of a single two-legged bird. Or perhaps there was a large flock of one-legged birds, each of which touched down on the sand just once, all perfectly coordinated to leave the track.

Each hypothesis explains the track. However we do not take the various, one-legged bird hypotheses seriously. How might we defend this judgment? The one bird hypothesis is by far the simplest. In ordinary discourse, merely declaring that might be a sufficient defense. If our methodology is at issue, then merely declaring that it is the simplest might not be enough to secure it. If we need more, we can turn to the great Isaac Newton. At the start of Book III of his magisterial *Principia* he asserted four "Rules of Reasoning in Philosophy," which would guide the subsequent analysis. The first two rules are (Newton, 1726, p.398):

> Rule I
>
> *We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.*

> To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.
>
> Rule II
>
> *Therefore to the same natural effects we must, as far as possible, assign the same causes.*
>
> As to respiration in a man and in a beast; the descent of stones in *Europe* and in *America*; the light of our culinary fire and of the sun; the reflection of light in the earth, and in the planets.

These two rules remain the clearest and firmest pronouncement of a methodological principle of parsimony in science.

Applied to the birds, Rule I tells us immediately that we should use the one bird hypothesis, for it is a truth that there are two-legged birds and their behavior is sufficient to explain the tracks. We do not need the many bird hypothesis, so it should not be admitted. In so doing, we conform with Rule II by assigning the same cause, a single bird, to the many footprints.

So far, all is well. Simplicity has provided the principled justification for our intuitive judgment. That will not last, however. We now proceed farther down the beach and come to a place where the smoothed sand is criss-crossed by very many tracks, in evident confusion (Figure 2).

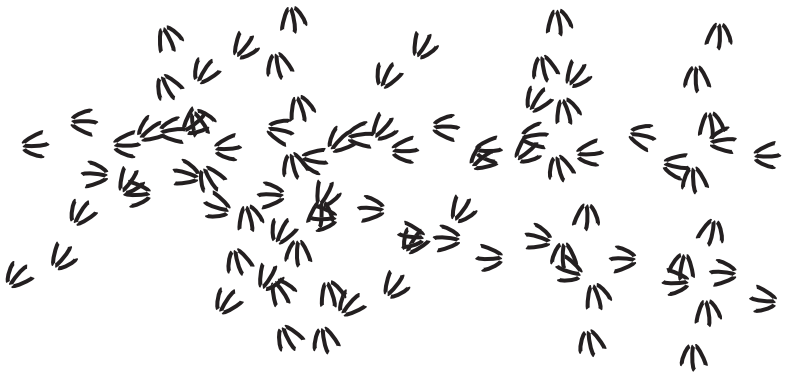


Figure 2. More Bird Tracks

We would normally posit that many birds alighted on the sand, each leaving just one track. However there is another hypothesis: the tracks were left by just one, very busy bird. It waddled over the sand; flew to another spot; waddled some more; and so on, until the final set of tracks was formed.

A mechanical application of Newton's rules leads us directly to the one busy bird hypothesis. We are, as before, assigning the same cause, one bird, to the same effects, the one set of tracks. Few of us would accept this outcome. We would be satisfied with one bird hypothesis

for the single track but expect a good analysis to return a many bird hypothesis for this case of many tracks. We would surely reason something like this. In the case of the single track, we rule out the many, one-legged bird hypothesis because we know that one-legged birds are rare and, if they were on the beach, it is unlikely that they would follow each other around in just the way needed to produce a single track. For the case of many tracks, we know that it is possible for one bird to be very busy and produce the multiple tracks. However we rarely if ever see such a lone, busy bird, whereas flocks of birds producing tracks like this are quite common.

These further reflections show that our initial analysis was not merely based on the invocation of simplicity. We chose the one bird hypothesis for the single track on the basis of our relatively extensive knowledge of birds. It is a shared knowledge, so we generally feel no need to explain in tedious detail why we rule out other possible, but unlikely hypotheses: two one-legged birds hopping, many one-legged birds alighting once, a mutant four-legged bird, and so on. We can dismiss all these far-fetched notions with a breezy wave towards the simplest hypothesis.

In short, what we offer as a conclusion governed by some general principle of parsimony is really a conclusion dictated by our knowledge of background facts. We use an appeal to simplicity as a convenient way of circumventing the need to explain in detail these background facts, whose details can become quite convoluted. My claim here is that all epistemically efficacious invocations of simplicity have this character.

3. Pragmatic Justifications of Simplicity

Let us return to the standard view that treats a preference for simplicity as a methodological principle of universal character. What justifies it? What precisely does the principle assert? My interest in simplicity is restricted to the case in which simplicity functions epistemically as a marker of truth; we are to choose the simpler hypothesis or theory because, we are assured, it is more likely to be true. I will argue below that a principle of this form has no precise content and no good justification. However, before we proceed, we need to dispense with two distracting special cases that lie outside my concerns. In them, simplicity is sought merely for pragmatic reasons.

3.1 Simplicity for Economy of Search

In seeking to understand some new phenomenon, scientists commonly deal with many hypotheses or theories. How should they go about searching among them and testing them? A common recommendation is that they should start with the simplest hypothesis or theory. The simplest are easiest to deal with and, if they are incorrect, likely to be refuted by new evidence sooner than a more complicated one.

In the 1920s, it was found that distant galaxies recede from us with a speed that increases with distance. In 1929, Hubble proposed that the speed of recession was linearly proportional to the distance. In principle, he could have fitted a complicated, tenth order polynomial function to his data. The linear dependency, however, was easier to deal with formally. If it is the wrong relation, new data would be likely to show the error much faster than with a more complicated function. A tenth order polynomial is able to contort itself to fit a larger range of data, so that considerably more data may be needed to refute it.

This sort of circumstance is common. One of the simplest hypotheses concerning an ailment is that it is caused by a specific pathogen. Famously, in the mid nineteenth century, John Snow was able to localize the cause of a cholera outbreak in London to drinking tainted water, drawn from a public water pump at Broad Street. More recently, the cause of AIDS—acquired immune deficiency syndrome—has been identified in the HIV virus. Once the simple hypothesis was pursued, it was readily affirmed. Were definite pathogens not responsible, the simple hypothesis could likely have been ruled out fairly quickly by the appearance of cases in which no exposure to the conjectured pathogen was possible. Matters are quite different with ailments such as cancer. Multiple factors can make a cancer more likely, including carcinogenic chemicals, ionizing radiation, certain viruses and even specific genes. Dealing with this multiplicity of causal factors and discerning which are operating when, is considerably more difficult.

These simple observations have been incorporated into analyses of scientific discovery. Karl Popper (1968, Ch. VII) urged that science proceeds through a continuing cycle of the conjecture of new hypotheses and their refutation. He identified the simpler hypotheses with the more falsifiable. It follows that the cycle advances faster if the scientists investigate more falsifiable hypotheses, that is, simpler hypotheses. A mathematically more sophisticated analysis of the role of simplicity in heuristic search has been provided by Kelly (2007). In the context of a formal learning theoretic analysis of the evidence-guided search for hypotheses, he shows that favoring simpler hypotheses is a more efficient way of getting to the truth.

How are these considerations relevant to our present concerns? One might ground the favoring of simplicity in searching in two related suppositions: that nature is ontically simple or that nature is descriptively simple in our languages. In both these cases, further discussion must be deferred to later sections of this chapter, where I argue that both suppositions are epistemically efficacious only in so far as they make indirect appeals to background assumption.

However these assumptions are not needed to ground the heuristic recommendation. It is still good advice to investigate the simplest first in a world that is indifferent to the simplicity of hypotheses. Whether the world is more likely to give us a linear function or a tenth order polynomial, the linear function will still be dealt with more easily and more quickly. Whether

ailments are more likely to be caused by a single pathogen or by many factors, we still proceed most expeditiously by checking the single pathogen hypothesis first.

In short, simplicity can remain a good heuristic in hypothesis searching without any need for nature to be governed by a general principle of simplicity or parsimony.

3.2 Simplicity as Mere Economy of Expression

Ernst Mach famously held the view the scientific laws were merely compact summaries of our experience. He said in an 1882 address to the Imperial Academy of Sciences in Vienna "The goal which it [the intellect] has set itself is the *simplest* and *most economical* abstract expression of facts." (Mach, 1898, p. 207) The idea can be put crudely as follows. Galileo asserts that the distance fallen by a body varies with the square of time of fall. In Mach's view, all that Galileo is allowed to assert is that each pair of distances and times we have measured for falling bodies conforms to this relation.

In so far as this is all that is asserted, then the role of simplicity is merely that of convenience. One seeks the least troublesome way of summarizing the facts at hand. In more modern terms, the exercise is essentially one of data compression. We could report all the numerical data pertaining to the fall of many bodies; or we could report merely that these data all conform to Galileo's relation without loss of anything that matters.

This may seem an extreme view that is, nowadays, well out of the philosophical mainstream. However much engineering practice conforms to it. That is because engineering commonly deals with systems dependent on many variables and the systems are sufficiently complicated that a fundamental analysis is precluded. To deal with this problem, the behavior of the system is measured experimentally under widely varying circumstances and the collected data reduced to as compact a form as possible.

One of the best-known examples is the treatment of fluid flow in pipes. Even this simple problem involves many variables: the fluid's speed, density and viscosity; the pressure drop in the paper; and the pipe's diameter and surface roughness. Once the flow becomes turbulent, this empirical approach is the only tractable one. Moody (1944) presented a now famous chart summarizing the outcomes of many experiments. See Figure 3.

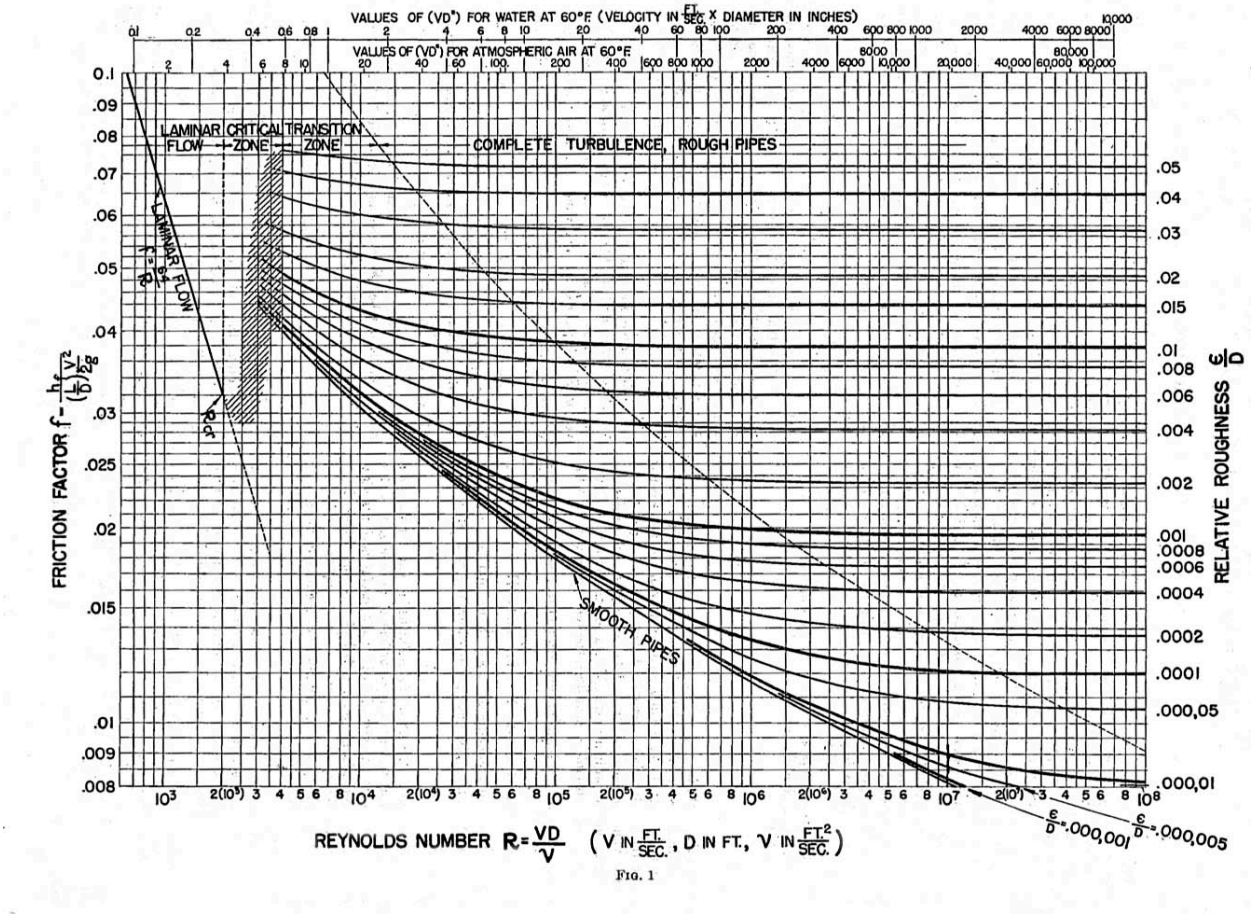
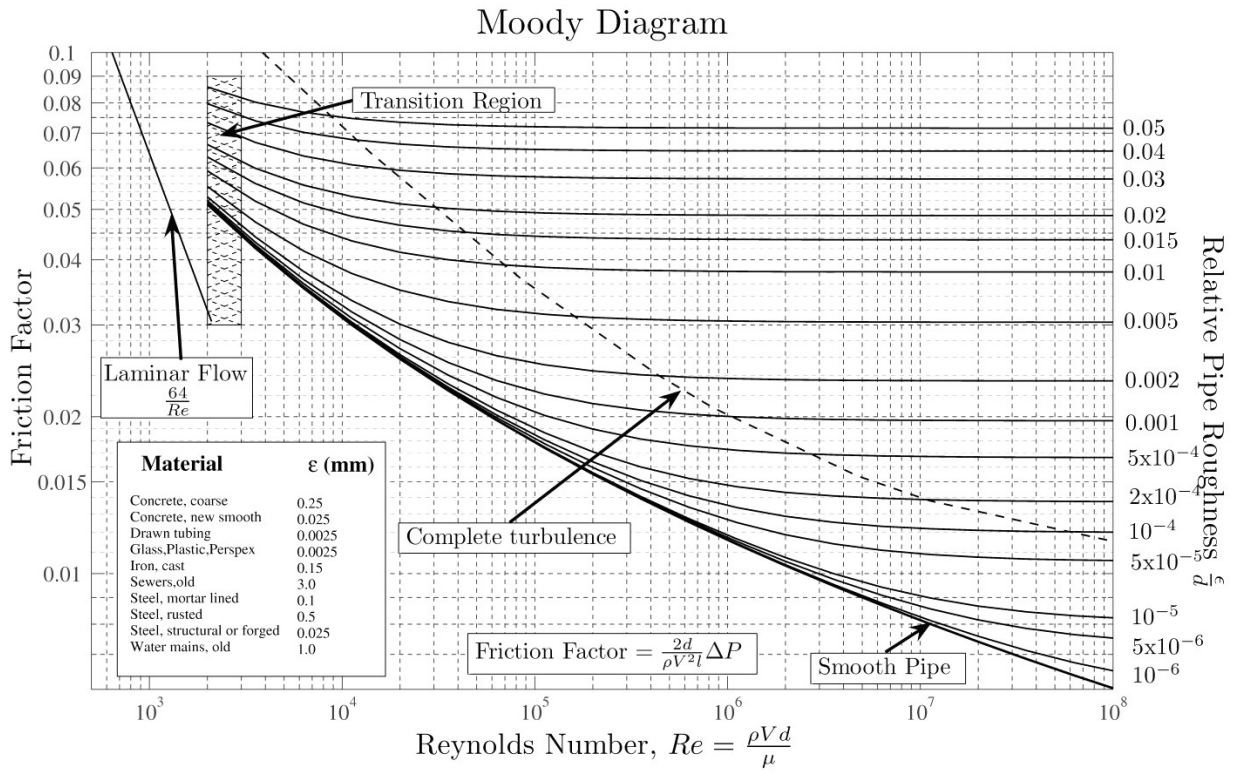


Fig. 1

Figure 3. Moody Chart

(Note: This Moody diagram is the original from the 1944 paper. It may still be under copyright. If so, it can be replaced by



from
http://en.wikipedia.org/wiki/File:Moody_diagram.jpg
which is released under the GNU Free Documentation License.)

In this one chart, one can read the pressure drop associated with the flow of fluid of specified speed, density and viscosity in a pipe of specified diameter and surface roughness.

In so far as the chart merely summarizes the outcomes of experiments already performed, it is free of entanglement with the concerns of this chapter. One need make no reference to background facts when one reports that a simple formula generates a curve that happens to pass through the data points near enough. I will complain shortly of the ambiguity in the notion of simplicity. That ambiguity is untroubling here. We can use any formula that generates a curve that fits the data well enough. The choice is purely pragmatic.

This purely pragmatic use of simplicity is an extreme case. I believe that it is rarely and possibly never realized in all purity. The examples above do not realize it fully. The Moody chart is a summary of past experience. But it is also a great deal more. Its primary use is as an instrument of prediction. The presumption is that, if an engineer constructs a piping system with flow rates, fluid densities, and so on, matching conditions covered by the chart, then the relevant curve will reflect the pressure drop that will be found. That can only happen if the chart is properly adapted to broader facts about fluid flow in pipes in general.

These facts have the character of simplicity assumptions. We must assume that the variables included are all that matters. Temperature does not enter into the chart explicitly; it is assumed that thermal effects are fully captured by the unrepresented dependence of density and viscosity on temperature. We must assume that the curves fitted to the data points interpolate correctly between them so that the chart makes good predictions for cases whose precise combination of variables have never been observed.

In so far as the descriptions seek to go beyond past experience, they seek the type of epistemic success to which the subsequent discussion applies.

4. Principles of Parsimony: Ontic Simplicity

The notion that parsimony can successfully guide us epistemically has many expressions and one might despair of categorizing them all successfully. There is, however, a broad division between ontic simplicity and descriptive simplicity. I will discuss ontic simplicity first and later turn to descriptive simplicity.

In this ontic version of the principle, we are guided to the truth by favoring accounts that posit the fewest entities or processes in the world. The locus classicus of this notion is "Ockham's razor." Its now universal formulation is

Entia non sunt multiplicanda praeter necessitatem.

Entities must not be multiplied beyond necessity.

Curiously this formulation is not to be found in the writings of the fourteenth century scholastic, William of Ockham. His closely related pronouncements include[60]

It is useless to do with more what can be done with fewer.

A plurality should not be assumed without necessity.

It has been an historical puzzle to locate the source of the popular formulation.[61] Another puzzle is that Ockham's name should be so exclusively attached to this maxim of simplicity, for it was an idea that, according to Mauer (1999, p. 121), was used commonly from the thirteenth century, after being gleaned from Aristotle.

The greater puzzle is why modern thinkers would look to a fourteenth century scholastic for this sort of guide. His understanding of the demand of parsimony was rather different from its modern use in science. He felt it not binding on god. He felt, as Maurer (1999, p. 120) reports,

---

[60] Quoted from Maurer (1999), p. 121. The Latin is "Frustra fit per plura quod potest fieri per pauciora." "Pluralitas non est ponenda sine necessitate."

[61] See Thorburn (1918).

that "God is not bound by it; he does many things by more means which he could do by fewer, and yet this is not done uselessly, because it is God's will."[62]

The better-formulated statement of the sentiments in Ockham's razor are Newton's two rules of reasoning as quoted in the last Section. The notion is advanced explicitly as a rule of reasoning; and Newton provides a justification. "Nature does nothing in vain." and "Nature is pleased with simplicity, and affects not the pomp of superfluous causes." The justification is dressed in anthropomorphic garb. Nature, surely, is never literally pleased or displeased. Without that garb, Newton is enjoining us to infer to the simpler case of fewer causes because the world is simpler, harboring fewer causes. This is a factual claim about the world.

This justification seems routine. We find it reported in Aquinas (1945, p. 129) (writing before Ockham's birth):

> If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices.

4.1 Its Difficulties

This ontic version of the principle of parsimony faces many difficulties. The most immediate is that we have no general prescription for how to count the entities, processes or causes to which the principle is applied. It is not hard to find ambiguity sufficiently severe as to compromise the principle.

How do we count entities when we compare a continuum and a molecular theory of gases? The continuum theory represents the gas as a single, continuous fluid. The molecular theory represents it as a collection of very many molecules, of the order of 10^{24} in number for ordinary samples of gases. Do we count one entity for the continuum gas and 10^{24} for the molecular gas, so that the molecular gas posits many more entities? Or do we invert the count? A continuum is indefinitely divisible into infinitely many parts.[63] The molecular gas consists of

---

[62] My reaction to this puzzle is that we have fallen into introducing a defective principle of parsimony with the faux dignity of a pedigreed Latin maxim in the hope that it will deflect a skeptical demand for justification. We may be unprepared to justify just why the two entity hypothesis is better than the three. But we can plant the idea that it is what *everyone* thinks and has done so since the fourteenth century as a way of stalling skeptical challenges. On a superficial survey of its use, it appears that this subterfuge is working pretty well.

[63] In statistical physics, this gives a continuous entity such as a field infinitely many degrees of freedom and is responsible for the "ultraviolet catastrophe" of classical electromagnetic fields.

finitely many molecular parts. Has the continuum now infinitely many more parts than the molecular gas?

This discussion in terms of entities can be converted into the causal conception of Newton's rules. What causes the pressure exerted by a gas jet impinging on a surface? Do we count the impact of the continuum gas as one cause? Or do we count an infinity of causes for the infinitely many impacts of its infinitely many parts?

What of the justification for this ontic principle? Newton asserts the world is simpler in employing fewer causes. That assertion is empty in so far as the counting of causes is ill-defined. However even setting that concern aside, the claim is still unsustainable. Nature is not simple. Traditional alchemical theories posited three or four elements in an attempt to account for chemical appearances. We now know that this count is far too low. A tractable chemistry requires over ninety elements. Perhaps Nature is pleased with chemistry, but surely not for the simplicity of the count of elements.

The existence of isotopes is especially damaging to Newton's justification. For one can explain the chemistry of carbon quite well just by assuming that there is one element, carbon. Hence, Newton's rules urge, we should infer to there being just one element carbon since Nature "affects not the pomp of superfluous causes." That is, we should infer to the one element and not to the existence of multiple types of chemically identical carbon, because that is the way Nature is. Yet that is not the way Nature is. Carbon exists in multiple, chemically identical isotopes, Carbon-12, Carbon-13 and Carbon-14. The recommendation to infer to just one type of carbon may well be good advice as far as the chemistry of carbon is concerned. I do not wish to impugn this recommendation or to suggest that an inductive rule is defective because it sometimes leads us astray. That is part of the risk we take whenever we carry out inductive inference. Rather the issue here is the justification. While the recommendation may be good, it cannot be justified by a supposition that factually there is just one type of carbon. Factually, there is not.

4.2 Rescue by the Material Theory: the Principle as an Evidential Truism

This ontic form of the principle of parsimony is troubled. Yet it has figured and continues to figure prominently in successful science. There must be something right about it. The clue to what is right lies in the ambiguous qualifications found in every formulation. We are not to multiply entities "beyond necessity." We are to admit no more causes than are "sufficient to explain…" We assign the same cause "as far as possible." These qualifications mean the

---

Correspondingly, a molecular gas has finitely many degrees of freedom as a result of its finitely many molecules.

principle is not self-contained. Something more supplies the sense of necessity, possibility and sufficiency.

Newton's formulation gives us the clearest indication of what that something is. We are not to proceed beyond that which is "sufficient to explain the[ir] appearances." That gives us some flexibility. We can add to the causes we admit as long as they are sufficient to explain the appearances. We are not to go beyond that sufficiency. Since we routinely infer inductively to that which we count as explaining the appearances, this amounts to telling us to infer to no more than that for which we have inductive authorization. Understood this way, the principle is revealed to be a truism of inductive inference, which says:

> We should not infer to more than that for which we have good evidence.

It is a corollary of another truism: we should infer inductively only to that for which we have good evidence.

How did an inductive truism become enmeshed with the muddle of the metaphysics of simplicity? The key relevant fact is that the truism is not an independent inductive principle; it is a meta-inductive principle. That is, it is not a principle *of* an inductive logic. Rather, it is a principle *about* how other inductive logics should be used. That this is so is harder to see if one conceives of inductive inference formally. The principle is entangled with assertions about how the world is factually. If one understands inductive inference materially, however, that entanglement is expected. Moreover it clarifies how the original principle can be a good epistemic guide.

We can see this entanglement in Newton's first use of his Rules I and II. In Book III of *Principia*, Proposition IV Theorem IV asserts that the force of gravity that draws objects near the earth's surface is the same force that holds the moon in its orbit. He assumes that the force acting on the moon intensifies with decreasing orbital radius according to an inverse square law, as it does with other celestial objects. It follows that were the moon to be just above the earth's surface, it would fall to earth with the same motion as ordinary bodies fall by gravity. He continued:

> And therefore the force by which the Moon is retained in its orbit becomes, at the very surface of the Earth, equal to the force of gravity which we observe in heavy bodies there. And therefore (by Rule 1 & 2) the force by which the Moon is retained in its orbit is that very same force which we commonly call gravity; for were gravity another force different from that, then bodies descending to the Earth with the joint impulse of both forces would fall with a double velocity…

Newton here invokes his rules to complete the inference. However the inference is already fully controlled by a factual presumption: that the matter of the moon is the same as the matter of the earth and, if brought to the surface of the earth, would behave like ordinary terrestrial matter.

That factual assumption already authorizes Newton's conclusion and he gives the reason. Were there to be some additional celestial force that acts on the matter of the moon but not on ordinary terrestrial matter, then the moon would fall with double the motion of ordinary terrestrial matter. That contradicts the assumption that the matter of the moon behaves just like that of the earth. This is a kind of simplicity assumption: contrary to ancient tradition, there is no difference between terrestrial and celestial matter. But its comprehension and use make no appeal to abstract metaphysical notions of simplicity. It is a specific factual statement and it powers the inductive inference, as the material theory requires.

We also see this entanglement in the illustrations Newton supplies for his Rules. To illustrate Rule II, he considers the "light of our culinary fire and of the sun." We are to assign the same cause to both. We now know that this is an erroneous conclusion. Culinary fires generate light from combustion; the sun generates light by a different process, nuclear fusion. What makes the inference appear unproblematic for Newton is that he is really relying on a tacit background assumption: that it is very unlikely that there is a process that produces intense light other than combustion. That fact powers the inductive inference.

In short, this ontic formulation of the principle of parsimony fails as a universal principle of inductive inference. It is too vague to be applied univocally and efforts to give it a foundation in a supposed general, factual simplicity of the world founder. Its successes, however, can be understood in so far as it is the meta-inductive principle that one should infer inductively to no more that for which one has good evidence. The assertions of simplicity are veiled invocations of relevant facts that authorize the inductive inference, in accord with the material theory.

5. Principles of Parsimony: Descriptive Simplicity

These descriptive versions of the principle do not directly address the numerical simplicity of entities or causes. Instead we are enjoined to favor the simplest *description* of processes pertaining to them. It may not be possible to effect an absolute separation between the descriptive and the ontic versions of the principles of parsimony. For descriptive simplicity is tacitly supposed to reflect some sort of ontic simplicity. However the explicit focus on language introduces sufficient complications to need a separate analysis.

The best-known application of descriptive simplicity is curve fitting. In its simplest form, we are given a set of data points, that is, many measured values of a variable x and a variable y. These are represented as points on a graph, as shown below. We then seek the curve that fits them best. It is routine in curve fitting to start with a constant relation, then a linear one, then a quadratic, and so on, seeing how much better is the fit of the curve as we proceed to higher order polynomials, as shown in Figure 4.

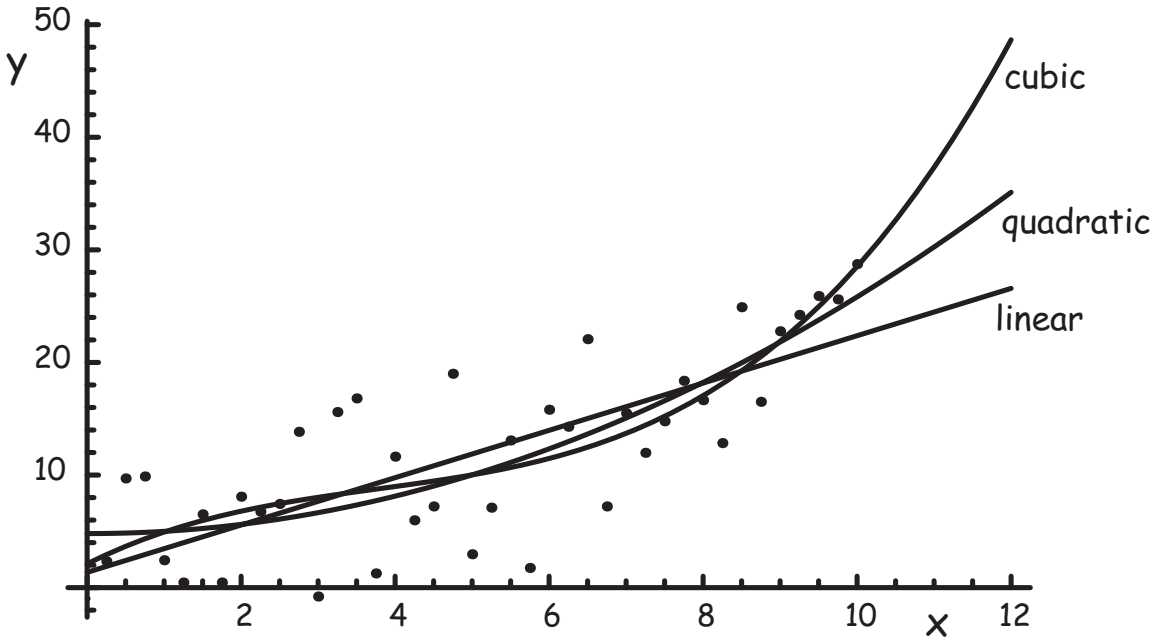


Figure 4. Polynomial Curve Fitting

The fit will always improve as we increase the order. The higher the order of the polynomial, the larger the repertoire of curves available and hence the more likely we are to come close to the data points.

Eventually, however, this greater flexibility will cause trouble. For the data is routinely assumed to be a compound of the true curve sought and confounding error. If the true law sought is merely a linear curve, the error will scatter the data around the true straight line. Higher order polynomial curves will have little trouble adapting to the random deviations due to noise. This will lead the fitted curve to deviate from the true curve as it responds to the vagaries of the noise. Figure 5 shows the best fit of linear and eighth order polynomial curves to a data set.

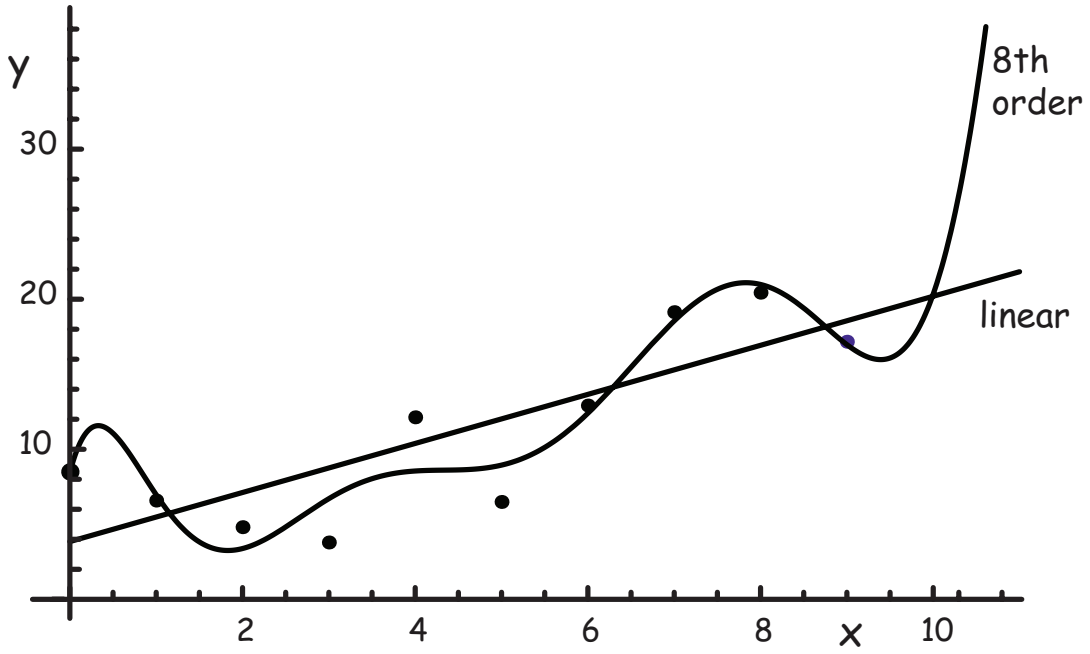


Figure 5. Overfitting

The apparently better fit of the higher order curves is spurious. This phenomenon is known as "overfitting."

The primary burden in curve fitting is to find a balance of the two effects: the simplicity of the curves that fit less well against the better fit of more complicated curves that overfit. The simplicity of a curve is derived from its description. The polynomial family consists of smaller, nestled families of curves:

Constant: $y = a$

Linear: $y= a + bx$

Quadratic: $y = a + bx + cx^2$

Cubic: $y = a + bx + cx^2 + dx^3$

Quartic: $y = a + bx + cx^2 + dx^3 + ex^4$

and so on.

That is, the formulae that describe the curves have more parameters as we proceed to the less simple, higher order polynomials. The constant curve has one parameter, a. The linear curve has two parameters a and b. The quadratic curve has three parameters, a, b and c. And so on. Built into the practice of curve fitting is a simplicity postulate: favor those curves whose descriptions require fewer parameters.

The preference for simpler descriptions has been applied more broadly. Perhaps its best-credentialed proponent is Albert Einstein. The laws of fundamental theories of physics employ constants. Newton's theory of gravitation employed the gravitational constant G. Einstein's special theory of relativity employed the speed of light c and his general theory employs both c

and G. Quantum theory employs Planck's constant h, as well as numerous quantities characteristic of the various particle interactions, such as the charge of the electron e. The standard model of particle physics now employs nearly 20 such constants. Some of these constants can be absorbed into the system of units used. The speed of light c can be suppressed merely by measuring distance in light years; then the speed of light reduces to unity.

Einstein (1949, p. 61-63) grounded his hope for a physics free of all further constants in a belief in the simplicity of nature

> If one considers this [suppression] done, then only "dimension-less" constants could occur in the basic equations of physics. Concerning such, I would like to state a theorem which at present cannot be based upon anything more than upon a faith in the simplicity, i.e., intelligibility, of nature: there are no *arbitrary* constants of this kind; that is to say, nature is so constituted that it is possible logically to lay down such strongly determined laws that within these laws only rationally completely determined constants occur (not constants, therefore, whose numerical value could be changed without destroying the theory).

While the freedom from these constants reflects something factual in the structure of the world, Einstein expresses it in terms of the descriptions of that structure, that is, in terms of the constants appearing in the equations that describe it. Just as curve fitting should favor smaller numbers of parameters, Einstein favors laws with the fewest arbitrary parameters.

These sentiments come from Einstein later in his life. By then he had abandoned his earlier allegiance to positivistic approaches. He had become a mathematical Platonist and that was a doctrine, he assured us, he had learned from his experiences in physics.[64] His 1933 Herbert Spenser lecture, "On the Methods of Theoretical Physics," offers an explicit and powerful manifesto:

> Our experience hitherto justifies us in believing that nature is the realisation of the simplest conceivable mathematical ideas. I am convinced that we can discover by means of purely mathematical constructions the concepts and the laws connecting them with each other, which furnish the key to the understanding of natural phenomena. Experience may suggest the appropriate mathematical concepts, but they most certainly cannot be deduced from it. Experience remains, of course, the sole criterion of the physical utility of a mathematical construction. But the creative principle resides in mathematics. In a certain sense, therefore, I hold it true that pure thought can grasp reality, as the ancients dreamed.

---

[64] For an account of precisely how Einstein's experience with general relativity led him to this, see Norton (2000).

Einstein continues to detail how we can use mathematical constructions to make discoveries in physics:

> The physical world is represented as a four-dimensional continuum. If I assume a Riemannian metric in it and ask what are the simplest laws which such a metric can satisfy, I arrive at the relativistic theory of gravitation in empty space. If in that space I assume a vector-field or an anti-symmetrical tensor-field which can be derived from it, and ask what are the simplest laws which such a field can satisfy, I arrive at Maxwell's equations for empty space.

The recipe is one of descriptive simplicity. In each context one writes the simplest admissible equations and thereby recovers the law.[65]

5.1 Its Difficulties

The difficulty with any principle expressed in terms of descriptions is that it can be rendered incoherent by merely altering the descriptive system used. In so far as the simplicity principle of curve fitting merely requires us to favor the curves with fewer parameters, it is unsustainable. Merely rescaling the variables used can overturn its judgments completely, as will be illustrated in the following section.

The idea that we get closer to the truth by writing mathematically simpler laws has the imprimatur of Einstein. However it unsustainable for the same reasons that trouble curve fitting. Judgments of just what is descriptively simple are too malleable. Einstein's own general theory of relativity illustrates the problem. When the theory first came to public notice after the eclipse tests of 1919, it was notorious for its abstruse difficulty. The eminent astronomer George Ellery Hale confided in correspondence:[66]

> …I confess that the complications of the theory of relativity are altogether too much for my comprehension. If I were a good mathematician I might have some hope of forming a feeble conception of the principle, but as it is I fear it will always remain beyond my grasp.

---

[65] Here are the technical details for those who want them. The simplest non-trivial structure in the derivatives of the metric tensor g_{ik} is the Riemann curvature tensor, R^{i}_{kmn}. Its vanishing requires the flatness of spacetime, which is too restrictive. The vanishing of its unique first contraction, R_{ik}, is the Einstein gravitational field equation for empty space. The vector field is the vector potential A_i and the tensor field mentioned is the Maxwell field tensor $A_{i;k} - A_{k;i}$. Setting its divergence to zero returns the source-free Maxwell equations.

[66] February 9, 1920. Quoted in Clark (1984, pp. 299-300).

The *New York Times* of November 19, 1919, reported an incredible tale, reflected in the partial headline "A book for 12 wise men":

> When he [Einstein] offered his last important work to the publishers he warned them there were not more that twelve persons in the whole world who would understand it, but the publishers took it anyway.

The fable took root. It is repeated in the publisher's introduction to Lorentz's (1920, p.5) popularizations of relativity theory.

As the decades passed, general relativity was absorbed into mainstream physics and opinions began to migrate. By the 1970s, the standard textbook for the theory came to a startlingly different conclusion (Misner, Thorne and Wheeler, 1973, pp. 302-303):

> "Nature likes theories that are simple when stated in coordinate-free, geometric language."…According to this principle, Nature must love general relativity, and it must hate Newtonian theory. Of all theories ever conceived by physicists, general relativity has the simplest, most elegant geometric foundation … By contrast, what diabolically clever physicist would ever foist on man a theory with such a complicated geometric foundation as Newtonian theory?

How is a reversal of this magnitude possible? The key is the centrality of "coordinate-free, geometric language." One finds general relativity to be the simpler theory when one adopts the appropriate language. As a result, the principle of descriptive simplicity, as enunciated by Einstein, is incomplete. Without a specification of the right language to be used, it can give no direction at all.

Perhaps one might hope that somehow mathematics provides the natural descriptive language for our science and physics in particular. A cursory glance at the interplay of mathematics and science shows things to be different. There is no unique language for physical theories. New physical theories commonly appear mathematically difficult and even messy. That is followed by efforts by mathematicians and the scientists themselves to simplify the presentation and manipulations of the theory. As I have argued in Norton (2000, pp. 166-68), what results are new mathematical methods and new formulations of the theories that become successively simpler.

Newton's development of his mechanics employed the ancient methods of Euclidean geometry. His contemporaries required considerable insight and facility in geometry to follow and emulate his difficult demonstrations and proofs. Over the subsequent centuries, Newton's theory was re-expressed in terms of algebraic symbols and the calculus. Many of what were once abstruse results became natural and simple. Quantum mechanics developed in the first quarter of the 20th century. The theory that resulted in the late 1920s was a complicated mess of differing approaches and techniques: matrix mechanics, wave mechanics, Dirac's c- and q-numbers.

Subsequent efforts showed all the theories to be variant forms of a single theory that found its canonical mathematical formulation in von Neumann's 1932 classic, *Mathematical Foundations of Quantum Mechanics*. Even Einstein's general relativity benefited from this reforming. His original methods did not include the now key notion of parallel displacement. This notion was introduced in response to the completion of the theory by the mathematician Levi-Civita in 1917.

5.2 Rescue by the Material Theory: Adaptation of Language as a Factual Judgment

As before, we must acknowledge that there is something right in the idea that descriptive simplicity is a guide to the truth. What is right is already clear from the discussion above. The real epistemic work is done in finding and developing a language or descriptive apparatus appropriate to the systems under investigations. What makes that apparatus appropriate is precisely that the truths concerning the system find simple expression in it. Then it is automatic that seeking simple assertions in the language or descriptive apparatus leads us to the truth.

That is, the principle of descriptive simplicity guides us to the truth in so far as the language we use is properly adapted to the background facts. Hence what is really guiding us is not some abstract notion of simplicity but merely the background facts as reflected in our choice of descriptive language. This inductive guidance from background facts is, of course, precisely what is called for by the material theory of induction. This idea will be illustrated with the example of curve fitting in the next section.

6. Curve Fitting and the Material Theory of Induction

As a mode of discovery, curve fitting is based on the idea that fitting a simple curve to data can guide us to the truth. The material theory of induction requires that these inductive inferences are warranted by background facts. Here I will describe in greater detail the character of these background facts. We will see that the vague and incomplete injunctions to seek the simpler curve translate into more precise constraints, expressed in terms of these background facts. The characteristics of background facts found in many but not all cases of curve fitting can be grouped under three headings, as below.

6.1 The Error Model

When curve fitting is used to seek some underlying truth or law, the presumption is that the data to which the curve is fitted have been generated by a standard error model of the form:

$$\text{Error laden data} = \text{true curve} + \text{error noise}$$

That curve fitting operates with data produced by this model is so familiar that it is easy to overlook its importance. The techniques of curve fitting are designed to strip away confounding

noise. Thus the assumption of the standard error model must be true if these techniques are to guide us towards the truth.

A quick way to see its importance is to consider the curve fitting problem shown in Figure 6. We seek the value of the quantity a that gives the best fit of y = 1/[log(x)-a] to the data.

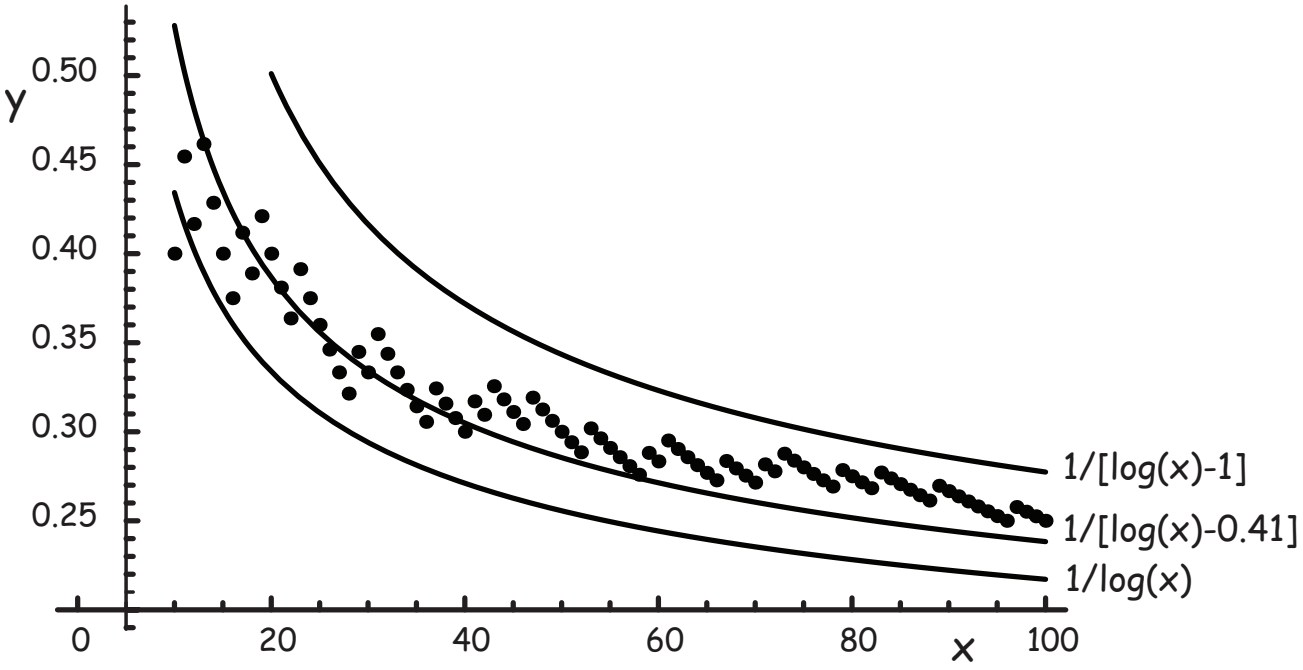


Figure 6. A Curve Fitting Problem

The optimum value turns out to be the value shown, a = 0.41.

Superficially, the problem looks traditional. However this curve fitting problem does not conform to the standard error model. The data represent the density of prime numbers between x= 10 and x=100. The first datum at x=10 reports the number of primes between 1 and 10. There are four of them: 2, 3, 5 and 7, so that y= 4/10 = 0.4. The prime number theorem assures us that the density y(x) approaches 1/log(x) for large x. A corollary is that the density also approaches 1/[log(x) – a] for *a* some constant, for the two quantities, 1/log(x) and 1/[log(x) – a] approach one another for large x. The curve fitting problem is to ascertain which value of a gives the best fitting curve for values of x in the range 10 to 100 covered by the data. The result is 0.41 as shown.[67]

Instead of the standard error model, this problem conforms to a non-standard error model in which truth and error are permuted:

True data = error laden curve + error noise

[67] This is specifically for primes in the range specified. The optimum value for all primes is a = 1.

This means that, epistemically, the exercise is different. We are not seeking truth. We already have the complete truth in the data that report the true density of prime numbers. Instead we are seeking a summary that has least deviation from the truth, where the notion of "least deviation" is one we can choose conventionally. In this case, I chose a fit that minimized the sum of squared deviations.

Engineering applications, such as the Moody diagram above, illustrate a second way that we may deviate from the standard error model. In so far as we are merely seeking a compact summary of past experience, there is no real error model in use at all, for there is no hidden truth sought. Our ambitions, however, are rarely so limited. For example, as noted above, the Moody diagram is typically not intended merely as a compact historical report. It is also intended as a predictive tool. In so far as that is the case, the standard error model is presumed.

However the practice is somewhat protected from the full rigors of the model by the fact that engineering practice rarely requires perfectly exact prediction of pressure drops in pipes. A prediction correct to within a few percent is more than sufficient for most applications. This affords great protection when we seek predictions for conditions that interpolate between those used to create the original chart. Fitting just about any family of curves will interpolate to the requisite level of accuracy. In effect we are conforming the data to a weaker model:

Error laden data = near enough to true curve + error noise

Near enough to true is good enough for interpolation.

This protection is lost when we seek to extrapolate to new conditions outside those used to create the diagram. For then two curves that each interpolate among the data equally well may diverge markedly when we extend beyond the condition in which the data were collected. Then we need to find which is the true curve on pain of uncontrolled errors entering our predictions. This divergence is illustrated in Figure 4 above. The polynomials in the figure interpolate the data comparably well in the range of $x = 0$ to $x = 10$. They diverge rapidly outside the range in which the data was collected, giving markedly different results in the range $x = 10$ to $x = 12$.

6.2 The Parameterization

Descriptive simplicity can only be a good epistemic guide to the truth, I have urged, if the language of description is chosen so that the truths correspond to simple assertions. In the case of curve fitting, that condition translates into a matching with background facts of the parameterization used and the family of its functions from which the curves are drawn.

We are free to rescale the quantities used to describe our measurements; and we do. We may compare cars by their speeds or, equivalently, by the times they take to cover some fixed distance. Since one parameter is the inverse of the other, fitting some family of curves to speed will in general give different results from fitting the same family of curves to times. This

reparameterization is common. In acoustics, we measure loudness of sounds in decibels, which is a logarithm of the power of the sound. In astronomy we measure the apparent magnitude of the stars on a scale that is the logarithm of the intensity (that is, the energy flow per unit area).

To see how the parameterization we choose makes a difference, we will develop an example in which the data are generated by a true linear relation y = x, as in Figure 7. The data has been simulated with very little noise, so the best fitting straight line[68] comes so close to y=x.

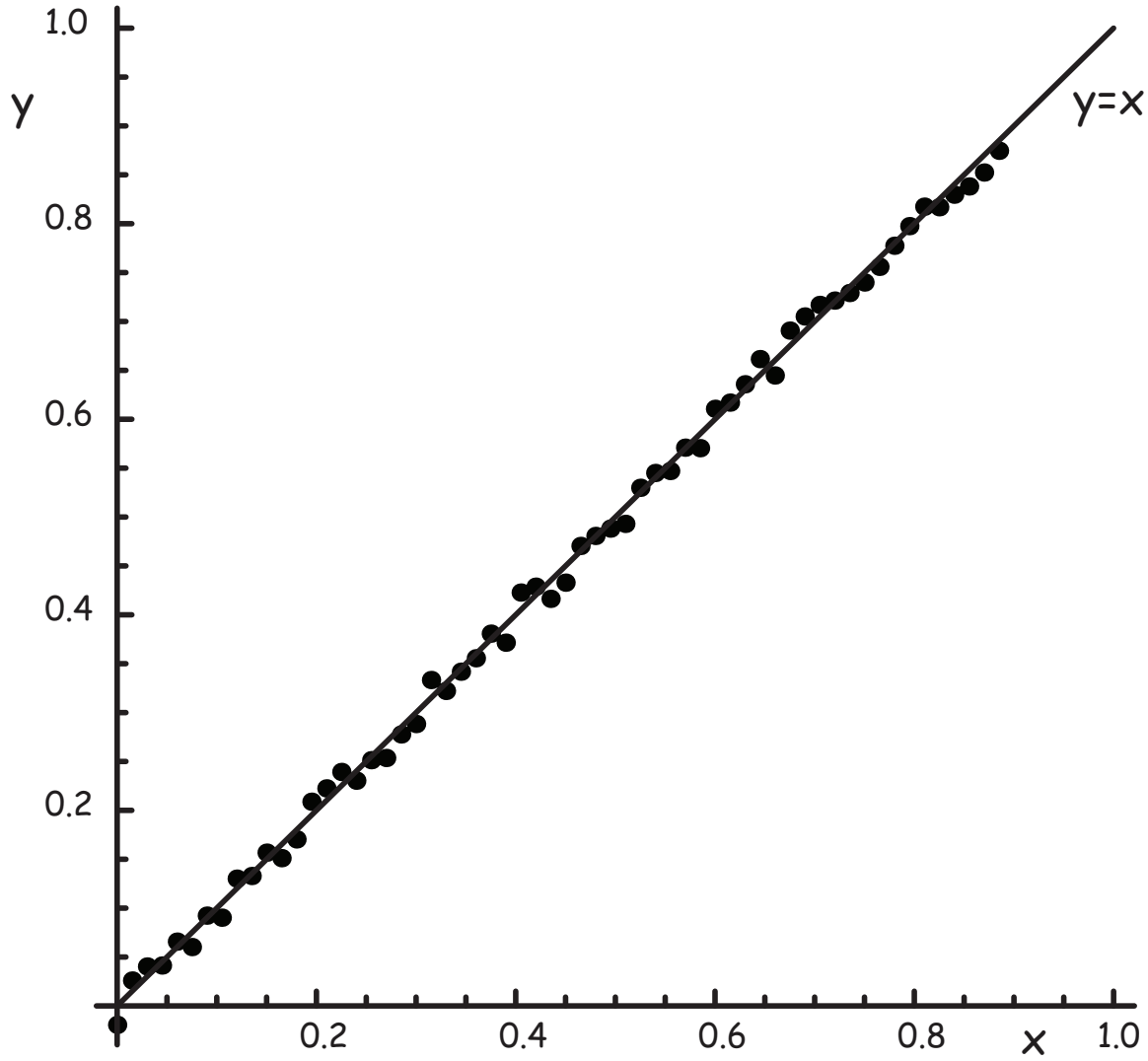


Figure 7. Data generated from true curve y=x

Rescaling the x variable to $z = \sin^{-1}(x)$, means that the true curve rescaled to y = sin(z). However a polynomial curve fit between y and z will never return this curve, for y=x is equivalent to a polynomial of infinite order in z:

[68] The best fitting straight line is y = -0.000403379 + 0.996917 x. It is not shown in Figure 7 since it is too close to the true curve y=x to be separated visually.

$$y = \sin(z) = z - (1/3!)z^3 + (1/5!)z^5 - (1/7!)z^7 + \ldots$$

A curve fitting algorithm that proceeds up the family of polynomials in z will necessarily halt at some finite order and so cannot return the true curve. Finding polynomials of best fit for the rescaled data of Figure 7 shows how poorly the polynomial fit performs. The best fitting linear, quadratic, cubic and quartic polynomial curves interpolate the data well. However, as Figure 8 shows, they fail immediately on extrapolation beyond the domain x = 0 to x = 0.9 in which the data was generated.[69]

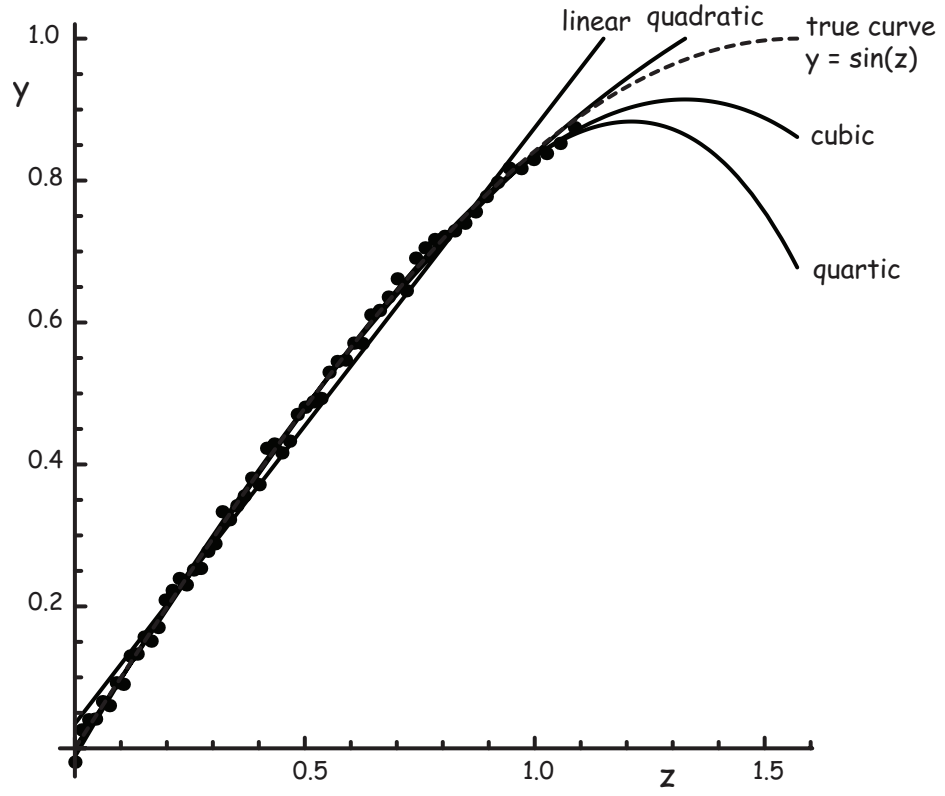


Figure 8. Failure of polynomial curve fit on reparameterized data

The problem is the same in the reverse direction. If y=z is the true curve, then this true curve corresponds to an infinite polynomial if we parameterize the data using x, for

$$y = \sin^{-1}(x) = x + (1/6)x^3 + (3/40)x^5 + (5/112)x^7 + \ldots$$

for $|x|<1$. Once again ascending the family of finite polynomials will never return the true curve.

Choosing the right parameterization and family of curves amounts to properly adapting them to the background facts. If the attenuation of the intensity I(r) of some signal with distance r is due to an absorptive medium, then the signal attenuates as $I(r) = I(0) \exp(-\lambda r)$, for λ some

[69] The domain x= 0 to x = 0.9 corresponds to z = 0 to $z = \sin^{-1}(0.9) = 1.12$. The largest x shown, x=1, corresponds to $z = \sin^{-1}(1) = \pi/2 = 1.57$.

constant. The exponential dependence of I(r) amounts to another infinite order polynomial in r. If we rescale, the relation reduces to a simple linear dependence of the logarithm of I(r) on r, for then the attenuation follows log I(r) = constant – λr. If, however, the attenuation is due to spreading it space, signal intensity will attenuate according to $I(r) = A/r^2$, for some constant A. This corresponds to log I(r) = A – 2 log(r), which once again corresponds to an infinite order polynomial of in r. However, if we use both log I(r) and log(r) as our parameters, then the true curve is linear and its slope, -2, conveys the fact that the attenuation follows an inverse square law.

Perhaps the clearest example of this adaptation of the parameters and curves to the background facts arises when we have processes that are periodic in time. We should then use the time t as the parameter. The family of curves to be fitted should not be polynomials, since they are not periodic. Rather we should use the family of periodic trigonometric functions, sin(t+a), sin(2t+b), sin(3t+c), etc, where the a, b, c, … are constant phase factors. We learn from Fourier analysis, that this family is sufficiently rich to represent all periodic curves of likely interest to curve fitters.

6.3 The Order Hierarchy

We must have an adaptation of the descriptive language to background facts if descriptive simplicity is to be an effective guide to truth. In important cases the adaptation can be especially tight. Curve fitting proceeds with some collections of families of curves, such as the constant, linear, quadratic, etc. In these important cases, the families of curves fitted correspond directly to particular processes. Then fitting a curve from a family farther up the hierarchy corresponds to the inclusion of more processes in the account developed of the phenomena. Further the adaptation has to be such that curves fitted earlier in the procedure correspond to stronger or more probable processes.

This adaptation will be illustrated in the following two sections with the cases of fitting trajectories to celestial objects and the harmonic analysis of the tides.

7. Fitting Orbital Trajectories

The standard method of curve fitting is to find that curve that minimizes the sum of the squares of deviations of the curve from the data. This least squares technique was introduced around the start of the 19th century in astronomy to assist the fitting of orbits to celestial objects in our solar system. This application illustrates a tight adaptation on the curve fitting method to the background assumptions that are the surrogates of simplicity. The family of curves fitted reflects the particular trajectories that background assumptions select. Moreover ascending the

order hierarchy reflects pursuit of trajectories according to their likelihood and the strength of the processes that form them.

7.1 Ellipses, Parabolas and Hyperbolas

A new celestial object—a new planet or comet, for example—is sighted. The astronomers' first task is to find the orbit that fits the positions seen. Astronomers do not follow the curve fitters' generic procedure of seeking first to fit a straight line and then proceeding up through higher order polynomials. Rather the family of curves chosen is provided by gravitation theory. The initial model is provided by the "one body problem": the motion of a free body attracted to a central point by a force that varies inversely with the square of distance r to the point. That is, the attracting force is k/r^2 for k a suitable constant. The familiar result, given early in any text on celestial mechanics,[70] is that the trajectory is one of the three conics sections: an ellipse, a parabola or a hyperbola.

Select polar coordinates (r,θ) in the plane of the orbit with the origin r=0 at the center of force in the sun and set $\theta=0$ at the perihelion, the point of closest approach to the sun in the orbit. A single formula that covers all three curves is

$$r = \frac{L}{1 + e\cos(\theta)}$$

where e is the eccentricity and L is the semi-latus rectum that, loosely speaking, fixes the width of the figure. (More precisely, it is the distance from a focus to the curve along a line perpendicular to the major axis.) We pass among the conic sections with equal semi-latus recta by changing the eccentricity e. A circle is e=0, an ellipse is $0<e<1$; a parabola is e=1; and a hyperbola is $e>1$.[71]

Figure 9 shows trajectories of the three types of conic section with equal semi-latus rectum:

---

[70] I happen to be using Sterne (1960, §1.3) here.

[71] The semi-latus rectum is related to the semi-major a by $L = a(1-e^2)$ for both an ellipse and an hyperbola if we adopt the convention that a is positive for an ellipse and negative for an hyperbola.

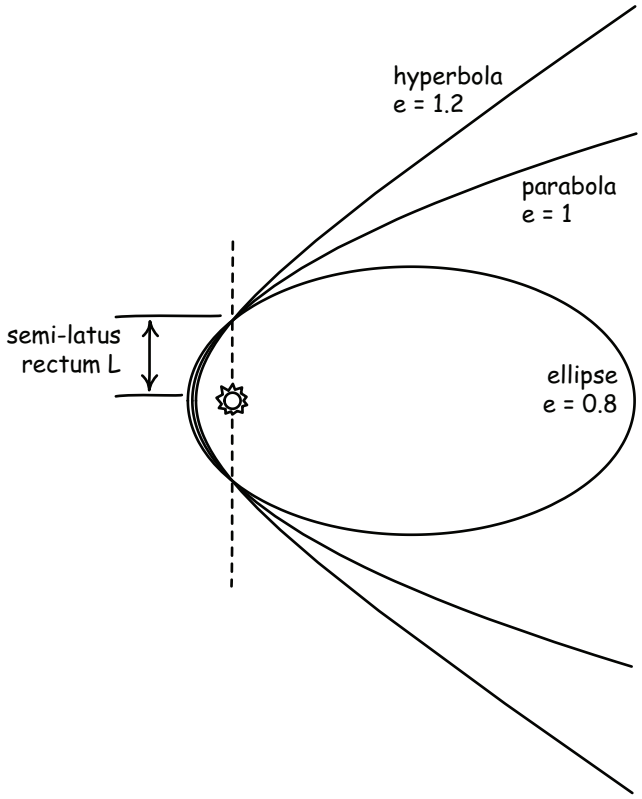


Figure 9. Conic Sections with Equal Semi-Latus Rectum

Once the semi-latus rectum L and eccentricity e are fixed, so is the orbital motion. Kepler's area law affirms that the radial line from the center of force to the object sweeps area at a constant rate with time t. That is the areal velocity $h=r^2 d\theta/dt$ is constant. This areal velocity is related to L by $h = \sqrt{kL}$. Hence the angular speed of the object in its orbit, $d\theta/dt$, is fixed once the trajectory is fixed.

Consider once again the default of the straight line of generic curve fitting. This familiar default is precluded by the background assumptions of gravitation theory. It can arise for objects in the solar system if they are moving at speeds very much greater than those normally encountered. Then the object will follow a hyperbolic trajectory with a very large eccentricity that is practically indistinguishable from a straight line.

The three conic sections provide the basis for the family of curves employed. When we allow for perturbations produced by gravitational attractions from other objects in the solar system as we shall see below, the family is enlarged by allowing a slight motion in the curve. For example, the major axis of the ellipse along which the object moves may rotate slowly. Representing that slow rotation introduces further parameters and provides the full family of curves used in virtually all accounts of orbital motion.

Fitting an orbit to a celestial object involves moving up this hierarchy of curves until a suitably close fit is obtained. One might try to describe this ascent as guided by some principle of parsimony that requires starting with the simplest curve and then moving up to more complicated ones. However it is hard to see just which abstract notion of simplicity might here lead us to identify conic sections as the simplest case with straight lines an extreme case never implemented. Fortunately no such notions of simplicity are needed to explicate the procedures used. The selection of curves and their order are guided by background assumptions on the likelihood of certain trajectories and on the likelihood and strength of processes that modify them in prescribed ways.

7.2 Comets

An easy illustration is provided by the methods used to fit orbits to newly discovered comets. That is, the illustration is easy if we limit ourselves to the methods routinely used in the nineteenth century. Watson (1861, pp. 163-169) describes the methods then customary. They depend essentially on the following background fact: Comets typically have very eccentric orbits and we get to observe them when they are in the vicinity of the sun. There, as Figure 9 above suggests, it becomes quite difficult to separate the ellipses and hyperbolas with eccentricity close to unity from each other and from a parabola.[72] This fact leads to the procedure described by Watson (1861, p.164):

> It is therefore customary among astronomers, when a comet has made its appearance unpredicted, to compute its orbit at first on the supposition that it is a parabola; and then, by computing its place in advance, find from a comparison of the actual observations, whether this hypothesis is the correct one. Should it be found to be impossible to represent the observed positions of the comet by a parabola, an ellipse is next computed and when this also fails, recourse is had to the hyperbola, which, provided the previous computations are correct in every particular, will not fail to represent the observations within the limits of their probable errors.

[72] The details: If the trajectory is a parabola with semi-latus rectum L, then the distance to the sun at perihelion is L/2. For a very eccentric ellipse or hyperbola, e is close to 1; that is $e = 1-\varepsilon$, where ε is small. (It is positive for an ellipse and negative for an hyperbola.) Hence $1-e^2 = 1-(1-2\varepsilon + \varepsilon^2) \approx 2\varepsilon$ to first order. Hence the semi-latus rectum $L = a(1-e^2) \approx 2a\varepsilon$. The distance to the sun at perihelion is $a(1-e) = a\varepsilon \approx L/2$, which agrees with the parabolic case.

That is, the known fact of the high eccentricity of comets directs a choice of a parabola to fit the initial data. The astronomers then collect more data and move to a nearby ellipse and then hyperbola if the deviations from the parabola are sufficient.

The sequence of shifts reflects a definite physical assumption about the energy of the comet. Adopting an ellipse amounts to assuming that the comet's total energy—kinetic plus potential—is negative so that it is bound to the sun and can never escape. Adopting the hyperbola amounts to assuming a positive energy sufficient to enable the comet to escape the sun's gravity. The case of the parabola is the intermediate case of zero energy, which is the minimum level at which escape from the sun's gravity is possible. Watson does not mention it here, but I believe the decision to try an ellipse after the parabola rather than an hyperbola reflects the prevalence of comets bound in elliptical orbits. Such comets will return periodically and thus are more likely to be seen by us. Adopting the hyperbola amounts to assuming that the comet will pass just once through our solar system, so that this is our one chance to see it. If the orbit is elliptical, we will get many chances.

There is a small element of arbitrariness in the procedure. Instead of fitting a parabola initially, the astronomers could have chosen an ellipse with an eccentricity imperceptibly different from unity. (That trajectory would be near indistinguishable from a parabola in the vicinity of the sun.) Whichever is chosen as the first curve, the choice is driven by the background physical assumption that the comet is just on the energetic edge of being gravitationally bound permanently to the sun. Further data then directs a decision to one or other side, within or beyond the edge. The selection of the curves fitted reflects this background physics.

7.3 Perturbed Orbits and New Planets

The conic sections discussed so far are grounded physically in Newton's law of gravity through the one body problem. However they are not the complete family of curves fitted to bodies moving in our solar system. Very careful measurements show that a planet will trace out an almost perfectly elliptical orbit only over the shorter term. If it is tracked over the longer term, however, deviations will appear. They are sufficiently small that they can be represented as changes in the elliptical orbit that the planet is following. If, for example, the axis of the ellipse rotates in the plane of the orbital motion of a planet, then the orbit actually traced out takes on the flower petal shape seen in Figure 10. It is an advance of the planet's perihelion, its point of closest approach to the sun, with each orbit of the sun. Or at least this is the shape traced out if we depict an advance unrealistically fast for any planet.

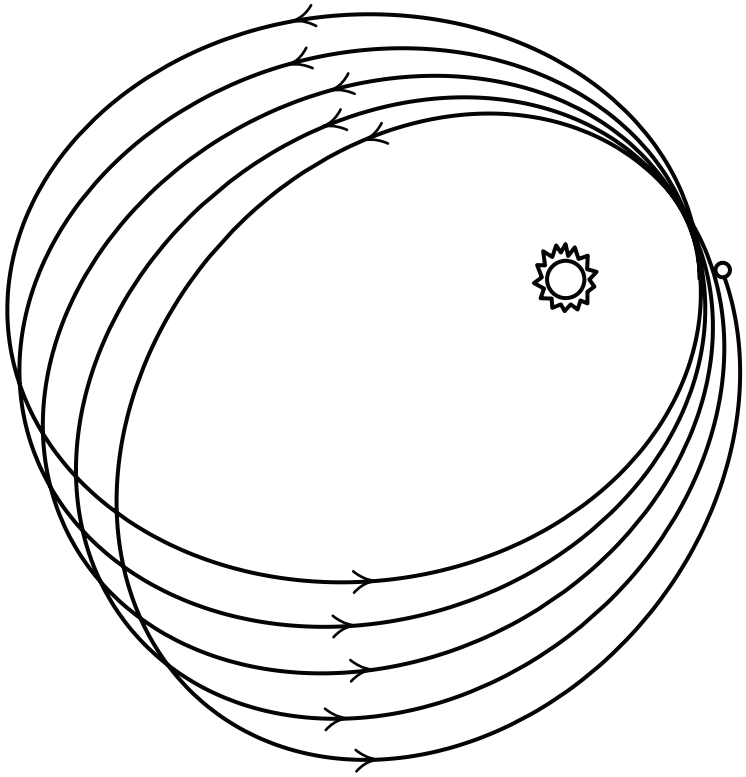


Figure 10. Advancing Perihelion Motion of a Planet

In general, the small deviations from the perfect elliptical orbits are represented by slow changes in the six Keplerian elements that characterize each elliptical orbit. The first two elements are the most familiar: the semi-major axis and eccentricity specify the shape of the ellipse. The remaining elements locate that ellipse in space and place the planet on it at the appropriate time.[73] The families of curves associated with these perturbed elliptical orbits are the ones fitted to the observed positions of the planets.[74]

These perturbed ellipses are fitted to the planets initially because they are found to fit. However, from the earliest time of Newtonian gravitation theory, the challenge has been to locate the physical cause for the perturbation, which is almost invariably sought in the perturbing

[73] The inclination and longitude of the ascending node locate the orientation of the orbital plane holding the ellipse in relation to the reference plane, which is usually the ecliptic. The argument of the periapsis locates the orientation of the orbit's major axis in the orbital plane. Finally, the mean anomaly at epoch fixes the position of the planet in its orbit at one time. (If that position is known, the dynamics of gravitation theory can be used to determine its position at other times.)

[74] Our moon's motion is greatly perturbed, so that this approach is less successful for it and other methods are used in the historically troublesome lunar theory. See Brown (1896, p. 66).

gravitational influence of bodies other than the sun. Jupiter, the most massive of the planets, is a common source of perturbations. It exerts a large perturbing influence on Mercury for example. The axis of Mercury's orbit advances and the ellipse is more eccentric when Jupiter is in line with this axis. The axis regresses and is less eccentric when Jupiter is perpendicular to the axis.[75]

This need to give a physical foundation for the perturbed ellipses fitted is uncompromised. One might initially find that some perturbed ellipse fits the motion. However that fit remains tentative until the physical basis is located. Only then can the astronomers know how well the perturbed ellipse will continue to fit the planet's motion. More importantly, the perturbations to the ellipse can be adjusted according to the subsequent movements of the perturbing body.

Perhaps the most vivid illustration of the need for a physical basis for the changing elements of a planet's ellipse arises in the celebrated discovery of the planet Neptune. The need for the physical basis is inverted to become a means for discovery, in this case, of a new planet. By the early 19th century, the orbit and perturbations of the planet Uranus had been established. However not all of the perturbations could be explained by the gravitational action of known planets. In 1845, Adams and Le Verrier independently pursued the possibility of another hitherto unknown planet outside the orbit of Uranus that would be responsible for the perturbations. They predicted the position of this planet. After an easy telescopic search in 1846, the planet was found and was eventually given the name Neptune.

That astronomers *require* the variant curve forms to have a physical foundation is seen most clearly when these efforts fail. The orbit of Mercury was also well established in the nineteenth century and the bulk of its perturbations could be accounted for by the gravitational effects of the other planets. However they could not be accounted for completely. Recalling the success with Neptune, Leverrier (1859) proposed that these further perturbations could be accounted for by another new planet orbiting closer to the sun than Mercury. The new planet, which had come to be known as Vulcan, was never found.

That failure was discouraging. Nonetheless, astronomers could not abandon the idea that the perturbations were generated by some attractive mass somewhere. By the end of the century, many proposals were under investigation. Newcomb's (1895) treatment became the authoritative analysis. Its Chapter VI assesses a list of possible locations for new masses that might account for the anomalous motion of Mercury. They include masses located in a slight non-sphericity of the sun, in rings of masses or groups of planetoids inside Mercury's orbit, or planetoids between the orbits of Mercury and Venus and the possibility of a masses associated with the zodiacal light, a diffuse glow seen around the sun.

---

[75] Or so Airy (1884, p.113) reports.

More intriguing was a proposal by the astronomer Asaph Hall (1894). If the force of gravity does not dilute as the inverse square $1/r^2$ with distance r, but slightly faster, then the orbit of a planet would trace an ellipse that was advancing slightly, as Mercury's was observed to do. Hall noted that a very slight adjustment to the exponent in the inverse square was all that was needed to accommodate the anomalous motion of Mercury. He found that $1/r^{2.00000016}$ would suffice. Newcomb (1895, pp. 118-121) gave a more precise $1/r^{2.0000001574}$. None of these proposals survived Newcomb's and later astronomer's scrutiny.[76]

What is interesting for our purposes in Hall's hypothesis is that it altered the default repertoire of curves to be fitted to planetary motions. The one body problem no longer gives a fixed conic section as the simplest curve. Rather, under Hall's modified law of attraction, it gives very slowly rotating ellipses for bound orbits. These become the default curves to be fitted to planetary motions. The choice has a physical grounding in Hall's modified law of attraction.

While Hall's hypothesis did not survive scrutiny, that a law slightly different from Newton's prevails in the solar system soon proved to be the way to accommodate the anomalous motion of Mercury. In November 1915, an exhausted Einstein was putting the finishing touches onto his nascent general theory of relativity. He discovered to his jubilation that the new theory predicted precisely the anomalous advance of the perihelion of Mercury. He computed an advance of 43 seconds of arc per century, noting that the astronomers' values lay in the range of 40-50 seconds. With the adoption of Einstein's theory, it became automatic to include a relativistic correction to the Newtonian orbits; that is, under the physical grounding of Einstein's theory, the default curves to be fitted to planetary motions became very slowly rotating ellipses.

8. Harmonic Analysis of Tides

On an oceanic coast, the level of the seawater rises and falls periodically, with about two high tides and two low tides each day. Beyond this crude description are many variations. There is some variability in the timing of highs and lows; and there is considerable variation in just how high is a high and just how low is a low tide. This variability is also somewhat periodic over longer time scales, but the exact cycles are hard to pin down precisely merely by casual inspection of some portion of a tide's history.

Accurate tidal prediction is important and even essential for coastal navigation. Since the ebb and flow of the tide can produce considerable currents in coastal estuaries and bays, reliable advance knowledge of the tides can be the difference between easy and hard exits from one's

[76] The continuation of this episode, including Einstein's successful account of the motion of Mercury, is discussed further in the chapters on inference to the best explanation.

port. Reliable tidal prediction can make the difference between a successful return to one's home port or running aground in unexpected low water.

These factors make reliable long-term tidal prediction highly desirable. Since the behavior of the tides varies so much from place to place, the problem of prediction is best tackled as a curve fitting problem. Start with a good history of tides at each place on a coast. For each, find the curve that best fits the history and use it for prediction. Since the tides are periodic phenomena, one would expect that the families of functions to be fitted are based on the periodic trigonometric functions: sines and cosines of time. The natural method is straightforward Fourier analysis, which is the celebrated mathematical method for representing periodic functions in terms of sine and cosine harmonic constituents. To apply it, we would assume that the dependence of the water height on time t is given by the series:

$$a_0 + a_1\sin(t) + b_1\cos(t) + a_2\sin(2t) + b_2\cos(2t) + a_3\sin(3t) + b_3\cos(3t) + \ldots$$

with the series continued as far as needed. We know from the theory of Fourier series that any credible behavior of the tides over some nominated time can be represented arbitrarily closely by this expression. We merely need a suitable scaling for t, a suitable selection of the a and b coefficients and the inclusion of enough terms from the series.

While this is the obvious approach, in the past century and a half of work on tides, I have found no serious effort to provide this sort of analysis. The core difficulty, I conjecture, is that the dominant harmonic constituents present do not have the frequencies 1, 2, 3, … of the generic Fourier analysis. Combinations of these dominant constituents could still be captured by Fourier analysis with components of frequency 1, 2, 3, … However a large number of these components would be needed to represent accurately the summation of a few dominant harmonics whose frequencies are not in this set.

Instead, from the first moments, a physical basis has always been demanded for the harmonic constituents fitted to observed tidal histories. William Thomson (later Lord Kelvin) introduced the method of fitting harmonic constituents to tidal histories in 1867. Writing in his report to the British Association (Thomson, 1869), he noted that previous methods had merely recorded the times of high and low water. He proposed that fuller records be kept to which harmonic constituents would be fitted. The particular constituents he proposed were drawn directly from the background theory of the origin of the tides in the gravitational interaction of the earth, sun and moon.

The elements of this theory are widely known. The moon's gravity pulls on the waters of the earth's oceans. The pull is stronger than the average on the side of the earth nearer the moon and weaker on the side farther from the moon. The net effect is an elongation of the oceans into a spheroid that bulges away from the earth on both sides in line with the earth-moon axis, as shown in Figure 11.

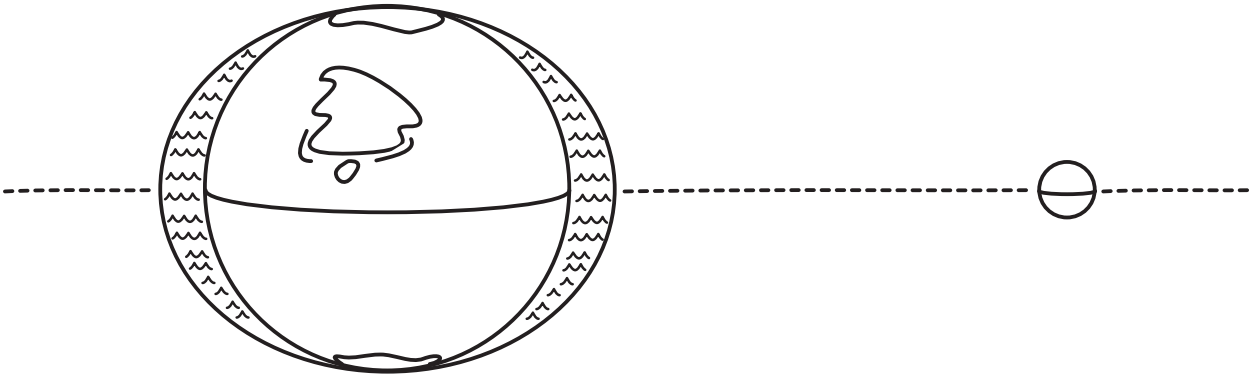


Figure 11. Tidal Bulge of Oceans Raised by the Moon.

The earth rotates daily under this bulged shape. As some location on the earth passes each bulge, that location registers a high tide. Since there are two bulges, each location registers two high tides and two low tides each day. The cycle is only roughly daily since it completes when a point on the earth returns to its original position in relation to the moon. The moon orbits the earth once a month and moves in the same direction as the earth rotates. So to return to its starting position in relation to the moon, the earth must rotate slightly more than the full rotation of 24 hours. It requires roughly 24 hours and 50 minutes. In this time, two tide cycles are completed. Half this process gives us the most important harmonic constituent: the "principal lunar semi-diurnal [=half-daily]," written as M_2, where the 2 denotes two cycles per day. It has period of about 12 hours and 25 minutes.

Superimposed on this semi-diurnal cycle is another semi-diurnal cycle. It results from the gravitational attraction of the sun on the waters of the oceans. The sun's attraction also distorts the ocean waters into a spheroid elongated along the line of the earth-sun axis, or so it would if there were not greater distortions due to the moon's gravity. The bulge produced would be a little less than half that raised by the moon. It takes 24 hours exactly for a point on the earth to return to a position with same relation to the sun. There are once again two bulges passed in this time, so we cycle between them each 12 hours. This contributes another harmonic constituent, the "principal solar semi-diurnal" S_2, whose period is 12 hours.

That these two harmonic constituents have periods differing by about 25 minutes is of the greatest consequence for the tides. At the full or new moon, when the sun and moon align, the two bulges add and we have especially high tides, known as the "spring" tides. They are so named since more waters are imagined as springing forth. The effect is shown in Figure 12.

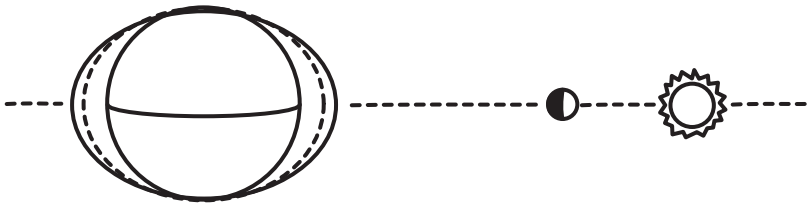


Figure 12. Spring Tides

Each 12 hours, the high water of lunar semi-diurnal cycle will lag behind that of the solar semi-diurnal cycle by about 25 minutes. This lag accumulates. After about a week, at the time of the half moon, the tidal bulges of the moon and sun are aligned roughly perpendicularly. The outcome is a lowering of the high tide and an elevation of the low tide, producing the more modest “neap” tides of Figure 13.

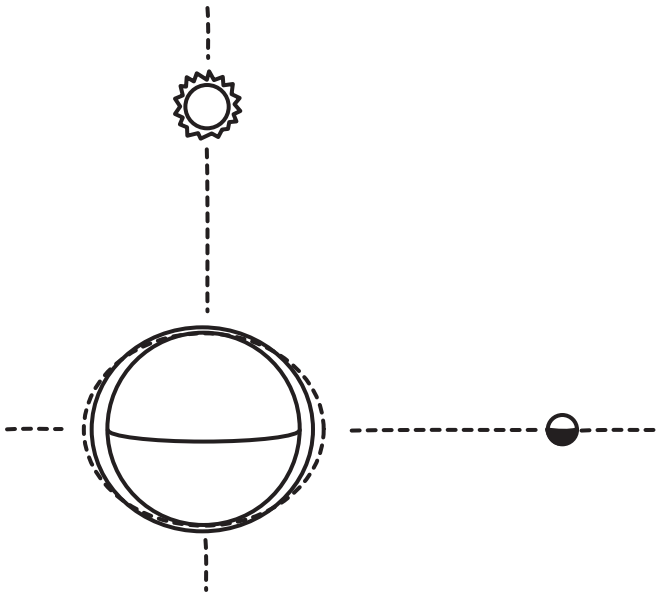


Figure 13. Neap tides

The combining of the two cycles to produce this further cycle of spring and neap tides is shown in Figure 14.

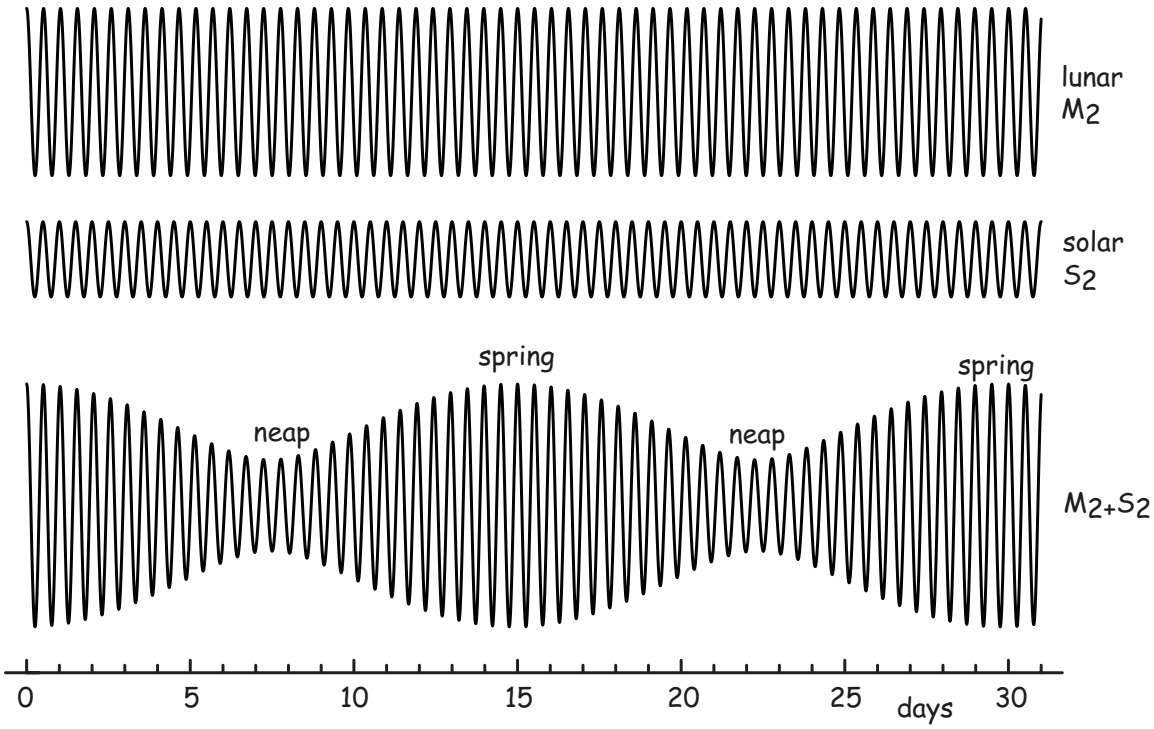


Figure 14. Spring and Neap Times from Combining Harmonic Constituents

This is one of the most important outcomes of the combining of the tidal harmonic constituents.

One might expect that there is little more to the harmonic analysis of the tides. What we have seen so far is adequate for an informal understanding of the gross behavior of the tides. It associates tidal levels with the phases of the moon in a qualitatively correct way. However it falls far short of what is needed for successful quantitative prediction of the tides. Many more physical factors must be considered.

The sun and moon also move to the North and South, carrying their tidal bulges with them. In the course of a solar year, the sun completes one cycle around the ecliptic, moving between 23.5º North and 23.5º South of the stellar equator. The moon's monthly motion carries it along a plane that is inclined at about 5º to the plane of the ecliptic. Its resultant motions carry it North and South of the stellar equator, between maximum elongations from the stellar equator that vary from 18.5º to 28.5º. As the sun and moon change their longitude, they carry with them the tidal bulges that they raise. This affects the heights of the tides and does it differently at each location on the earth.

If the sun and moon are directly over the equator, the two tidal bulges will pass symmetrically over some fixed terrestrial location on the equator in the course of a day. In so far as these processes are concerned, successive tides will have equal height. If, however, the sun and moon have moved together to a position far to the North, then the two tidal bulges will be shifted towards different hemispheres. One will be massed in the Northern hemisphere and the

other in the Southern hemisphere. As a result, a point on the earth away from the equator will meet with deeper and shallower portions of the successive bulges, adding a diurnal (daily) cycle to the semi-diurnal cycles so far mentioned. In extreme cases, the sun and moon pass overhead sufficiently far from the equator that, at some locations, one bulge may be missed entirely. These locations experience a single high tide per day. That is, their tides are on a diurnal cycle.

There are further complications. The size of the tidal bulge raised depends on the distance from the earth to both the sun and moon. Since the orbits are elliptical, the sun and moon approach and recede from the earth as they complete their cycles, annually and monthly, respectively. In addition, this cyclic effect is compounded by the perturbations induced by the sun on the moon's orbit. Those perturbations alter the eccentricity of the moon's ellipse introducing further variation in the distance of the moon from the earth. All these astronomical effects happen on regular cycles, readily predictable in advance. They are incorporated into tidal analysis by adding more harmonic constituents of the appropriate form.

These astronomical effects may seem overwhelming. However they are merely the most reliably regular of the influences on the tides. If the earth were a perfectly smooth spheroid, a tidal bulge of the ocean would wash over it as a uniform wave. However the earth is not a perfectly smooth body and all sorts of irregularities in its surface obstruct the uniform passing of the tidal wave. These obstructions are great in coastal areas, which is precisely where we are seeking predictions. The problems are even worse if we wish to predict tides in bays and estuaries. For the rising and falling of the tide will be delayed by the need for water to flow in and out of the bay as the tidal wave passes. Enclosed bodies of water have their own natural frequencies with which water oscillates to and fro within them. The coming and going of tidal waves couples with these oscillatory process. All these processes are represented by further harmonic constituents.

The shallow-water constituents are of two types: overtides and compound tides. The first are the analog of harmonic overtones in music. For example, the principal lunar semi-diurnal M_2 consists of 2 high tides per day. It raises shallow-water overtides M_4 and M_6 with four and six peaks each day. Compound tides arise with a frequency that is the sum or difference of the components from which they are derived. The shallow-water terdiurnal MK_3 is derived from the principal lunar semi-diurnal M_2 and the lunar diurnal K_1. It sums their two and one peaks per day to give three peaks.

Finally, meteorological facts can have a major influence on tides. Strong winds can materially affect them. These factors, however, are the hardest to address. Accurate weather prediction is difficult even a day in advance, whereas tide tables are prepared years in advance. There is some small effort to allow for these meteorological effects by means of the solar

components, S_a, S_{sa} and S_1; that is, the solar annual, solar semi-annual and solar diurnal, which have periods of a year, half a year and a day.[77]

In sum, the harmonic analysis of tides is complicated and difficult, even when we seek a sound physical basis for the harmonic constituents. Many are needed. This was already apparent to Thomson (1869, p. 491), who initially listed 23 constituents. Many more can be needed. The most difficult locations for prediction are complex estuaries, such as Anchorage, Alaska, and Philadelphia, Pennsylvania. An adequate analysis requires over 100 harmonic constituents. (Hicks, 2006 , p.40.) The United States National Oceanic and Atmospheric Administration (NOAA) employs a standard set of 37 constituents for its tidal predictions for coastal regions in the US. Here is an illustration of their use.

The table shows the harmonic constituents used by NOAA for Annapolis, Maryland, in the Chesapeake Bay:[78]

	Constituent Symbol	Constituent Name	Amplitude	Phase	Speed
1	M2	Principal lunar semidiurnal	0.457	291.6	28.9841042
2	S2	Principal solar semidiurnal	0.071	319.5	30
3	N2	Larger lunar elliptic semidiurnal	0.095	270.5	28.4397295
4	K1	Lunar diurnal	0.194	356.7	15.0410686
5	M4	Shallow water overtides of principal lunar	0.012	58.3	57.9682084
6	O1	Lunar diurnal	0.157	6	13.9430356
7	M6	Shallow water overtides of principal lunar	0.011	159.6	86.9523127
8	MK3	Shallow water terdiurnal	0	0	44.0251729
9	S4	Shallow water overtides of principal solar	0	0	60
10	MN4	Shallow water quarter diurnal	0	0	57.4238337
11	NU2	Larger lunar evectional	0.021	268.5	28.5125831
12	S6	Shallow water overtides of principal solar	0	0	90
13	MU2	Variational	0	0	27.9682084
14	2N2	Lunar elliptical semidiurnal second-order	0.013	246.7	27.8953548
15	OO1	Lunar diurnal	0.006	347.3	16.1391017
16	LAM2	Smaller lunar evectional	0.011	318	29.4556253
17	S1	Solar diurnal	0.065	290.5	15
18	M1	Smaller lunar elliptic diurnal	0.011	1.2	14.4966939

[77] For further discussion of these harmonic components, see Shureman (1958, pp. 39-48).

[78] Amplitude is measured in feet, phase in degrees and speed in degrees per hour. Source: http://tidesandcurrents.noaa.gov/data_menu.shtml?stn=8575512%20Annapolis,%20MD&type=Harmonic%20Constituents accessed August 10, 2012.

19	J1	Smaller lunar elliptic diurnal	0.011	340.9	15.5854433
20	MM	Lunar monthly	0	0	0.5443747
21	SSA	Solar semiannual	0.119	44.5	0.0821373
22	SA	Solar annual	0.338	128.4	0.0410686
23	MSF	Lunisolar synodic fortnightly	0	0	1.0158958
24	MF	Lunisolar fortnightly	0	0	1.0980331
25	RHO	Larger lunar evectional diurnal	0.012	29	13.4715145
26	Q1	Larger lunar elliptic diurnal	0.025	331.6	13.3986609
27	T2	Larger solar elliptic	0.004	318.3	29.9589333
28	R2	Smaller solar elliptic	0.001	320.6	30.0410667
29	2Q1	Larger elliptic diurnal	0.004	15.1	12.8542862
30	P1	Solar diurnal	0.065	348.8	14.9589314
31	2SM2	Shallow water semidiurnal	0	0	31.0158958
32	M3	Lunar terdiurnal	0	0	43.4761563
33	L2	Smaller lunar elliptic semidiurnal	0.033	308.1	29.5284789
34	2MK3	Shallow water terdiurnal	0	0	42.9271398
35	K2	Lunisolar semidiurnal	0.021	317.9	30.0821373
36	M8	Shallow water eighth diurnal	0	0	115.9364166
37	MS4	Shallow water quarter diurnal	0	0	58.9841042

Table 1. Harmonic Constituents used by NOAA for Tidal Predictions at Annapolis, Maryland.

These thirty seven constituents fix the family of thirty seven component functions whose sum is to be fitted to the tidal history in Annapolis. Each consists of a cosine wave whose amplitude, phase and speed are to be determined either from background assumptions or by fitting to the tidal history. The resulting parameters, given in the last three columns of the table, are used to compute NOAA's tidal prediction. Figure 15 shows the result of combining them for the week of August 7, 2014.[79]

[79] The height predicted is above MLLW = mean lower low water. Source: http://tidesandcurrents.noaa.gov/noaatidepredictions/NOAATidesFacade.jsp?timeZone=2&dataUnits=1&datum=MLLW&timeUnits=2&interval=6&Threshold=greaterthanequal&thresholdvalue=&format=Submit&Stationid=8575512&&bmon=08&bday=07&byear=2014&edate=&timelength=weekly accessed August 10, 2012.

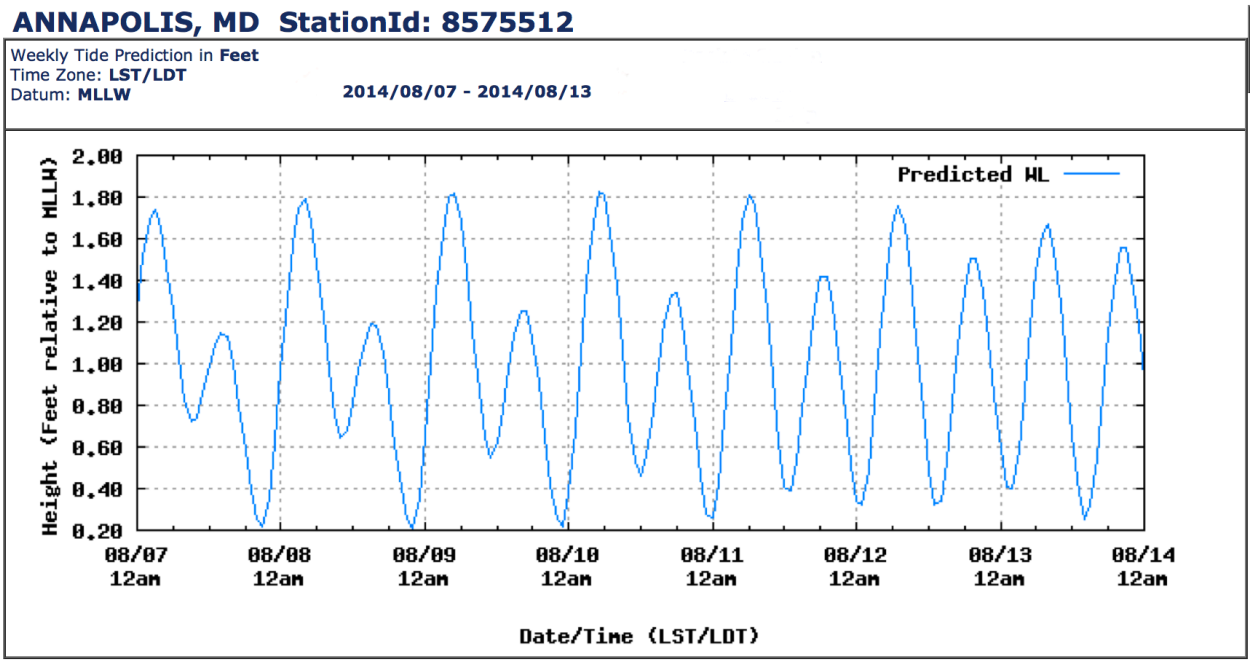


Figure 15. Tidal Prediction for Annapolis, August 7-13, 2014

References

Airy, George B. (1884) *Gravitation: An Elementary Explanation of the Principal Perturbations in the Solar System.* 2nd. Ed. London: MacMillan and Co.

Aquinas, Thomas (1945) *Basic Writings of Saint Thomas Aquinas*. Ed. Anton C. Pegis. Random House; Reprinted Indianapolis, IN: Hackett, 1997.

Brown, Ernest W. (1896) *An Introductory Treatise on the Lunar Theory*. Cambridge: Cambridge University Press.

Clark, Ronald W. (1984) *Einstein: the Life and Times.* New York: Avon.

Einstein, Albert (1933), "On the Methods of Theoretical Physics," in *Ideas and Opinions*. (New York: Bonanza, 1954), 270-276.

Einstein, Albert (1949), "Autobiographical Notes," in P. A. Schilpp, , ed., *Albert Einstein: Philosopher-Scientist*. Evanston, IL: Library of Living Philosophers.

Hall, Asaph (1894) "A Suggestion in the Theory of Mercury," *The Astronomical Journal.* **14**, pp. 49-51.

Hicks, Steacy D. (2006) "Understanding Tides," Center for Operational Oceanographic Products and Services, National Oceanic and Atmospheric Administration. http://www.co-ops.nos.noaa.gov/publications/Understanding_Tides_by_Steacy_finalFINAL11_30.pdf

Kelly, Kevin (2007) “A New Solution to the Puzzle of Simplicity,” *Philosophy of Science,* **74**, pp. 561-73.

Le Verrier, Urbain (1859), , "Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète", *Comptes rendus hebdomadaires des séances de l'Académie des sciences (Paris)*, **49**, pp. 379–383.

Lorentz, Hendrik A. (1920) *The Einstein Theory of Relativity: A Concise Statement.* 3rd ed. New York: Brentano’s.

Mach, Ernst (1898) “The Economical Nature of Physical Inquiry,” pp. 186-213 in *Popular Scientific Lectures*. Trans. T. J. McCormack. 3rd ed. Chicago: Open Court.

Maurer, Armand A. (1999) *The Philosophy of Ockham in the Light of its Principles.* Toronto, Canada: Pontifical Institute of Mediaeval Studies.

Misner, Charles W.; Thorne, Kip S.; Wheeler, John Archibald (1973), *Gravitation*. San Francisco: W. H. Freeman

Moody, Lewis. F. (1944), "Friction factors for pipe flow", *Transactions of the American Society of Mechanical Engineers*, **66** (8), pp. 671–684.

Newcomb, Simon (1895) *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy.* Washington: Government Printing Office.

Newton, Isaac (1726), *Mathematical Principles of Natural Philosophy*. 3rd ed. Trans. Andrew Motte, rev. Florian Cajori. University of California Press, 1962.

Norton, John D. (2000), “ ‘Nature in the Realization of the Simplest Conceivable Mathematical Ideas’: Einstein and the Canon of Mathematical Simplicity,” *Studies in the History and Philosophy of Modern Physics*, 31, 135-170.

Popper, Karl (1968) *Logic of Scientific Discovery*. New York: Harper and Row.

Shureman, Paul (1958) *Manual of Harmonic Analysis of Tides*. Washington: United States Printing Office.

Sober, Elliott (1988), “The Philosophical Problem of Simplicity,” CH. 2 in *Reconstructing the Past: Parsimony, Evolution, and Inference.* Cambridge, MA: Bradford, MIT Press.

Sterne, Theodore E. (1960) *An Introduction to Celestial Mechanics*. New York: Interscience Publishers.

Thomson, William (1869) “Committee for the purpose of promoting the extension, improvement, and harmonic analysis of Tidal Observations,” *Report of the Thirty-Eighth Meeting of the British Association for the Advancement of Science, Norwich August 1868.* London: John Murray, Albemarle St. pp. 489-510.

Thorburn, W. M. (1918) “The Myth of Ockahm’s Razor,” *Mind*, **27**, pp. 345-53.

Watson, James C. (1861) *A Popular Treatise on Comets*. Philadelphia: James Challen & son.

June 29, 2012
Minor formatting December 15, 2017; typos June 17, 2018

Chapter 7

Simplicity in Model Selection

1. Introduction

In philosophical analyses, simplicity is most commonly introduced as a rather abstruse metaphysical notion whose application in theory appraisal is important but troublesome. For the invocation of simplicity seems to require the highest level of human insight, as opposed to the mechanical application of an unambiguous, even algorithmic rule. Hence it was quite a revelation in the philosophy of science literature when Forster and Sober (1994) pointed out that the model selection literature in statistics had succeeded in incorporating a simplicity condition into rules for model selection that are applied mechanically, that is, without the need for higher level human insight.

This example of model selection is important and interesting. However, my sense is that Forster and Sober were too optimistic in just what they thought we can learn from it. They passed too readily from the case of model selection to broader morals pertaining to other cases in which there are invocations of simplicity, such as the decision between Copernican and Ptolemaic astronomy. That is an overreach. The model selection literature shows how simplicity considerations arise in solving a quite specific problem: the discerning of the true relation obscured by random, statistical noise. The simplicity considerations in Copernican and Ptolemaic astronomy are not dependent essentially on error noise. There is a loose similarity between the two cases, but much more needs to be said before general morals can be recovered from the one case of model selection.

My goal in this chapter is more modest. I seek to recover no universal claims about simplicity from this example. Rather I merely want to show how the literature on model selection provides an important illustration of the central claim of the last chapter: that there is no epistemically potent, universal principle of parsimony and that simplicity considerations in theory appraisal are really surrogates for background facts. I will look at hypothesis selection governed by the Akaike Information Criterion, discussed by Forster and Sober. The criterion directs us to evaluate an hypothesis by determining how likely it makes the data at hand. The danger of overfitting is greater, the larger the hypothesis space of the model from which the

hypothesis is drawn. The criterion directs that we correct for this overfitting merely by subtracting the dimension of the hypothesis space from the statistic that expresses the likelihood of the data. This correction is its notable property for it rewards models for their simplicity. However, I will argue, the criterion provides no comfort for metaphysicians of simplicity since:

- The criterion is deduced from straightforward assumptions about the systems investigated. These assumptions include no posit of simplicity and no principle of the parsimony of nature.
- The criterion deduced is simply a formula used to weight the performance of various models in narrowly specified condition. No general principle of parsimony is inferred such as could be applied elsewhere.
- Considerations of simplicity need not enter into the discussion at all. They arise only because we metaphysically-minded readers see a particular formula and find it comfortable to interpret one term in it as a reward for simplicity (or punishment for being complicated).

Finally we shall see that the simplicity correction is merely a surrogate for a correction derived from a background assumption. The most potent of the governing assumptions is that the data are generated by an hypothesis in the model under test.[80] That assumption proves strong enough to allow us to estimate how much overfitting the model permits and, as a result, to correct for it in an especially simple way. We then interpret this correction as what simplicity requires, although that notion played no role in its generation.

The chapter will introduce model selection and the Akaike Information Criterion. It is merely one of many such criteria. For our purposes of identifying how generally simplicity considerations enter model selection, it is as good as any.[81] The early part of the chapter will introduce model selection and try to explain how the criterion is able to generate the simplicity correction. The chapter will then turn to a fully worked out example of the criterion in action and conclude with an account of its relation to the material theory of induction.

---

[80] For a good account of the Akaike Information Criterion, see Konishi and Kitagawa (2008, ch. 3) and especially their Section 3.3 for an account of additional terms needed if the truth is not assumed to be one of the hypotheses under test.

[81] There is, for example, an extended version of the Akaike criterion modified to correct for small data sets and large numbers of parameters. (Burnham and Anderson, 2004). Other related criteria include the Bayes Information Criterion ("BIC"), which arises in a Bayesian analysis of model selection (Wasserman, 2000).

2. Model Selection

Model selection deals with data generated by some probabilistic system. A model consists of a set of hypotheses such that each is a candidate description of the probabilistic system. A primary application is the example of curve fitting discussed in the last chapter, in which data is generated by a function confounded by statistical noise. The models are the different families of functions that may be fitted: linear functions, quadratic function, and so on and their associated error distributions. However the methods can deal with more general cases and can be applied whenever data are generated probabilistically. If, for example, one samples the heights, weights, genders and so on of a population, the resulting data are generated by a probability distribution that covers these features of the population. In this case, the models are sets of possible distributions and the parameters sought are means, variance, covariances and other parameters of the distributions.

The model selection literature seeks ways to see past the statistical noise in the data to the true system that generated it. For any particular data set, one can always find a better fitting model by sacrificing simplicity. The more complicated model fits better since it can conform to the confounding statistical noise. The larger the model, that is, the more hypotheses it contains, the greater its ability to conform to the data and the greater the danger of overfitting. The remedy is to forgo some goodness of fit in favor of a simpler model.

A crude illustration is the problem of identifying the daily arrival times of a bus. We may find the bus to arrive at 11:58, 12:04 and 12:02 on successive days. These data are accommodated well enough by the hypothesis that the bus arrives roughly at 12:00. However if we allow more complicated descriptions, we can find a hypothesis that fits the data perfectly. We might propose that the bus arrival times cycle successively through 11:58, 12:04 and 12:02, thereby eliminating any mismatch between our hypothesis and the data at hand. Informally, we would judge the improvement in goodness of fit to be spurious, a result of overfitting, and revert to the "roughly 12:00 arrival" hypothesis as simpler.

3. Maximum Likelihood Criterion

The Akaike Information Criterion "AIC" (Akaike, 1974) is an elaboration of another simpler criterion, the maximum likelihood criterion. Assume we have some probabilistic system that produces data and we wish to infer back to the properties of the system. We identify those properties through the parameters characteristic of the system. These would be the coefficients in the functions we fit to the data in curve fitting; or they might be means and variances if we are trying to find the population parameters from the data of a population sample. To start, we

presume some model, that is, some set of hypotheses indexed by the sorts of parameters we believe characteristic of the system. In curve fitting, it would be, say, a linear or quadratic curve confounded by error noise. Different parameters in the model pick out different hypotheses that will make the data actually recovered more or less probable. That conditional probability is called the likelihood L:

$$L = P(\text{data} \mid \text{model parameters})$$

Which parameters should we choose? An obvious choice is those parameters that make the data most probable; that is, we choose to maximize the likelihood L and the resulting parameters are know as "maximum likelihood estimators." It turns out to be convenient not to work with the likelihood L directly but with its logarithm, log L. Since the logarithm function is strictly increasing, maximizing L is equivalent to maximizing log L. And maximizing log L is equivalent to minimizing –log L. This gives us:

Maximum Likelihood Criterion:
Seek the parameters that maximize log L,
that is, that minimize –log L

This criterion works well until we try to use it to compare models with different numbers of parameters. You might expect that we can compare two models by looking at the maximum log-likelihood each supplies. What if best fitting hypothesis H of model M_1 yields a higher log likelihood of the data than does best fitting hypothesis K of model M_2? It would seem straightforward that we should pick the H of model M_1 over the K of model M_2.

That straightforward conclusion is too hasty because the log likelihood delivered by one model can be spuriously inflated by overfitting. For example, in curve fitting, if we use a model with linear functions $y=A+Bx$, we fit just two parameters, A and B, as well as any parameters characterizing the error noise distribution. If we move to a model with quintic equations $y=A+Bx+Cx^2+ Dx^3+Ex^4+Fx^5$, these two parameters are replaced by six parameters, A, B, C, D, E and F. The larger number of parameters in the second model gives it more flexibility and that gives it an unfair advantage over the first model. The data is generated probabilistically and, as a result, will not perfectly reflect the probabilistic system that generated it. A sample mean will typically differ slightly from a population mean. A maximum likelihood estimator can increase the likelihood of the data by tracking these slight deviations. Selecting the sample mean as the estimator for the population mean will render this particular data set more probable than selecting the true population mean. This unwanted effect is overfitting, once again. As the number of parameters in the model grow, the model becomes more flexible and the extent of overfitting increases.

4. Akaiki Information Criterion

How can we guard against overfitting? Qualitatively, we might seek to protect ourselves by favoring simpler models, that is, models with fewer parameters. That solution is correct at the level of vague generality, but it does not translate into a quantitative procedure with a precise justification that tells us just when to abandon the models with more parameters.

Akaike approached the problem by considering not just performance with the particular data at hand. Instead he asked that we choose estimators that perform well on average over all the data sets that might be produced by the probabilistic system. The motivation is that overfitting produces estimators that work well for one data set to which they are tuned, but will generally fare worse for others that the probabilistic system may produce. A model with a larger set of parameters is more flexible and thus more likely to be overfitted to the data. So, if we seek models that perform well on average, we must penalize the performance of models with larger numbers of parameters to compensate for the inflation in their performance due to overfitting. What Akaike found was that the requirement of best performance on average over all data sets led to a remarkably simple correction to the Maximum Likelihood Criterion. That is, he found that overfitting inflates the log likelihood of the data by the dimension d of the parameter space. We correct the log likelihood function for overfitting merely by subtracting this dimension d from it. What results is:

Akaike Information Criterion (AIC):
Seek the parameters that maximize log L –d;
that is, they minimize[82] –log L + d.

The penalizing factor d automatically favors models with lower numbers of parameters. It expresses in quantitative form the qualitative notion that we should favor the simpler over the more complicated model.

4.1 How it Works: The Essential Assumption

The criterion works by asking not merely how well the estimator performs with the particular data set at hand. Rather it asks how the estimator performs on average with all possible data sets and rewards and penalizes the various models accordingly. For example, if we suspect a population is exactly 50% female, we would not be surprised to find that there are 57 females in a random sample of 100 from the population. We might be tempted by this datum to posit that 57% of population overall is female. The posit would make the datum of 57 females in the

[82] Akaike's original proposal was to minimize - 2log L + 2d, but I have dropped the factor of two since it confounds the simplicity of the formula without any gain.

sample more probable than the supposition that 50% are female. However, we would likely hesitate. In *this* sample, we might allow, we found 57 females. But what might happen if we draw another random sample of 100; and another; and another? Over the repeated samplings, if the 50% hypothesis is correct, we would find a range of sample results scattered around 50 females. The hypothesis of 57% would perform poorly over this range and, on average, the true hypothesis of 50% female would perform best.

The Akaike Information Criterion arises when we correct the performance of an estimator for how it is likely to perform on average over all possible data sets. The great difficulty with this correction is that we do not know the full properties of the true probabilistic system, so, it would seem, we cannot know what all possible data sets are. It is true that we cannot know this without further assumption. We must assume something more. Otherwise the analysis would be performing impossible magic.

The key assumption of the analysis is that the *true probabilistic system lies within the model under consideration*, where a model is simply some collection of hypotheses.[83] So if we are fitting a linear curve $y = A + Bx$ to data, then we assume that some values of A and B are the true values of the system. The remarkable thing about Akaike's analysis is that this assumption is sufficient to allow the analysis to proceed. We do not need to know which values of A and B are the true values. We merely need to assume that there are some values of A and B that coincide with the truth.

What results is a correction to the Maximum Likelihood Condition of impressive simplicity. That simplicity comes at some cost, for it arises only after we have made strong assumptions about the background system and our sampling of it. In addition to the assumption noted above, we also assume that the data set is sufficiently large for the central limit theorem of statistics to be applicable. Nonetheless, it is striking that such a simple correction formula is possible under any conditions. The penalizing factor d merely records the dimension of the space of parameters. The two parameters A and B of linear functions provide two dimensions; the six parameters A, B, C, D, E and F of quintic functions provide six parameters. Nothing else in the details of the space matters.

4.2 Kullback-Leibler Discrepancy, Predictive Accuracy and the Truth

This discussion has been kept as simple as possible, so a technical note is required, for those who want it. This characterization of how the Akaike Information Criterion works will at

[83] This is an awkwardness of the application of AIC. This assumption can fail for at least some of the models we may compare. It must fail, for example for all but one, when we compare models with disjoint sets of hypotheses.

first seem different from the way the criterion is normally motivated. Akaike (1974) and later authors (e.g. Zucchini, 2000; Konishi and Kitagawa, 2008, ch.3) employ what is variously called the Kullback-Leibler discrepancy or the Kullback-Leibler information. In seeking to identify a probabilistic system, we seek to identify the probability the system assigns to each possible outcome datum $\mathbf{x}$, where the datum $\mathbf{x}$ is a vector since it will, in general, consist of several numbers. That true but unknown probability is labeled as the probability density $g(\mathbf{x})$. The models we fit are also probability densities over the same space of possible outcomes, $f(\mathbf{x}|\boldsymbol{\theta})$, where the vector valued $\boldsymbol{\theta}$ is the set of parameters characterizing the model. The Kullback-Leibler discrepancy is

$$I(g;f) = \int_{\text{all } \mathbf{x}} g(\mathbf{x})[\log g(\mathbf{x}) - \log f(\mathbf{x}|\boldsymbol{\theta})]\, dx$$

It measures how closely the model $f(\mathbf{x}|\boldsymbol{\theta})$ comes to the target $g(\mathbf{x})$. It achieves its minimum value of 0 when $g(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})$ almost everywhere. The goal is to find that $f(\mathbf{x}|\boldsymbol{\theta})$ that achieves this minimum value. Since the target $g(\mathbf{x})$ is fixed, this goal is equivalent to maximizing the integral

$$\int_{\text{all } \mathbf{x}} g(\mathbf{x}) \log f(\mathbf{x}|\boldsymbol{\theta})\, dx$$

This integral computes a measure of average performance. The term $\log f(\mathbf{x}|\boldsymbol{\theta})$ is the log-likelihood of some particular datum $\mathbf{x}$. The density $g(\mathbf{x})$ tells us how frequently that datum will appear in repetitions of whatever procedure or experiment generates the data. So the integral is the average log likehood of a datum over many repetitions. Selecting a parameter $\boldsymbol{\theta}$ that maximizes the integral identifies that density $f(\mathbf{x}|\boldsymbol{\theta})$ that will have the best performance on average in the sense that it renders the data we expect in multiple repetitions most probable.

The $f(\mathbf{x}|\boldsymbol{\theta})$ that is selected by this performance criterion is commonly described as selecting the probability density that has the best "predictive accuracy." In general, it will not be the distribution that makes the data at hand most probable. That distribution may have been eliminated by a penalty for a larger number of parameters. However the one selected will have the property of making the accumulated data most probable over very many repetitions of the procedure. Since these procedures have yet to happen, this feature is labeled "predictive accuracy."

While predictive accuracy is desirable, it is less than the goal of finding the truth. False theories can enjoy considerable predictive accuracy. The Demeter-Persephone myth of ancient Greece successfully predicts endless repetitions of fertile and barren seasons. Also some model selection problems may preclude prediction. At an archaeological site, we may collect and map the positions of bone fragments. We want to know if their spatial distribution has one or two

peaks, which would correspond to one or two sources. In this problem, we are indifferent to prediction, since there are no further bone fragment locations to be predicted. All we really want is the true distribution.

In the particular case of the Akaike Information Criterion, we can see that the maximization is a condition that will return the true probability distribution to us. For the Akaike Information Criterion proceeds from the assumption that the true distribution $g(\mathbf{x})$ coincides with one of the distributions in the model. That is,

$$g(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}_0)$$

for $\boldsymbol{\theta}_0$ the true parameter value. Then we seek to optimize the integral

$$\int_{\text{all } \mathbf{x}} f(\mathbf{x}|\boldsymbol{\theta}_0) \log f(\mathbf{x}|\boldsymbol{\theta}) \, dx$$

and this integral achieves its maximum value when we set $f(\mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}_0)$.[84]

The common justification of the Akaike Information Criterion is that it selected the probability distribution that has greatest predictive accuracy. We can now see that this undersells the criterion. It is designed to seek the true probability distribution. Its justification should be given in terms of truth not predictive accuracy.

5. How It Works: An Oversimplified Analogy

That the Akaike Information Criterion can correct for overfitting may seem mysterious and even magical. It is not so. The correction results from implementing a prosaic standard: seek the best performance over all data on average. The correction does not explicitly set out to reward simplicity. That is does so is merely a consequence of the analysis. A greatly oversimplified analogy shows that this sort of correction is far from mysterious.

In this analogy, we will consider the near trivial problem of fitting linear, quadratic, cubic, … curves to data *without error*. That is, we require that the fitted curve must pass through all the data points without error. We seek a criterion that directs us to the unique curve appropriate to the data. We might initially choose the scoring criterion

Number of hits

That is not a good criterion. If we have three data points for (x,y): {(0,0), (1,1), (2,2)}, then the straight line y = x scores three hits. But so do many cubic curves (as shown in Figure 1) and so do many more quartics.

---

[84] This follows since the Kullback-Leibler discrepancy I(g:f) has its minimum value of zero when $g(\mathbf{x}) = f(\mathbf{x})$ almost everywhere.

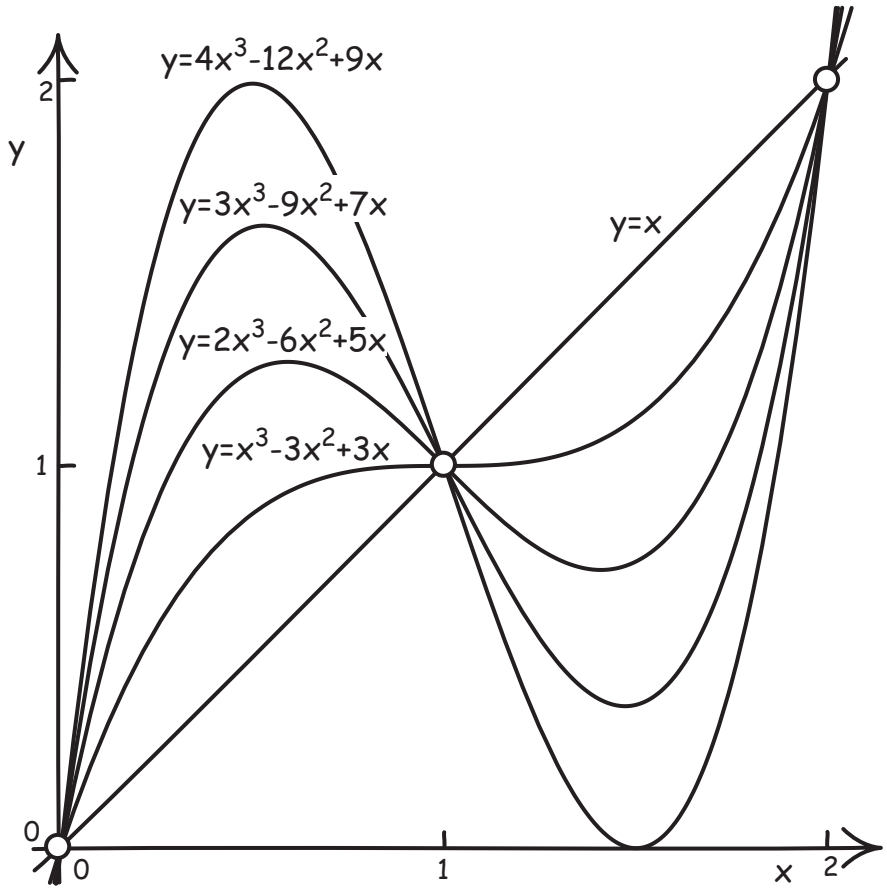


Figure 1. Linear and cubic curve fits

They score equally—3 hits—but they are not equally successful. We discount the cubic and quartic curves since they are not uniquely selected. Cubic curves $y = A + Bx + Cx^2 + Dx^3$ have 4 free parameters and thus many cubic curves can hit just three data points, but there is only one that can hit four. Quartic curves have five free parameters. Many can hit can three data points, but only one can hit five.

If our interest is uniqueness, instead of counting the number of hits, we should assess whether the number of hits are sufficient to ensure a unique curve. That leads to the new score:

Score = Number of hits – Number of parameters

We have uniqueness if this Score is greater than or equal to zero. For each of the d parameter families of curves mentioned above return a unique curve only when it has a curve that hits d or more points.

This new Score discriminates the linear model from the others in the above case. The linear curve has a Score of 3-2=1, the cubic 3-4=-1 and the quartic 3-5=-1. Only the linear curve has a Score greater than or equal to zero.

The example is elementary, but it manifests two features of model selection methods:

- The score was not derived from a metaphysics of simplicity that demands that more complicated models must be penalized for their lack of simplicity.

Rather all models were held to the same standard: the scoring rewards them only when they produce a unique curve. The result of this requirement was an automatic penalizing of the more complicated models.

- The success of the scoring system depends on background assumptions.

In this case, the curve scoring zero or more is assured to be unique only if the true curve lies in the same model. In the example, if the true curve were actually in the cubic model, then the uniqueness of the straight line y=x for the linear model would be insufficient to assure us that we have found the unique curve. Since we have only three data points, it could be any of many curves in the cubic model.

6. A Coin Tossing Illustration of the Akaike Information Criterion

That the simple correction of the Akaike Information Criterion suffices does seem too good to be true. That it does suffice, under the right conditions, is found merely by working through the statistical analysis that leads to the result. Since this analysis is quite difficult, I have provided a simple application of AIC here and in the Appendix that is designed to display the full analysis and show how it is that a correction merely in the dimension of the parameter space d can be deduced from the requirement of maximizing average performance.

The example pertains to tossing coins. Let us say that we toss N coins and find n heads. What is the chance p of a single toss coming up heads? Our estimation problem is to find that chance. Let us consider models with differing numbers of parameters. Each model assumes independence of the tosses.

6.1 Zero-Parameter Model

The simplest model just posits that our best estimate of p, $\hat{p}$, is 1/2. It is a rather inflexible model since it allows only one value, but just that is what makes it a zero parameter model. The likelihood L of n heads in N tosses in this model is

$$L_0(1/2) = (1/2)^n (1-1/2)^{n-N} = 1/2^N.$$

So we have the log likelihood $\log L_0(1/2) = N \log (1/2)$. AIC directs us to maximize

$$L_0(1/2) = N \log (1/2)$$

where no dimensional correction is applied since d = 0.

6.2 One Parameter Model and its Problems

The next simplest model has one parameter, p, which is the chance of a head. The log likelihood of n heads in N tosses is

$$\log L_1(p) = \log (p)^n (1-p)^{n-N} = n \log p + (N-n) \log (1-p)$$

and (as is shown in the Appendix), the value of p that maximizes the log likelihood is

$$\hat{p} = n/N.$$

This model already admits a small amount of overfitting. If, for example, the true value of p is 0.5 = 1/2 and we have N=100 tosses, then n is less likely to be 50 exactly. Rather it will be somewhere in the neighborhood of 50, say n=42 or n=55. Choosing $\hat{p} = 0.42$ or 0.55 in these two cases will produce log likelihoods that exceed the log likelihood returned by the zero parameter model, even though, in this case, our supposition is that the zero parameter model happened to have hit upon the true value of p.

Here are the values. The zero parameter model yields

$$\log L_0(1/2) = 100 \log (1/2) = -69.31$$

The one parameter estimators do better when employed with the data sets to which they are tuned:

$$\text{For } n=42, \log L_1(.42) = 42 \log (.42) + 58 \log(.58) = -68.03$$

$$\text{For } n=55, \log L_1(.55) = 55 \log (.55) + 45 \log(.45) = -68.81$$

The one parameter estimators yield greater (i.e. less negative) log likelihoods that does the presumed true zero parameter estimator.

The estimators $\hat{p} = 0.43$ or .55 have performed better in these two cases of n=43 or n=55 since they have been tuned specifically to these two cases respectively. They each perform worse than the zero parameter model, however, if we reverse cases and use $\hat{p} = 0.42$ for the case of n=55 and use $\hat{p} = 0.55$ for the case of n=42.

$$\text{For } n=55, \log L_1(.42) = 55 \log (.42) + 45 \log(.58) = -72.23$$

$$\text{For } n=42, \log L_1(.55) = 42 \log (.55) + 58 \log(.45) = -72.73$$

That is, successes of $\hat{p} = 0.43$ or .55 are inflated by overfitting to the specific data at hand. They will perform worse if we employ them with other data sets to which they are not tuned.

6.3 One Parameter Model Repaired

These effects indicate how we can correct our assessments for overfitting. We give up the goal of merely maximizing log likelihood for the data at hand. Instead we seek to optimize the log likelihood over all possible data sets, appropriately weighting each set for its probability. Finding the estimators that perform best by this standard is basis of the Akaike Information Criterion. This fundamental idea is important enough to bear restatement:

Seek the estimator that gives the best log likelihood when averaged over all possible data sets.

To proceed, we need to know which are all possible data sets. For that we assume:

There is a single true chance of a head, p^*, within the hypotheses of the one parameter model.

As I noted above, this is the non-trivial assumption of the analysis, for it says that the truth lies somewhere within our present one parameter space of hypotheses.[85] Our calculations are also greatly simplified with the assumption that the number of tosses N in each data set is very large. That means the central limit theorem of statistics can be called up to assure us that the number of heads n is normally distributed around a mean of p^*N with a variance $N\, p^*(1- p^*)$.

Let us fix some particular maximum likelihood estimator $\hat{p} = \pi$ that is derived from one data set. We can ask how the log likelihood of that particular value π will fare over all possible data sets. That is, we compute the expectation

$$E_{\text{all data}}(\log L_1(\pi)) = N[p^* \log (\pi) +(1- p^*) \log (1-\pi)]$$

where the Appendix gives the computation.

We are interested not just in the performance of one particular estimator π, but in all. So we now average over all estimators. Since $\hat{p}= n/N$, we know that $\hat{p}$ will inherit its distribution from n. It is normally distributed about a mean p^* with variance $p^*(1- p^*)/N$. The expectation over all data and over all $\hat{p}$ yields

$$E_{\text{all } \hat{p}, \text{ data}}(\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) - 1/2 \qquad (1)$$

The first term on the right is the average log likelihood using the true chance p^* over all data:

$$E_{\text{all data}}(\log L_1(p^*)) = N[p^* \log (p^*) +(1- p^*) \log (1- p^*)]$$

The average in (1) is the quantity that measures the success of the maximum likelihood estimators in the one parameter family. It tells us how their log likelihoods fare on average over all possible data sets and thus is corrected for overfitting. We compare this quantity with the corresponding quantity from other families in choosing our final estimate. We read from (1) that the maximum likelihood estimators fare slightly worse overall than the true value p^*, indicating that we have successfully corrected the overfitting of the maximum likelihood estimators.

However, we are not yet in a position to use (1) since we do not know the value of $E_{\text{all data}}(\log L_1(p^*))$. We need to have some estimate of it since it will vary from parameter space to parameter space and thus affect our choices. We will not be able to determine it exactly. The

[85] It could fail in many ways. The true chance of heads my vary with different tosses; or there may be correlations between successive toss outcomes.

true value p^* is precisely what is unknown and sought. However there is an indirect way that we can recover a good estimate of $E_{\text{all data}}(\log L_1(p^*))$. We use the fact, that for each particular data set, the maximum likelihood estimator $\hat{p}$ tuned to that data set will always outperform the true value p^*.

The extent of overperformance will vary from case to case and will be unknown to us in any particular case. However we can compute its average. To do this, we average over a different set from the one used in (1). That is, we average over pairs of data sets and the estimator best tuned to the data set. That is, we look at a data set and the estimator tuned to it and compare that estimator's log likelihood with that of the true value p^*; and we repeat for many cases. The average that results is expressed by the expectation

$$E_{\text{all } \hat{p}, \text{ data tuned to } \hat{p}}(\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) + 1/2 \qquad (2)$$

The Akaike Information Criterion is recovered by combining equations (1) and (2). Equation (2) tells us that, on average in the data sets for which it is computed, the log likelihood $\hat{p}$ will yield a log likelihood greater by 1/2 than that of the true chance p^* averaged over all data. Hence we can use $\log L_1(\hat{p})$-1/2 as an estimator of $E_{\text{all data}}(\log L_1(p^*))$. Inserting this into (1), we find that $\log L_1(\hat{p})$-1/2-1/2 = $\log L_1(\hat{p})$-1 is an estimator of the quantity we seek to optimize, $E_{\text{all } \hat{p}, \text{ data}}(\log L_1(\hat{p}))$. That is, $\log L_1(\hat{p})$-1 is an estimator of the average log likelihood of $\hat{p}$, averaged over all possible data sets. Maximizing this quantity $\log L_1(\hat{p})$-1 is what AIC calls for in the case of a one dimensional parameter space.

6.4 d Parameter Model

It might seem that a major step must be taken from this last case of a one parameter model to the case of a d parameter model. However all the hard work has already been done in computing the one parameter case. It is a small step to a d parameter case. To get there, we divide the N tosses into d subsets of tosses. We posit different true chances, p^*_1 for the first M_1 tosses, p^*_2 for the next M_2 tosses, …, p^*_d for the final M_d tosses. We have now introduced a d parameter model, with parameters $p_1, p_2, \ldots, p_d$. Each subset of tosses can be treated as a separate one dimensional parameter space problem. So in each subset of tosses M_i, we estimate the average of the maximum likelihoods of $\hat{p}_i$ by computing $\log L_1(\hat{p}_i)$-1. The estimate for the average maximum likelihood associated with all d parameters is just the sum of these individual estimators, that is

$$\sum_{i=1}^{d} \log L_1(\hat{p}_i)\text{-}1 = \log L_d(\hat{p}_1, \ldots, \hat{p}_1)\text{-}d$$

But this last quantity is just the quantity to be maximized in applying Akaike's Information Criterion in the d dimensional parameter space of a d parameter model.

The result still depends upon restrictive assumptions: all of the M_i must be large enough for the central limit theorem to take effect; and we have assumed that some set of values for the p_i expresses the truth exactly. What the calculation also shows is that the character of the parameter space is of lesser importance. The particular magnitudes of the subsets M_i played no role in the final result. They can each be different in size, as a long as they are each large enough to support an application of the central limit theorem. All that matters is that they open new dimensions in the parameter space. It is this fact that enables the criterion to be expressed so simply in terms of parameter space dimension only.

6.5 Akaike Information Criterion Computed

The analysis is specific enough for us to be able to use AIC to compare the zero and one parameter models in a context in which we have an independent intuitive grasp of the competing factors. In 100 coin tosses, if the coin is fair so that the chance of a head is 1/2, we expect the number of heads to lie in the range 40 to 60.[86] When do we choose the hypothesis from the zero or the one parameter models?

For the zero parameter model, the quantity maximized in the Akaike Information Criterion is

$$\log L_0(1/2) = 100 \log (1/2) = - 69.31$$

For the one parameter model, it is

$$\log L_1(\hat{p})-1$$
$$= 100([\hat{p} \log (\hat{p}) +(1- \hat{p}) \log (1- \hat{p})] -1$$
$$= 100([(n/100) \log (n/100) +(1- n/100) \log (1- n/100)]) - 1$$

where n is the number of heads and $\hat{p} = n/100$. If we plot these two quantities as a function of n, we find Figure 2.

[86] The mean number is 50 and the standard deviation is $\sqrt{100.1/2.1/2} = 5$, so the two standard deviation interval 40-60 and will contain the outcome with probability 0.954.

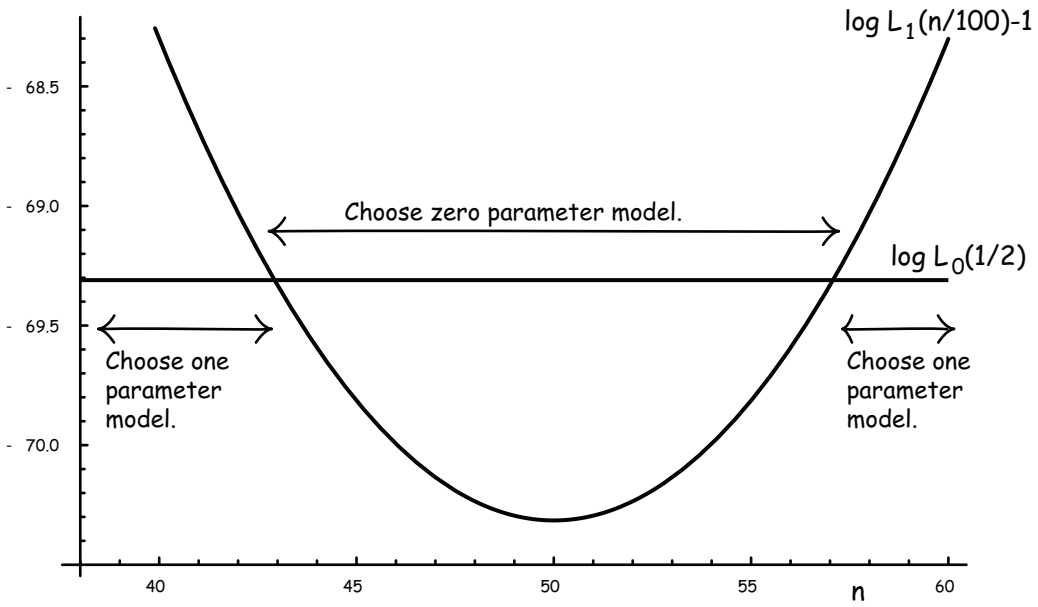


Figure 2. Comparing the zero and one parameter models.

From it, we see that the zero parameter model returns a higher value when n lies between 43 and 57, so we choose the zero parameter estimator p=1/2 for those values. Otherwise, when n falls outside this range, we choose the one parameter estimator $\hat{p}$= n/100.

Here is how we can interpret these results. When we have a datum n=49, the outcome is close enough to the expected value n=50 of the zero parameter model that we prefer the zero parameter model. The one parameter model would give us $\hat{p}$= 0.49 and, as a result, a log likelihood of the data slightly greater than that of p=1/2. However the gain is due to overfitting and not sufficiently great to lead us to switch from the zero parameter value of p=1/2.
If, however, the outcome were to be n=40, then the situation is reversed. The one parameter model gives us $\hat{p}$= 0.40 and a log likelihood for the data that so exceeds the one from p=1/2 that we switch to the one parameter model. These decisions conform with what our vaguer notions would dictate in this case.

7. Relation to the Material Theory of Induction

The main ideas of the connection have already been seen above. I collect them and develop them here. The material theory of induction denies that there is any universal schema for inductive logic. A candidate for such a schema is the idea that we should choose the simpler hypothesis over the more complicated. We have already seen the difficulty with positing this as an independent rule. We still lack any universal characterization of what is simple. At best we

can identify the simpler cases on an ad hoc basis according to domains we encounter. The schema also raises the deeper issue of whether it requires us to presume some sort of metaphysics of simplicity. It would assert that the world is, in its essential construction, parsimonious. Are we willing to accept this metaphysics of simplicity? If not, how do we justify the universal schema just described?

The material theory of induction asserts that we should not accept this simplicity schema as universal. Rather it asserts that any schema for inductive inference is warranted by facts and the schema is applicable only in the domains in which those facts obtain. In the case of the Akaike Information Criterion, the essential posit is that the true hypothesis lies somewhere among the hypotheses of the model that we seek to fit. That assumption in turn gives us sufficient access to all possible data sets that the true probabilistic system may generate for us to correct for overfitting by the models.

The derivation of the criterion makes no prior supposition of parsimony or simplicity of the world. It merely asks that we choose estimators that perform well over all possible data sets, not just the ones to which they were initially tuned. The Akaike Information Criterion then follows. That there is any connection to simplicity understood as a general and abstract notion is an interpretation we supply after the analysis is complete. We look at the correction factor d applied to the log likelihood. It reminds us of a vaguer idea that we find it apt to penalize more complicated models with larger numbers of parameters. So it may seem to us that the criterion is somehow vindicating some broader metaphysics of simplicity. That is an illusion and a mistake. The success of the criterion supplies nothing of the sort. We make a mistake in connecting a statistical data analysis procedure, grounded in quite specific assumptions about the case at hand, to some ill-formulated and dubious metaphysics of simplicity.

The following consideration shows how dependent the approach is on the selection of models and how little it can be said to understand deeper notions of simplicity and complexity. Consider two models. The first is a two parameter model with parameters p_1 and p_2. Call the model $M_2(p_1, p_2)$ and assume that the AIC directs us to select the particular hypothesis with parameters $\hat{p}_1$ and $\hat{p}_2$, chosen since they maximize the penalized log likelihood $\log L_2(p_1, p_2) - 2$. Now consider a second, one parameter model M_1 defined by

$$M_1(p_1) = M_2(p_1, \hat{p}_2)$$

where the log likelihoods of the two models will be related by

$$\log L_1(p_1) = \log L_2(p_1, \hat{p}_2)$$

It is immediately clear that AIC will direct us to favor the one parameter model M_1 over the two parameter model M_2. We can readily find values for which the one parameter model's penalized

log likelihood outperforms that of the two parameter model. For example, if in both we set p_1 to the same value $\hat{p}_1$ returned for the two parameter model, we find

$$\log L_1(\hat{p}_1) - 1 > \log L_2(\hat{p}_1, \hat{p}_2) - 2$$

since $\log L_1(\hat{p}_1) = \log L_2(\hat{p}_1, \hat{p}_2)$.

From our elevated perspective, we know that the case is an unfair contrivance. The model M_1 is really just the same as M_2 with one of its parameters artificially hidden by the contrivance of setting it to the estimator value in advance. We would want to say that it is unfair to ask any method to do well against examples precisely contrived to confound them. But that is the point. Calling up some higher perspective, *we* know that the example is contrived. The AIC analysis itself has no way of knowing that. All it can know is that there are two models, a one parameter M_1, and two parameter M_2, which it treats by its rules. The method has no access to which model is really simple and which is maliciously contrived to look simple and has no provisions for treating them differently.

Finally, Forster and Sober's introduction of the Akaike Information Criterion into philosophy of science attracted some spirited responses. For example, De Vito (1997) urged that it could not overcome the language dependence brought by "grue-like" problems. Myrvold and Harper (2002) have pointed out cases in which AIC fails to pick hypotheses that successfully extrapolate.

These are all worthy complaints in so far as they are leveled against the idea that the AIC has somehow vindicated a broader metaphysics of simplicity. However, once one realizes that the real power and proper ambitions of the AIC analysis are much more modest, these concerns pass. Forster (1999) has responded that variant, grueified descriptions cannot change the dimension of the parameter space that is central to the AIC analysis. Also, I will note here, we can only expect the hypothesis selected by an AIC analysis to fare well in extrapolations if the true hypothesis lies within the models considered. Counterexamples in which the AIC selection fails in extrapolation are easily found merely by contriving examples in which the true hypothesis lies outside the models. Then failure of extrapolation is untroubling since the AIC approach, properly understood, has no power to estimate a truth that lies outside its compass. Understood materially, an AIC analysis can only achieve ends authorized by the assumptions made in the analysis. These assumptions fall far short of the positing of a metaphysics of simplicity that can provide universal guidance whenever philosophical issues of simplicity of are raised.

Appendix: Computations for the Akaike Information Criterion in a Simple Coin Tossing Problem

A coin is tossed N times, where N is very large, and the outcome of n heads is reported as the data. In the one parameter model, we assume that the probability of a head in each toss is equal to some undermined probability p, so that the probability of a tail is (1-p). With independence of the tosses, it now follows that the probability of n heads in N tosses is $(p)^n(1-p)^{n-N}$. Hence the one parameter log likelihood is

$$\log L_1(p) = \log (p)^n (1-p)^{n-N} = n \log p + (N-n) \log (1-p)$$

The maximum likelihood estimator is that value of p that maximizes this likelihood. That is, $\hat{p}$ solves the equation

$$0 = (d/dp) \log L_1(p) = n.(1/p) - (n-N).(1/1-p)$$

which leads to

$$\hat{p} = n/N.$$

Thus, the log likelihood of any data set with n heads according to this estimator is

$$\log L_1(\hat{p}) = N [(n/N) \log \hat{p} + (1-n/N) \log (1-\hat{p})]$$

We now seek to assess how well some particular estimator, say $\hat{p}=\pi$, fares when we consider all possible data sets. We assume that the true value of p is p^* and that n/N will differ from its mean value p^* by an amount δ. Writing $n/N = p^*+\delta$, we have

$$\begin{aligned} \log L_1(\pi) &= N [(n/N) \log \pi + (1-n/N) \log (1-\pi)] \\ &= N [(p^*+\delta) \log \pi + (1-p^*-\delta) \log (1-\pi)] \\ &= N [p^* \log \pi + (1-p^*) \log (1-\pi)] + \delta \log (\pi/(1-\pi)) \end{aligned}$$

We now average this quantity over all possible data sets. The number of heads n/N is distributed about the mean p^*. Hence $\delta = n/N - p^*$ has a mean of 0 and vanishes under the expectation operator $E_{\text{all data}}$. Thus we find:[87]

$$E_{\text{all data}}(\log L_1(\pi)) = N[p^* \log (\pi) + (1-p^*) \log (1-\pi)]$$

This expectation depends explicitly on the value of $\hat{p}=\pi$. To suppress it, we now average over the possible values of $\hat{p}$. Writing $\hat{p} = p^*+\Delta$, where we now assume that Δ is small, we have

$$E_{\text{all data}}(\log L_1(\hat{p})) = N[p^* \log (p^*+\Delta) + (1-p^*) \log (1-p^*-\Delta)]$$

[87] This computation does *not* require the assumption that N is large and that n is normally distributed.

We expand the two log terms in a power series:

$$\log (p^*+\Delta) = \log p^* + \log (1 + \Delta/p^*) \approx \log p^* + \Delta/p^* - (1/2)(\Delta/p^*)^2$$

$$\log (1 - p^*- \Delta) = \log (1-p^*) + \log (1 - \Delta/(1-p^*)) \approx \log (1 - p^*) - \Delta/(1-p^*) - (1/2)(\Delta/(1-p^*))^2$$

After substituting, multiplying terms and saving terms up to Δ^2, we have

$$E_{\text{all data}}(\log L_1(\hat{p})) \approx N[p^* \log (p^*) +(1- p^*) \log (1- p^*)] - (1/2)N\Delta^2/(p^*(1-p^*))$$

The quantity Δ is a random variable that inherits its probability distribution from n. When N is large, n is normally distributed[88] with a mean p^*N and a variance $Np^*(1-p^*)$. Since $\hat{p} = n/N$ and $\Delta = \hat{p}-p^* = n/N-p^*$it now follows that $Z = \Delta/\sqrt{p^*(1-p^*)/N}$ is a standard normal variable with mean 0 and variance 1. Hence $Z^2 = \frac{N\Delta^2}{p^*(1-p^*)}$ is chi–squared distributed with one degree of freedom. This distribution has the property that its mean is unity. Hence taking the expectation of $E_{\text{all data}}(\log L_1(\hat{p}))$ over all values of $\hat{p}$, we recover:

$$E_{\text{all } \hat{p}, \text{ data}} = N[p^* \log p^* +(1-p^*) \log (1-p^*)] -1/2$$

To identify the first term on the right hand side, note that that the likelihood of n heads according to the correct chance p^* is

$$\log L_1(p^*) = N [(n/N) \log p^* +(1-n/N) \log (1- p^*)]$$

We also have the expectation

$$E_{\text{all data}}(n/N) = p^*$$

so that

$$E_{\text{all data}}(\log L_1(p^*)) = N[p^* \log p^* +(1-p^*) \log (1-p^*)]$$

Combining, we have

$$E_{\text{all } \hat{p}, \text{ data}}(\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) - 1/2 \qquad (1)$$

of the main text.

To arrive at (2) we compute the behavior of $\log L_1(\hat{p})$ over the data sets to which each $\hat{p}$ is tuned. To limit ourselves to these data sets, we set $n/N = \hat{p}$ in

$$\log L_1(\hat{p}) = N [(n/N) \log \hat{p} +(1-n/N) \log (1-\hat{p})]$$

and write $\hat{p} = p^*+\Delta$ as before, so that

$$\log L_1(\hat{p}) = N [(p^*+\Delta) \log (p^*+\Delta) +(1-p^*-\Delta) \log (1- p^*-\Delta)]$$

---

[88] This follows since the exact distribution of n is a binomial distribution with these same parameters. The central limit theorem tells us that this distribution approaches a normal distribution of the same mean and variance for large N.

Expanding the log terms as a power series in Δ as before, multiplying out terms and saving terms up to Δ^2, we have

$$\log L_1(\hat{p}) \approx N[p^* \log p^* + (1-p^*) \log (1-p^*)] + N\Delta \log (p^*/(1-p^*)) + (1/2)N\Delta^2/(p^*(1-p^*))$$

From above, we have that Δ is a standard normal variable with mean zero and $N\Delta^2/(p^*(1-p^*))$ is chi-squared distributed with one degree of freedom and thus has a mean of 1. Hence we recover the expectation:

$$E_{\text{all } \hat{p}\text{, data tuned to } \hat{p}} (\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) + 1/2 \qquad (2)$$

The quantity to be maximized in AIC is recovered from (1) and (2) as described in the main text.

References

Akaike, Hirotugu (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control*, **19** (No. 6): pp. 716–723.

Burnham, Kenneth P. and Anderson, David R. (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods and Research*, **33**, pp. 261-304.

De Vito, Scott (1997) "A Gruesome Problem for the Curve Fitting Solution," *British Journal for the Philosophy of Science,* **48**, pp. 391-396.

Konishi, Sadanori and Kitagawa, Genshiro (2008) *Information Criteria and Statistical Modeling*, New York, NY: Springer.

Forster, Malcolm (1999), "Model Selection in Science: The Problem of Language Variance," *British Journal for the Philosophy of Science*, **50**, pp. 83-102

Forster, Malcolm and Sober, Elliott (1994) "How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions," *British Journal for the Philosophy of Science*, **45**, 1–35.

Myrvold, Wayne C. and Harper, William L. (2002) "Model Selection, Simplicity, and Scientific Inference," *Philosophy of Science*, **69**, pp. S135-49.

Wasserman, Larry (2000) "Bayesian Model Selection and Model Averaging," *Journal of Mathematical Psychology,* pp. 92-107.

Zucchini, Walter (2000) "An Introduction to Model Selection," *Journal of Mathematical Psychology,* **44**, pp. 41-61.

Chapter 8

Inference to the Best Explanation: The General Account

1. Introduction

The project of this part of this book is to show how standard inductive inference forms are materially grounded in background facts as opposed to inductive inference schema of universal applicability. This chapter and the next address the inductive inference form known as "inference to the best explanation" or "abduction." The leading idea is that a theory or hypothesis must do more than merely accommodate or predict the evidence. If it is to accrue inductive support from the evidence, it must explain it. Since multiple explanations are possible, we are enjoined to infer to the best of them. That means that greater explanatory prowess confers greater inductive support. In 1964, Penzias and Wilson found puzzling residual noise in their radio antenna that turned out to be cosmic in origin. Subsequent investigation showed it to be thermal radiation at 2.7 degrees Kelvin. The radiation was explained by big bang cosmology as the much diluted and cooled thermal radiation left over from the hot big bang over 10^{10} years ago. The competing steady state cosmology and other now less well-known models could provide no comparably strong explanation. Cosmologist inferred to big bang cosmology as the best explanation.

Inference to the best explanation, however, has proven to be an especially troublesome case for my project. The difficulty does not lie with the material theory of induction. The difficulty lies with inference to the best explanation itself as an inductive inference form. Beyond the simple sketch just given, its elaboration is noticeably thin in the literature.

This thinness persists in spite of efforts to deepen our understanding of the inference form. The starting point is the notion of explanation itself in science. The general literature in philosophy of science has sought to elevate the notion beyond mere psychological satisfaction with some theory or hypothesis. It has become a core notion in philosophy of science and the subject of intense philosophical scrutiny. As far as abductive inference is concerned, the hope has been that this scrutiny will reveal something in the nature of explanation that makes it peculiarly potent in powering abductive inferences and that this in turn will enable a more precise statement of the general rule of abduction. This expectation has set the scene for decades

of frustration. Philosophical analysis of explanation has failed even to find a univocal sense of explanation at work in science. Instead it has found multiple, competing senses of explanation. This multiplicity indicates that the notion is a loose one, an umbrella term that collects several disparate notions. They have no common core such as might power a formal, inductive inference schema. As a result, the literature has provided no universal, formal account of abductive inference. Even the best developed accounts offer only superficial descriptions that use terms like "explains" and "loveliest" without giving them precise, formal definitions.

Most of the analysis of this and the next chapter, then, is devoted merely to an attempt to do better at understanding just how the inferences designated as abductive work. These efforts draw on a series of canonical examples of abduction in science, described in the next chapter. My initial hope was that these examples would reveal the secret ingredient in good explanation that rewards explanatory prowess with inductive support. I would then seek its material underpinning. The plan has failed. That secret ingredient has proven elusive. The inductive support proved time and again to come indirectly through weaknesses of competing explanations as opposed to coming from some special virtue of the preferred explanation. That omission led to the curious notion developed in this chapter of "inference to the best explanation without explanation."

The upshot is that inference to the best explanation is an overrated argument form. Its strength is its visceral appeal. We apply it when we have an hypothesis or theory that fits the evidence in a strikingly satisfying manner. It just feels right, even if that feeling is created retrospectively from sanitized textbooks accounts. What remains is to move our affection for the argument form from psychology to reason. That is, we need to find a unified account of just how the inference works and what warrants it.

If explanatory prowess is what powers the argument, then there is good reason to suspect that no such unified account is can be given. For we have to hope that a heterogeneous notion of explanation can somehow underwrite a homogeneous inductive power. In this case, inference to the best explanation will remain merely a label for a heterogeneous group of inferences powered more by visceral intuitions than good reasons.

In coming to these conclusions, I join a persistent, minority tradition in philosophy of science that has deprecated the importance of explanation in inferences identified as abductive. The conclusions conform with those of Day and Kincaid (1994, p. 282) in so far as they assert:

> In short, appeals to the best explanation are really implicit appeals to substantive empirical assumptions, not to some privileged from of inference. It is the substantive assumptions that do the real work.

They associate this view with the similar approach to arguments based on simplicity advocated by Sober and also developed here in the chapter on simplicity. Van Fraassen's (1977; 1980, Ch.

5) pragmatic deflation of explanation is well-known. More recently, Roche and Sober (2013) make their main claim clear in the title of their paper: "Explanatoriness is evidentially irrelevant, or inference to the best explanation meets Bayesian confirmation theory." Khalifa et al. (2017) argue that inference to the best explanation does not provide a fundamental argument form. Rather its instances are reducible to other inferences and these are not unifiable by a simple scheme.

Section 2 below recalls the identification of abduction as an argument form by scientists, most notably, Charles Darwin. Sections 3 to 6 provide a brief survey of the philosophical literature on inference to the best explanation. This literature is so large that the survey must be brief and incomplete.[89] The survey yields the unhappy result that this literature has done a poor job of developing inference to the best explanation as a general argument form. There are three problems.

First, the basic concepts invoked remain imprecisely defined. Worse, efforts to explicate these concepts trigger a death spiral of multiplying problems: clarifying one concept requires introduction of several new ones that in turn require their own clarifications.

Second, the selection of illustrative examples is commonly poor. Examples in science are often just named or glossed hastily and claimed to support some favored conclusion. We shall see in the next chapter that a closer examination of canonical examples commonly returns conclusions at variance with the existing literature. Most importantly, explanation proves to have a minor role in them.

Third, there is a strong tendency to employ illustrative examples that involve human action. They are poor surrogates for the corresponding scientific examples. In the human case, that the favored explanation is correct is obvious immediately and the exploration of alternatives at best a perfunctory exercise. There really are no credible, competing explanations for the origin of the bootprints in the freshly fallen snow. We might try to suppose that the snow just happened to settle into the shape of boot, complete with a boot's characteristic tread pattern. But the thought is too strained to bear serious consideration. Scientific examples are quite unlike this. It is far from obvious that the big bang is the unique, credible explanation of the cosmic background radiation. As we shall see in the next chapter, the real work in the examples is establishing with some effort that no other explanation can likely succeed.

---

[89] For another overview, see Douven (2016). I also do not explore the literature that investigates the inference to the best explanation from a Bayesian perspective, such as Iranzo (2008) and Henderson (2014). The reason is that Bayesian analysis cannot be applied everywhere, as later chapters this book take pains to show. Thus the Bayesian analysis has at best narrow applicability.

This unsatisfactory situation is resolved, I will argue here and in the next chapter, if we abandon the search for a single, unified formal account of these inferences. Instead, we approach the examples materially, on a case-by-case basis. We then find that there is commonly a clear warrant for the inferences in background facts, as required by the material theory of induction. We also see some similarities in how these facts are deployed to provide the warrant and it is these similarities that sustain the sense that inferences somehow belong together. The similarities, however, are not strong enough to support a formal schema, but just a loose resemblance. Most important, once we have found the warrant for the inferences in background facts, we have enough warrant. There is no longer any need to search fruitlessly within the very notion of explanation itself for some unifying, special constituent that confers inductive powers upon explanation.

What remains is to identify the loose similarities that connect the inferences commonly identified as abductive or as inference to the best explanation. Drawing on the inventory of examples in science in the next chapter, the result is summarized in Section 7. Abductions or inferences to the best explanation in actual science are carried out in two steps with some distinctive notion of explanation playing no role in either.

The first is a comparative step. The favored hypothesis or theory is shown to do better than one or more foils. We are to prefer, but not necessarily infer to, the better of them. We might call this "Preference for the better explanation." The way the favored hypothesis or theory does better turns out to be simple. While the preferred hypothesis or theory accommodates the evidence, the foil might just be contradicted by the evidence. Or the foil might require additional posits which do not themselves have evidential grounding. That lack amounts to what I shall call the incurring of an "evidential debt" not taken by the favored hypothesis or theory. It is then easy to see how the evidential judgments of this first step are supported by material facts, for the still elusive general notion of explanation plays no role. We prefer the theory that is not contradicted by the evidence; or the theory that accommodates the evidence without overt lacunae of support in its individual parts.

The second step is more fraught. We are to suppose that better is best; and that best is good enough to warrant commitment. Preference becomes commitment. The step is commonly grounded in a presumption that no other theory can do better than those explicitly considered. That presumption is so hard to justify that this second step is often left tacit and sometimes even omitted completely. For the step commonly relies merely on our human imaginative powers to sustain the conclusion that there is no better account just beyond our horizon. Kyle Stanford (2006) has effectively and powerfully described this problem of "unconceived alternatives."

Section 8 presents a conjecture on why, historically, inference to the best explanation rose in prominence as an argument form in the twentieth century. Section 9 offers a concluding comparison of the formal and material approach to abduction.

2. Scientists Explain

What gives inference to the best explanation solid credentials in philosophy of science is that scientists themselves often advertise the explanatory prowess of their theories and suggest it provides support for their theories. Here are two prominent examples.

Upland geese, Darwin (1876, pp. 142-43) reports, rarely go near water, but they have the webbed feet of great utility to aquatic birds. This curious fact, Darwin continues, is readily explained by natural selection as a residual from ancestral aquatic geese. It is poorly explained by the hypothesis of independent creation. Why create them with this unnecessary feature? Darwin made observations like this the explicit driver of his argument in *Origin of Species*. He concludes the final chapter with a defense of the argument form, not just in biology, but in ordinary life and the other sciences (1876, p. 421):

> It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified. It has recently been objected that this is an unsafe method of arguing; but it is a method used in judging of the common events of life, and has often been used by the greatest natural philosophers. The undulatory theory of light has thus been arrived at; and the belief in the revolution of the earth on its own axis was until lately supported by hardly any direct evidence. It is no valid objection that science as yet throws no light on the far higher problem of the essence or origin of life. Who can explain what is the essence of the attraction of gravity? No one now objects to following out the results consequent on this unknown element of attraction; notwithstanding that Leibnitz formerly accused Newton of introducing "occult qualities and miracles into philosophy.

In late 1915, Einstein was perfecting his general theory of relativity. It was then a highly speculative theory, operating at a level of abstraction and mathematical complexity remote from the other physical theories of his time. Einstein needed an evidential coup to secure the theory. That came in mid November 1915, when Einstein discovered to his delight that his new theory predicted the anomalous motion of Mercury. In a paper entitled “Explanation of the Perihelion Motion of Mercury from the General Theory of Relativity,” he wrote:

> In the present paper, I find an important confirmation of this most radical theory of relativity; that is, it turns out that the secular rotation of Mercury’s orbit in the

> direction of the orbital motion, discovered by Leverrier, which amounts to about 45" in a century, is explained qualitatively and quantitatively, without having to posit any special hypothesis at all.

This success was so striking that it is one of the most used illustrations in subsequent work in confirmation theory. We shall return to both examples below.

These two examples of Darwin and Einstein make at least a *prima facie* case that there is an interesting inductive argument form at hand that is somehow associated with a notion of explanation. One would expect that logicians and philosophers of science would be able to seize upon these clues and deliver a rigorous and logically tight account of the argument form. Alas, the brief survey below of the philosophical literature reveals one that is stalled in preliminary and inadequate sketches of the argument form. Worse, its prospects are limited at the outset by a near universal aversion to real examples in sciences. Instead, the literature favors examples in which the best explanation involves some human action, which makes the examples quite unlike the corresponding inferences in real science. The sections that follow will elaborate this grim assessment

3. Peirce and Abductive Inference

The philosophical literature attributes the first explicit discussion of abductive inference to C. S. Peirce. The much-quoted statement of it comes later in Peirce's writings, from his 1903 Harvard Lecture, "Pragmatism as the Logic of Induction" (1932, 5.189):

> Long before I classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis—which is just what abduction is—was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts of some of them. The form of the inference, therefore, is this:
>
> > The surprising fact, C, is observed;
> > But if A were true, C would be a matter of course,
> > Hence, there is reason to suspect that A is true.
>
> Thus, A cannot be abductively inferred, or if you prefer the expression, cannot be abductively conjectured until its entire content is already present in the premises, "If A were true, C would be a matter of course."

What is curious is the myopia in crediting Peirce. For Darwin's *Origin of Species* was already a tour-de-force of abduction. The inference form is used throughout *Origin*.[90] As we saw in the passage quoted from Darwin above, he was aware of the distinctive character of the argument form he was using and offered a defense of it as something used generally in common life and other great scientific discoveries.[91] What more can we ask? The inference form is identified explicitly at the same time as it is used repeatedly and powerfully in the canonical demonstration of one of science's greatest discoveries. In contrast, Peirce's development is labored. While it has the superficial appearance of a logical schema, key terms are not given precise definitions. Just what is means by "surprising" and "a matter of course"?[92]

4. Harman's Inference to the Best Explanation

Peirce's treatment also conforms to the nineteenth century tradition of combining inductive methods with discovery methods. Mill's methods were as much a way of discovering the causes of some phenomena as they were supporting them inductively. The same is true of Peirce's account of abduction. This procedural aspect is suppressed by the time of Harman's (1965) paper "Inference to the Best Explanation," from which the now popular label derives. His account of the inference is as follows (1965, p.89):

> In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference. Thus one infers, from the premise that a given hypothesis would provide a "better" explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true.

---

[90] In the final edition (Darwin, 1876), the word "explain" appears 108 times and "explanation" 44 times.

[91] What of Lyell's *Principles of Geology*? It contains a template for Darwin's argument in *Origin of Species* and Darwin studied it and drew inspiration from it. While we and, presumably, Darwin saw the argument form there, I will argue in the next chapter that, curiously, Lyell did not.

[92] Peirce's work is littered with citations to Darwin. I have not ascertained whether any of these credit Darwin's priority. Certainly the credit is not given prominently.

The account remains remote from a serviceable formal schema. What it is to explain is uninterpreted and "better" is introduced in scare quotes. Harman proceeds to concede that formulating a more precise account of which explanations are better is an open problem:

> There is, of course, a problem about how one is to judge that one hypothesis is sufficiently better than another hypothesis. Presumably such a judgment will be based on considerations such as which hypothesis is simpler, which is more plausible, which explains more, which is less *ad hoc*, and so forth. I do not wish to deny that there is a problem about explaining the exact nature of these considerations; I will not, however, say anything more about this problem.

The paper is short, a mere eight pages. It has no well-developed examples, but many are mentioned by brief allusion. The only example from a science is the single sentence (p.89) "When a scientist infers the existence of atoms and subatomic particles, he is inferring the truth of an explanation for various data which he wishes to account for." Otherwise all the examples mentioned pertain to human action:

> … a detective … decides that it *must* have been the butler ... (p. 89)
>
> … we infer that a witness is telling the truth (p. 89)
>
> … we infer from a person's behavior to some fact about his mental experience… (p. 89)
>
> … I read … that Stuart Hampshire is to read a paper at Princeton tonight… (p. 92)
>
> … obtaining knowledge from an authority… (p. 93)
>
> … knowledge of mental experience gained from observing behavior … (p. 93)

5. Thagard's Criteria

Paul Thagard's (1977) analysis is an exception to my dismal assessment of the philosophical literature on inference to the best explanation. It excels both in the range of real examples from science and in its dedication to clarifying just how inference to the best explanation works.

The range of examples deployed to illustrate and support the paper's claims is impressive. It includes:

> Darwin's long argument in his *Origin of Species;*
>
> Lavoisier's case for the oxygen theory of combustion;
>
> The wave theory of light, as developed by Huygens in the seventeenth century; and by Young and Fresnel in the nineteenth century;
>
> Newton's explanation of the motion of planets and satellites;

Halley's Newtonian prediction of the return a comet;
Young's account of di-polarization;
Fresnel's account of polarization through transverse waves;
General relativity's treatment of the anomalous perihelion motion of Mercury, the gravitational bending of light and the gravitational red shift of light; and
Quantum mechanics and its success with atomic spectra, magnetism, the solid state of matter, the photoelectric and the Compton effect.

This list is so long as to be too ambitious for a single paper. The accounts given are brief and often amount to mere mentions. However the laudable principle sustained is that Thagard's account is responsible to these real examples from science.

Thagard also recognizes that the inference form is in urgent need of elaboration and clarification; and he takes up the project. From the perspective of the material theory of induction, the project is ill-fated. For the arguments labeled "abductive" or "inference to the best explanation" form at best a loose unity. The individual arguments differ so much in their details that they can be grouped together only as long as the argument form is imperfectly specified. That means that any applicable notion of explanation must be kept vague enough so that it can be applied everywhere. Efforts to remove the vagueness in the notions "explanation" and "better explanation" will require further notions and possibly many of them, if the existing range of individual arguments is to be accommodated. Matters will get worse the further these efforts go, for each solution will generate new problems. An explosion of difficulties will be triggered. Yet the results of these efforts can never be secure. All it takes to overturn them is a new, troublesome instance of an abductive inference.

This fate befalls Thagard's project, as we shall now see. The project begins with a brief definition (p. 77):

> To put it briefly, inference to the best explanation consists in accepting a hypothesis on the grounds that it provides a better explanation of the evidence than is provided by alternative hypotheses. We *argue* for a hypothesis or theory by arguing that it is the best explanation of the evidence

Here the key term "explanation" is left undefined. This serious lacuna persists throughout the paper, until, on the concluding page (p. 92), we find a begrudging admission that "Explanation is a pragmatic notion." Instead, explicit analytic efforts are devoted to clarifying when one explanation—whatever that term may mean--is better than another. The clarification depends on three criteria: consilience, simplicity and analogy; with the difficulty that their evaluations may pull in different directions. Here we see the multiplication of problems. The project has now replaced the problem of clarifying one notion with the problem of clarifying three notions.

The notion of consilience, drawn from the work of William Whewell,[93] is glossed as (p. 79): "one theory is *more* consilient than another if it explains more classes of facts that the other does." The problem then is to specify how we are to count classes of facts, so that the "more" has an unambiguous meaning. Of course there is no simple solution and the analysis stalls with the inevitable difficulty (p. 83)

> In inferring the best explanation, what matters is not the sheer number of facts explained, but the variety, and variety is not a notion for which we can expect a neat formal characterization.

The notion of consilience then further mutates into a static and a dynamic notion.

The threat to a cogent notion of consilience is that any account can be made to embrace more facts if we are willing to make it more complicated. This is where Thagard's second criterion, simplicity, plays a role. Simplicity, he proposes, is measured by size and nature of auxiliary hypotheses needed by some theory to explain the facts. The fewer there are of these auxiliaries, the simpler and the better the explanation. Needless to say, trying to give a more precise account of simplicity leads to further problems and the wry conclusion (p. 88): "As has often been remarked, simplicity is very complex."

Finally the third criterion, analogy, enters apparently through no pressing conceptual need, but simply because the examples driving the analysis use it. Analogies, we are told, function to improve the explanations used. We have already seen in an earlier chapter that efforts to explicate analogical inference face a similar difficulty of multiplication of problems. Thagard's analysis only begins to probe this difficulty. After abandoning a classic definition of analogy, Thagard offers an alternative. If, for some entity A, property S explains why it has properties P, Q and R, then we can project to other cases. That is, if another entity B has properties P, Q and R, then we may "conclude that B has S is a promising explanation of why B has P, Q and R." (p. 90) Of course this characterization is only as good as the characterization of the notion of explanation, for which essentially nothing is offered.

With the close of the paper we are left with the unresolved problem of how to trade off the criteria (p. 92):

> Consilience and simplicity militate against each other, since making a theory more consilient can render the theory less simple, if extra hypotheses are needed to explain the additional facts. The criterion of analogy may be at odds with both consilience and simplicity, if a radically new kind of theory is needed to account simply for all the phenomena.

---

[93] Here Thagard draws on his earlier (1977) where he identified Darwin's use of Whewell's notion.

Leaving this problem unsolved means that we cannot unambiguously apply the rule of inference to the best explanation. Far from recovering a universally applicable rule of inductive inference, we have failed even to arrive at an unambiguous rule.

The material theory of induction was introduced in response to the pervasiveness in formal accounts of inductive inference of difficulties like these. Seeing the burden of multiplying problems drag down his account, I truly sympathize with Thagard's concluding lament (p. 92) "Application of the criteria of consilience, simplicity, and analogy is a very complicated matter."

6. Lipton's Monograph

Peter Lipton was the most prominent of more recent proponents of inference to the best explanation and his monograph (2004) has become the default, canonical source. His work (2000, 2004) provides no formula or schema that would improve on those in Darwin, Peirce, Harman or Thagard. Rather his detailed elaboration maps out just how open is the problem Harman and Thagard set aside. We have no notion of explanation or better explanation sufficiently well developed to convert what Lipton (2004) repeatedly calls the "slogan" of inference to the best explanation into formal schema

Take the notion of explanation. Efforts to clarify it lead to the same multiplication of problems we just saw in Thagard's project. It derives from the fact that there are multiple competing accounts of explanation. Some of them are surveyed in Lipton (2004, Ch. 2). To explain a phenomenon might mean to subsume it under a covering law; or to display those factors that increase its probability; or to display the causes that bring it about. Again, an explanation may unify many phenomena, hitherto thought disparate. Each of these notions captures a sense of explanation applicable in some circumstance. A fully elaborated schema of abduction would then need to accommodate all these further notions. What is a law as opposed to a general proposition? What is the origin of the probabilities? Just what do we mean by cause? How do we distinguish unification from mere conjunction? Needless to say, each of these is an unfinished project in its own right.

Prudently, Lipton does not take on the challenge of finding a schema that embraces all these senses of explanation. Rather his (2004, Ch. 3) develops the causal model of explanation, perhaps because its fits best with his favorite, elaborated example of Semmelweis and his discovery of the cause of childbed fever. However he concedes (2004. p. 3) that he can provide no analysis of the notion of causation and uses it as an unexplicated term.[94]

[94] For my pessimism concerning hope of any general account of causation that might serve his purposes, see Norton (2003).

We press on. How are we to judge which explanation is better? We could, Lipton urges, adopt the most likely explanation. However that reduces abduction to a circularity: the most likely explanation is most likely to be true. Lipton (2004, p. 59, p.121) introduces a distinct characteristic to replace "likeliest": we should infer to the "loveliest" explanation. It will then guide us to the likeliest explanation. What makes an explanation lovelier is, loosely, that it provides the most understanding (p. 59). This derives in turn from what Lipton (2004, p. 122) identifies as "explanatory virtues." They include: "mechanism, precision, scope, simplicity, fertility or fruitfulness, and fit with background belief" (p.122) as well as "unification" (p. 139). Once again, the single problem of determining which explanation is lovelier has multiplied into many, unresolved problems. We are quite far from any account of these virtues that would allow them a place in a formal schema of inductive inference.

Lipton does introduce what are, for present purposes, two important extensions to the notion of inference to the best explanation. The first is the recognition that an explanation must rise to some minimal level of success before we are authorized to infer to it. As a result, he is willing to relabel inference to the best explanation (2004, p.154):

"Inference to the Best Explanation if the Best is Sufficiently Good."

and (2000)

"inference to the best of the available explanations, when the best one is sufficiently good."

Second, Lipton introduces a contrastive notion. It is restricted to causal explanation and its key assertion is (p.42):

> Difference Condition: to explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the case of not-Q.

I will urge below the importance in all cases of distinguishing comparative judgments of which are better explanations from the absolute judgment that some explanation is best.

Unfortunately, Lipton's treatment of examples is superficial, with one exception, and the analysis suffers for it. Many examples are named, but mostly there is little or no explanation of the content of the example. This is a serious problem since we shall see in the next chapter that closer examination of canonical examples leads to conclusions other than those drawn by Lipton. Equally seriously, many of the examples are drawn from ordinary human situations. These are sufficiently unlike important examples in science that reliance on them is dangerous if one's goal is to understand inferences in science.

I have prepared a compendium of the examples as a way of assessing their distribution over types. The follow are examples from science:[95]

A drought may explain a poor crop.

The big bang explains background radiation.

Stress, fatigue etc. explains bridge collapse.

Velocity of recession explains galactic red shift.

Kinetic theory of gases explains thermal phenomena.

Natural selection explains the traits of plants.

Electronic theory explains current flow.

Echolocation explains bat navigation.

The same side of the moon faces us.

Why the planets move in ellipses.

Why leaves turn yellow in November.

Prior syphilis explains why someone contracted paresis.

Freudian wish-fulfillment explains a slip.

A field explains the deflection of a particle.

Chomsky infers language structure.

Lightning and thunder.

Perturbations in the orbit of Uranus explained by Neptune.

Mendeleyev predicts new elements.

Song employed by sparrows.

A double blind test of a drug's efficacy.

Gregor Mendel's peas.

Millikan's oil drops.

"We are more impressed by the fact the special [sic] theory of relativity was used to predict [sic] the shift in the perihelion of Mercury than we would have been if we knew that the theory was constructed in order account for that effect." (p.172)

Others examples are essentially dependent on human actions and thus unlike real examples in science:

Why you didn't come to the party (headache).

Peculiar tracks in the snow in front of my house (snowshoes).

A magician intuits the numbers I am thinking of.

I see a supposedly vacationing friend at the supermarket.

---

[95] This list is a mix of quotes and paraphrasing. No page numbers are given since many examples are repeated over many pages.

The rattle in the car.

Praise and punishment by Israeli airforce instructors.

Why a three year old threw his food on the floor.

Why Lipton went to see *Jumpers* rather than *Candide*.

Why Able rather than Baker got the philosophy job.

"Why did you order eggplant?"

Why Kate rather than Frank won the prize.

Why my horse won rather than yours.

Why Lewis went to Monash rather than Oxford in 1979.

A door opening triggers a bomb.

Why all men in the restaurant are wearing paisley ties.

The butler did it.

The patient has measles.

My front door has been forced open.

Why is my refrigerator not running.

Whether my car will start tomorrow.

Sherlock Holmes' dog that did not bark.

Movement of the mouse causes the movement of the cursor.

A crossword puzzle.

Successful navigation by means of a map.

Still other are intermediate between the two cases:

Sticks in a bunch thrown in the air more likely horizontal.

A spark causes a fire, but oxygen does not.

Why Mercury rose in the thermometer.

Why people feel heat more when humidity is high.

Kuhn infers normal science is governed by exemplars.

Opium puts people to sleep.

Data from flight recorder of crashed plane.

Kahneman and Tversky's Linda the bank teller; people told of a taxi involved in a hit and run accident.

A sympathetic powder that can cure wounds at a distance.

Methods of predicting future performance on the London Metal Exchange.

Persistence forecasting of the weather.

A scan of the lists indicates that the potentially misleading human examples have as much presence as the scientific examples.

Finally, we have one extended example in Lipton's text. It is the identification of the cause of childbed fever by Ignaz Semmelweis in the 1840s in Vienna's maternity hospital. The primary narrative spans seventeen pages (pp. 74-90). The example is well known in philosophy of science through its inclusion in Hempel's (1965; Ch. 2) widely read and highly accessible *Philosophy of Natural Science.* In brief, the maternity hospital had two divisions and, alarmingly, the death rate from childbed fever was markedly higher in one than in the other. Over a period of several years, Semmelweis checked all manner of differences between the two divisions in search of the cause of the difference. None answered until finally Semmelweis realized that the doctors and medical students in the higher mortality division only were delivering babies after performing autopsies elsewhere. He guessed that cadaveric material on the doctors' hands was the cause of the childbed fever. His guess was confirmed when he required the doctors to disinfect their hands with chloride of lime before delivering in the maternity ward, whereupon the differential death rate disappeared.

The case is a splendid example of dedicated scientific detective work and the powerful use of evidence. However as a case study intended to display the merits of inference to the best explanation specifically, the case study is a failure. For that, what is needed is a case study in which the evidential relations depend quite specifically on the distinctive merits peculiar to inference to the best explanation. It would do so in way that makes it unlikely that any other account of inductive inference could do as well. This is not that case study, for explanation plays little if any role in the analysis. Rather Semmelweis' investigation and analysis is a near perfect example of the application of Mill's methods.

The clearest application comes in the identification of the cause. Mill's method of difference applies when we have two instances, one in which the phenomenon of interest occurs and one in which it does not. If they differ only in one circumstance, that is the cause. This is precisely the case faced by Semmelweis. In the key experiment, the only change associated with the drop in mortality was that the doctors were disinfecting their hands from cadaveric material with chloride of lime. The eliminated cadaveric material was the cause.

We can see in Semmelweis' (2008, pp. 7-8) own narrative how his analysis was driven by just such considerations.

> As mentioned, the commissions identified various endemic factors as causes of the greater mortality rate in the first clinic. Accordingly, various measures were instituted, but none brought the mortality rate within that of the second clinic. Thus one could infer that the factors identified by the commissions were not causally responsible for the greater mortality in the first clinic. I assumed that the cause of the greater mortality rate was cadaverous particles adhering to the hands of examining obstetricians. I removed this cause by chlorine washings.

> Consequently, mortality in the first clinic fell below that of the second. I therefore concluded that cadaverous matter adhering to the hands of the physicians was, in reality, the cause of the increased mortality rate in the first clinic. Since the chlorine washings were instituted with such dramatic success, not even the smallest additional changes in the procedures of the first clinic were adopted to which the decline in mortality could be even partially attributed.

Clearly all that is at issue in Semmelweis' analysis is to find the difference that makes a difference and to identify it as the cause. Of course one can embed Semmelweis' analysis in a larger narrative replete with discussion of how the cadaveric material explains the childbed fever, as Lipton does. However to do so is unnecessary. Semmelweis' own analysis makes no essential use of explanatory notions.

The brief remarks above already indicate how well Semmelweis' methodology is captured by Mill's methods. Scholl (2013) has given a more thorough analysis of Semmelweis' methodology and finds extensive use of Mill's methods, including Mill's method of agreement and of concomitant variation. Scholl (2015) argues for the failure of Lipton's attempts to impugn the understanding of Semmelweis' analysis as an application of Mill's methods.

7. Inference to the Best Explanation without Explanation: Two Step Reconstruction

What do inferences commonly labeled abductive or inference to the best explain have in common? The examples of the next chapter are loosely bound together by a simple two-step scheme. The scheme does not require a sophisticated notion of explanation. Mere accommodation is all that is needed. Here we may conjecture that Lipton was not just unlucky in choosing as his major example a case in which explanation proved to play no special role. While Lipton's choice of the Semmelweis example was especially poor, it does reflect a problem that will be repeated in every example we shall examine in the next chapter: the more closely we look at the example, the less important is the role of explanation as a distinct notion of philosophers.

Step 1. Preference for the Better Explanation

What these examples have in common is that they all involve a comparison of a favored theory or hypothesis with one or more foils. The favored hypothesis is adequate to the evidence, most commonly in the sense that it deductively entails the evidence. The foils, that is, the alternatives, are judged inadequate in one of two ways:

Contradiction: The evidence at hand may directly contradict the alternative; or the evidence supplemented by specific background facts may contradict the alternative.

Evidential debt: to accept the alternative requires us to accept further assumptions for which we have no evidence.

The essential point is that the favoring invokes no explanatory notions, unless one accepts that the notions invoked here are a full, if thin, account of explanation. If the disfavoring consists of the alternative facing contradictions, it is simple logic. We prefer the logically consistent over the inconsistent. If the disfavoring is driven by evidential debt, a simple test shows that the presence of the evidential debt is fully responsible for the disfavoring. In one case in the next chapter, the explanation of the anomalous perihelion motion of mercury, we shall see that if the evidential debt could have been be discharged, what would have resulted is a fully admissible hypothesis or theory of a type that is much celebrated.

Step 2. From Comparative to Absolute: Better is Best

This first step just gives a reason to prefer one hypothesis or theory over another. That is not enough if we are to commit to the preferred hypothesis, as the inference scheme requires. We need more and that comes from an assumption that no other hypothesis or theory can do better that our preferred one.

Under any account of inference to the best explanation—material or otherwise—this is the fragile step. Whereas the comparative judgments of Step 1 are explicit in the scientists' narrative, this absolute judgment is not.

How can this step be warranted? The surest case arises when background assumptions assure us that the hypotheses or theories we have considered are exhaustive. Then there are no more credible candidates left, so the best of the ones considered must also be the best. These background assumptions are the assumptions that warrant the inference.

The most difficult case is the most common. It is when the inference from better to absolute best is made, even though the scientist have no clear grasp of the full range of hypotheses or theories possible. Then, at worst, the inference is unwarranted. Or perhaps, better, there is a tacit meta-argument at work. The argument works not at the level of theories but of theorizers. The assumption is that the theorizers are sufficiently inventive and perspicacious to have surveyed the full range of hypotheses or theories applicable; and that they have considered the most credible. Once again, the best of those considered then must be the absolute best. These background assumptions over the power of theorizers warrants the inference from better to best.

8. Why Inference to the Best Explanation?

Given that that the full two-step inference faces such difficulties, why has it come to prominence over the last century or so? It is because it has helped theorists solve a vexing evidential problem. In earlier theorizing, the theorists were often in the happy position that they could infer fairly directly from the evidence to the theory. Newton, for example, could infer quickly *from* Kepler's third law of planetary motion *to* an inverse square law of gravitational attraction for the planets. He spoke confidently of *deductions* from the phenomena. While that now sounds extravagant, Newton's inferences from the phenomena employed background assumptions that made them deductive. (See, for example, Harper, 2002.) Even as late as 1929, Hubble (1929) could arrive at his Hubble law for the speed of recession of the galaxies merely by fitting a straight line to a plot of velocity-distance data.

By Hubble's time, these happy days of easily supported theories were passing. This was especially clear with Einstein's general theory of relativity. It was a theory of such enormous complexity that no similar inference from phenomena to theory was possible. The gap was just too great. While the problem is not as stark, others faced similar problems. Darwin could not infer directly from his mass of evidence in natural history to natural selection. The relationship between it and his theory was just too complicated.

How can theorists close the gap? Perhaps a direct inference cannot be made from the evidence to the theory. However they sense that the theory fits the evidence so well that it must have something right. That sense is viscerally strong and can be communicated fairly easily by recounting the details of the example. It is often expressed compactly by the claim that the theory explains the evidence. The task remaining for the inductive logician, however, is to take the loosely articulated, but viscerally powerful sense and translate it into a transparent analysis of just how the evidence supports the theory. The project of translating it into a precise, general, formal schema remains unfinished and, if the arguments of this chapter and book are upheld, will remain so. However, if a material warrant is sought on a case-by-case basis, it can be found in background facts.

9. Conclusion

Does inference to the best explanation provide a serviceable, general rule of inductive inference? Its failure to do so is shown by a simple question: if we know that some hypothesis gives the best explanation of the evidence, should we infer to it? The answer, of course, is that without further details we simply cannot say. When we look more closely at the details, the strength of the inference becomes clearer. Since the strength of the inductive support can only be

assessed, in the end, by looking at the details of the case at hand, we can see that inference to the best explanation is not a self-contained rule of inductive inference. It is at best a loose guide in urgent need of development. We have seen in this chapter that prospects for development are meager. Efforts to develop the rule lead to a multiplication of problems. Each solution brings more problems than it solves. The more we try to clarify the general argument form, the less clear it becomes.

That inference to the best explanation should be troubled in just this way is quite expected according to the material theory of induction. For, according to it, there can be no universal formal rule covering all the cases. At best, the inferences grouped under the label "inference to the best explanation" form a loose unity that breaks once we look more closely at each inference. The most precise assessment of the inductive strength of any particular argument comes only when we fully take into account the background facts that warrant the inference. That one has an inference to the best explanation provides, in the end, only an indication of a loose similarity with other arguments and nothing more.

References

Darwin, Charles R. (1876). *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. 6th ed. London: John Murray.

Day, Timothy and Kincaid, Harold (1994) "Putting Inference to the Best Explanation in its Place," *Synthese*, **98**, pp. 271-95.

Douven, Igor, (2016) "Abduction," *Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2016/entries/abduction/.

Einstein, Albert (1915) "Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie," *Preussische Akademie der Wissenschaften, Sitzungsberichte*, 1915 (part 2), pp. 831–839.

Harman, Gilbert H. (1965) "The Inference to the Best Explanation," *The Philosophical Review*, **74**, pp. 88-95.

Hempel, Carl G. (1966). *Philosophy of Natural Science*. Upper Saddle River, NJ: Prentice Hall.

Henderson, Leah (2014) "Bayesianism and Inference to the Best Explanation," *British Journal for the Philosophy of Science*, **65**, pp. 687-715.

Hubble, Edwin (1929) "A Relation Between Distance and Radial Velocity Among Extra-Galactic Nebulae," *Proceedings of the National Academy of Science*, **15**, pp. 168–173.

Iranzo, Valeriano (2008) "Bayesianism and Inference to the Best Explanation," *Theoria*, **61**, pp. 89-106.

Khalifa, Kareem; Millson, Jared; and Risjord, Mark (2017) “Inference to the Best Explanation: Fundamentalism’s Failures,” pp. 80-96 in *Best Explanations: New Essays on Inference to the Best Explanation*, eds. K. McCain and T. Poston. Oxford: Oxford University Press.

Lipton, Peter (2000) “Inference to the Best Explanation,” pp. 184-93 in, W. H. Newton-Smith, ed., *A Companion to the Philosophy of Science*. Malden, MA: Blackwell.

Lipton, Peter (2004) *Inference to the Best Explanation.* 2nd ed. London: Routledge.

Norton, John D. (2003) "Causation as Folk Science," *Philosophers' Imprint* Vol. 3, No. 4 http://www.philosophersimprint.org/003004/; reprinted in pp. 11-44, H. Price and R. Corry, *Causation, Physics and the Constitution of Reality*. Oxford: Oxford University Press.

Peirce, Charles S. (1932). Collected Papers of Charles Sanders Peirce. Vol. V. Pragmatism and Pragmaticism. Cambridge MA: Harvard University Press.

Roche, William and Sober, Elliott (2013), “Explanatoriness is evidentially irrelevant, or inference to the best explanation meets Bayesian confirmation theory,” *Analysis*, **73**, pp. 659-668.

Scholl, Raphael (2013) “Causal Inference, Mechanisms, and the Semmelweis Case,” *Studies in History and Philosophy of Science*, **44**, pp. 66-76.

Scholl, Raphael (2015) “Inference to the Best Explanation in the Catch-22: How Much Autonomy for Mill’s Method of Difference?” *European Journal for the Philosophy of Science*, **5**, pp. 89-110.

Semmelweis, Ignaz (2008) “The Etiology, Concept and Prophylaxis of Childbed Fever: Classics in Social Medicine,” *Social Medicine*, **3**, pp. 4-12.

Stanford, Kyle (2006) Exceeding Our Grasp: Science, History and the Problem of Unconceived Alternatives. Oxford: Oxford University Press.

Thagard, Paul R. (1977) “Discussion: Darwin and Whewell,” *Studies in History and Philosophy of Science*, **8**, pp. 353-56.

Thagard, Paul R. (1978) “The Best Explanation: Criteria for Theory Choice,” *Journal of Philosophy*, **75**, pp. 76-92.

Van Fraassen, Bas (1977) “The Pragmatics of Explanation,” *American Philosophical Quarterly*, **14**, pp. 143-50.

Van Fraassen, Bas (1980) *The Scientific Image*. Oxford: Clarendon.

December 29, 2016; May 5, 2017
Minor typo corrections June 18, 2018

Chapter 9

Inference to the Best Explanation: Examples

1. Introduction

According to the material theory of induction, there can be no universally applicable schema that fully characterizes inference to the best explanation. At best we can find loose similarities that the canonical examples of inference to best explanation share. These loose similarities were codified in the last chapter (Section 7) in a two-step characterization. First is a comparative step in which one hypothesis or theory is favored over one or more foils. In the second step, this favoring is rendered absolute: we are authorized to infer to the favored hypothesis or theory. It was noted in the last chapter that this characterization was derived from a compendium of canonical examples of inference to the best explanation. That compendium is given in this chapter. The seven examples are developed in Sections 4 to 10. For ease of overview, their fit to the general characterization is summarized in a table in Section 3. Section 11 has a general conclusion. First, however, the next section offers reflections on the importance of the examples.

2. Examples Matter

What this chapter shows is that taking examples seriously is important. I have already lamented in various places in the last chapter how the present literature has often treated its examples in too great haste. There are two ways that this hastiness has obscured the evidential relations in the examples in science.

First, it is common to employ examples in which human action plays an essential role. This is appealing for, in these are examples, the analysis is easiest and most compelling. However that ease comes just because these examples are poor surrogates for the real examples in science, where the evidential relations are commonly less clear. Specifically, these examples mislead us since, unlike the examples in science, the role of the comparative foil in minimal. Lipton (2004, p. 6) gives the time-worn example:

> Faced with tracks in the snow of a certain peculiar shape, I infer that a person on snowshoes has recently passed this way.

Once one has seen the distinctive imprints left by snowshoes, there is really only one account to be given of their origin. We might invent fanciful scenarios just to drive home that there is no real choice. Lipton (2004, p. 56) shows how it is done:

> Of course, there is always more than one possible explanation for any phenomenon–the tracks might have instead been caused by a trained monkey on snowshoes, or by the elaborate etchings of an environmental artist–so we cannot infer something simply because it is a possible explanation. It must somehow be the best of competing explanations.

However, entertaining these alternatives rapidly becomes a perfunctory exercise in eliminating the fanciful. We might well dismiss them as comic relief.

Here these human examples are quite unlike the real scientific examples. The alternative hypotheses or theories in the scientific cases were not jokes. Prevailing over them is, almost everywhere, the greater challenge, as we shall see below. The wave theory of light struggled for centuries both with its own early weaknesses and the fact that the competing emission theory had been delivered by the authority of authorities, Isaac Newton himself. Darwin struggled to account for the eye, where his creationist opponents could readily explain the perfection of its design with their designer. The anomalous motion of Mercury's perihelion could be explained by Seeliger's zodiacal light, if only it could be determined that it held enough matter. It was essential to Einstein's general theory of relativity that this quite prosaic account fail, for nothing in the elegance of Einstein's theory could protect it if Seeliger's hypothesis proved workable. Thomson's particle theory of cathode rays had to overcome Lenard's ether wave theory. It is only in retrospect that we see how precarious was Thomson's victory, for the soon-to-emerge quantum theory did attribute wavelike properties to electrons.

Second, much of the literature on inference to the best explanation mentions examples in science but does not explore them fully. As a result, they draw on dangerously oversimplified caricatures and miss the real moral of the examples. Superficially, for example, big bang cosmology provides an account of Penzias and Wilson's observation of cosmic background radiation rich in explanatory virtue. As a result the inference to the big bang looks immediate and irresistible and can be drawn without much concern for other accounts. However, if one teases out the history, as is done below, one finds that that the explanatory virtue was initially less clear and less decisive. It took decades before the inference was secure; and only popular simplifications of the history could make the inference seem immediate and irresistible.

More importantly, the essential and delicate part of the analysis was not establishing that big bang cosmology could accommodate the result. Almost any cosmology could deliver

background radiation in one form or another. All it needed was to include electrically charged matter; and every viable cosmology must do this, else it cannot harbor stars that shine in the electromagnetic spectrum. Rather the burden was first to establish, with some effort over years, a particular thermal form for the background radiation and then to argue in some detail why competing accounts could not recover it. Then the evidential success looks less like a sudden explanatory coup of one theory than a slowly building and widespread failure of the competitors. This dynamic is repeated in many examples.

3. Synopsis of Examples

In the characterization of inference to the best explanation of the last chapter (Section 7), the principal burden is to establish superiority of the favored hypothesis or theory over a competing foil or foils. In the second step, the status of the favored hypothesis or theory can, but may not, be generalized from better explanation to best. The table below indicates in summary how the examples of this chapter instantiate this characterization.

Abduction	*Foil*	*Foil eliminated*	*Generalization from better to best*
Darwin on the origin of species	Independent creation	Refuted by traits without function	Tacit assumption of exhaustive choice
Lyell's uniformitarian geology	Geologies using presently unknown causes	Novel causes incur an undischarged evidential debt.	Known versus unknown causes is exhaustive
Thomson for cathode rays as charged particles	Cathode rays are processes in the ether.	Contradiction with experiment: Ether waves would not be bent by a uniform field	Tacit assumption of exhaustive choice
Lenard for cathode rays as ether processes	Cathode rays are processes in matter	Contradiction with experiment: cathode rays in evacuated tubes	Choice between matter and ether posed as exhaustive dilemma.
Einstein's explanation of Mercury's anomalous motion	Many. Modifications to Newtonian theory. Unobserved masses.	Contradiction with experience. Undischarged evidential debt.	Step not taken.
Cosmic background radiation from the big bang	Alternative cosmologies, especially steady state cosmology	Empirical failure	Taken tacitly
Lavoisier's oxygen chemistry.	Phlogiston chemistry.	Contradiction. Matter has weight (gravity), but phlogiston has levity.	Fact (matter has weight) is one of many warranting facts.
Wave theory of light.	Newtonian corpuscular theory.	Undischarged evidential debt. Contradiction with experiment.	Complicated.

Table 1. Summary of Examples

4. Darwin and The Origin of Species[96]

We saw in the last chapter that one of the earliest statements of what we now call inference to the best explanation appeared in Darwin's *Origin of Species.* My task here is focused narrowly on the argument as it is developed in this particular volume of Darwin's writings. My concern is not how the analysis may be developed in other of Darwin's writings. My concern is definitely not how we might presently make the case for the theory of evolution. The modern case rests on a much larger evidential base and has a greater reliance on supporting sciences, such as Mendelian genetic theory, unknown to Darwin. It resolves many of the problems troubling Darwin's development.

[96] I thank Zina Ward for helpful discussion of this section.

4.1 Darwin's Argument

The whole volume, Darwin tells us, develops what he calls "one long argument" (1876, p. 404). It is an argument that cannot be reproduced here with any fidelity. It depends on a lengthy, massively impressive recitation of detailed facts in natural history. They are explained by a wonderfully simple process. There is in nature a constant struggle for survival by living beings. They grow at a geometrical rate that outpaces the arithmetic growth of resources. Favorable variations give their bearers an advantage. Nature selects them for survival, just as domestic breeders select commercially desirable variations. Those selected flourish, leaving offspring with similar characteristics. Darwin (1876, pp. 102-103) offered a simple summary:

> This principle of preservation, or the survival of the fittest I have called Natural Selection. It leads to the improvement of each creature in relation to its organic and inorganic conditions of life:…

The development of the argument in *Origin* then follows a simple formula. Some feature of living beings is displayed and then an account is given of how it could arise through Natural Selection. A reader cannot but be overwhelmed by the sheer mass of facts in natural history that Natural Selection accommodates. No short selection here can do justice to it. Darwin (1876, p. 414) himself tries to convey its weight in a concluding chapter with a rapid recitation of successes:

> Many other facts are, as it seems to me, explicable on this theory. How strange it is that a bird, under the form of a woodpecker, should prey on insects on the ground; that upland geese which rarely or never swim, should possess webbed feet; that a thrushlike bird should dive and feed on sub-aquatic insects; and that a petrel should have the habits and structure fitting it for the life of an auk! and so in endless other cases. But on the view of each species constantly trying to increase in number, with natural selection always ready to adapt the slowly varying descendants of each to any unoccupied or ill-occupied place in nature, these facts cease to be strange, or might even have been anticipated.

These successes lead up to what is, for our purposes, the key evidential claim (1876, p. 421)[97]

> It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified.

[97] This confident claim was not present in the first edition and, presumably, was added as part of Darwin's response to his critics.

Darwin does not justify this key claim. It is, presumably, offered as self-evident. It is plausible, however, that Darwin was following William Whewell. The latter described the consilience of induction as arising when one theory proves, unexpectedly, to explain more classes of facts; and this, Whewell urged, is a powerful indicator of the truth of the theory. (For elaboration, see Thagard, 1977; 1978, Section II.)

In spite of its many successes, Darwin's theory faced serious difficulties. Darwin sought as well as he could to deal with them. They have the character of the taking on of an evidential debt: a supposition needed for the theory to succeed but for which evidence was then lacking. Here are two examples.

The first was that there was some doubt that the earth was sufficiently old for the extraordinary amount of time Darwin's theory required for natural selection to do its work. Darwin's (1876, p. 409) best hope was merely to keep the problem an open question, still to be decided:

> With respect to the lapse of time not having been sufficient since our planet was consolidated for the assumed amount of organic change, and this objection, as urged by Sir William Thompson, is probably one of the gravest as yet advanced, I can only say, firstly, that we do not know at what rate species change as measured by years, and secondly, that many philosophers are not as yet willing to admit that we know enough of the constitution of the universe and of the interior of our globe to speculate with safety on its past duration.

The second was the absence of intermediates. Darwin's theory required all variation to arise through very slow, small gradations. Yet nature has vast gaps between various forms. The evolution of the eye presented a special problem, since its perfection as an optical instrument was no naturally explained as the handiwork of a creator. Darwin strove for pages to make plausible that a light sensitive nerve in some being might eventually develop into an eye. However he could in the end do little better than to appeal to his reader's indulgence (1876, p. 145):

> He who will go thus far, ought not to hesitate to go one step further, if he finds on finishing this volume that large bodies of facts, otherwise inexplicable, can be explained by the theory of modification through natural selection; he ought to admit that a structure even as perfect as an eagle's eye might thus be formed, although in this case he does not know the transitional states.

The foil against which Darwin competed was independent creation: the thesis that "species were immutable productions, and had been separately created." (Darwin, 1876, p. xiii) On the development of eye, Darwin (1876, p. 146) was straining merely to match independent creation:

> It is scarcely possible to avoid comparing the eye with a telescope. We know that this instrument has been perfected by the long-continued efforts of the highest human intellects; and we naturally infer that the eye has been formed by a somewhat analogous process. But may not this inference be presumptuous? Have we any right to assume that the Creator works by intellectual powers like those of man?

Elsewhere, repeatedly, in the volume, Darwin sought to do better than the thesis of independent creation. The means depended on assuming just what he had suggested we had no right to do. That is, his arguments depended on assuming that a creator would only endow a being with a trait if that trait had some useful purpose; and that if there are similarities across species there must be some discernible purpose for them. With that assumption lingering behind his remarks, time and again Darwin could point out some feature that had no evident purpose, but arose naturally through the slow developments fostered by natural selection.

In besting the thesis of independent creation, Darwin was fond of the superlative "utterly inexplicable," using it at least four times:

> …we can clearly understand these analogies [clustering of species], if species once existed as varieties, and thus originated; whereas, these analogies are utterly inexplicable if species are independent creations. (p. 47)

and similarly

> This grand fact of the grouping of all organic beings under what is called the Natural System, is utterly inexplicable on the theory of creation. (p. 413)

> Such cases as the presence of peculiar species of bats on oceanic islands and the absence of all other terrestrial mammals, are facts utterly inexplicable on the theory of independent acts of creation. (p. 419)

> On the view of each organism with all its separate parts having been specially created, how utterly inexplicable is it that organs bearing the plain stamp of inutility, such as the teeth in the embryonic calf or the shrivelled wings under the soldered wing-covers of many beetles, should so frequently occur. Nature may be said to have taken pains to reveal her scheme of modification, by means of rudimentary organs, of embryological and homologous structures, but we are too blind to understand her meaning. (pp. 420-421)

As the examples of his deprecation of independent creation multiply, Darwin rarely speculates on the details of this competing theory. The tacit assumption everywhere is that independent

creation delivers immutable species all of whose traits have a purpose. An exception arises when Darwin (1876, pp. 130-31) recalls similarities between equine species:

> He who believes that each equine species was independently created, will, I presume, assert that each species has been created with a tendency to vary, both under nature and under domestication, in this particular manner, so as often to become striped like the other species of the genus; and that each has been created with a strong tendency, when crossed with species inhabiting distant quarters of the world, to produce hybrids resembling in their stripes, not their own parents, but other species of the genus. To admit this view is, as it seems to me, to reject a real for an unreal, or at least for an unknown, cause. It makes the works of God a mere mockery and deception; I would almost as soon believe with the old and ignorant cosmogonists, that fossil shells had never lived, but had been created in stone so as to mock the shells living on the sea-shore.

This is a less visible line of argument in Darwin's text: that there is something defective as a theory in positing a process of independent creation.

These thoughts develop into a direct assault on the explanatory viability of a creator. Darwin (1876, pp. 383) reports:

> On the ordinary view of the independent creation of each being, we can only say that so it is;—that it has pleased the Creator to construct all the animals and plants in each great class on a uniform plan; but this is not a scientific explanation.

and then again (p. 422)

> It is so easy to hide our ignorance under such expressions as the "plan of creation," "unity of design," &c., and to think that we give an explanation when we only re-state a fact.

To tease out Darwin's objection, imagine that we make up a huge list of species and their traits. To add the remark that the creator planned it so, adds nothing of explanatory value.

To sum up, Darwin's argument for his theory rests on its explanatory prowess with a huge array of facts in natural history. The meaning of the term "explanation" is not given. However the familiar covering law account of explanation fits his usage as well as any: the many facts are explained since they are entailed by his theory. More precisely, the *possibility* of the specific facts is entailed by his theory, for natural selection cannot predict specifically each of the many facts Darwin reports. The strength of the explanation resides in the breadth and variety of the facts covered. Perhaps this is a quiet echo of Whewell's notion of consilience of induction.

The foil against which Darwin rails is the independent creation of each species as immutable productions. His claim of its explanatory defects rests on the tacit assumption that each trait of a living creature must have a purpose; and that this is also the case for similarities

among different species. Without some assumption of this type, Darwin has no real basis for discarding independent creation. For without it the thesis is so incompletely defined that no evidential test is possible. What in nature might then count as favorable or unfavorable evidence? Finally, Darwin turns this difficulty against independent creation by suggesting that is it no explanatory theory at all.

Darwin's theory has its difficulties and these require him to take on some undischarged evidential debt, such as the supposition of long times for natural selection to work and there were transitional forms now not in evidence. We are to conclude, however, that the foil of independent creation is so troubled that Darwin's theory prevails.

4.2 What Powers the Inference

The delicate but central question in this analysis is just what powers Darwin's inference. Let us review some possibilities for a general account that employs a formal principle.

At some intuitive level, there is a sense of beauty and elegance in Darwin's theory and wonder that it embraces such a diversity of fact. This gives it the ring of truth. That feeling, however, falls well short of what an inductive logic, formal or even material, requires. Is the principle that the evidential support is strong merely if we genuinely and honestly feel it is so? That is not a sustainable principle of logic. Worse, how are we to deal with the case in which the feeling is not widely shared? That is our case. When Darwin announced his theory, the public debate was spirited. Darwin's critics were not swayed.

Might the inference be powered by the general result that Darwin himself cites: that the theory "explain[s], in so satisfactory a manner … several large classes of facts…"? I will continue to take the otherwise undefined term "explain" to mean "derive their possibility from a few posits of the theory." As noted above, the situation is more complicated. For that is not quite what Darwin's theory does. The derivation does proceed from a few simple posits. However it also draws upon suppositions that are themselves in great need of further evidential support. The theory requires an extraordinary amount of time for its operations to succeed; and many of them, such as the descent of eyes, are presumed possible while required intermediate states are not found. They are also presumed. These are evidential debts that, in other examples in this chapter, are sufficient to lead to the abandoning of a theory. There are evidential strengths and weaknesses to be balanced here before a final decision can be taken. Darwin delineates no general inductive principle to which his analysis conforms. There is no formal theory provided, even in vague outline, that negotiates the complexity of the balancing. In its absence, a formal analysis is unpromising.

The prospects for a material analysis are more promising. For, even lacking a formal theory, Darwin himself and his sympathetic supporters found powerful support for Darwin's

theory in his evidence. They had only facts to draw upon. Another promising sign for a material analysis is that contemporary commentators disagreed so pointedly. They have the same evidence and arguments before them. If these alone are compelling, then disagreement can only come from ineptitude in the logic. If however background factual assumptions also bear crucially on the cogency of the argument, then matters improve. For, if we allow that different commentators harbored different background assumptions, then the disagreement is intelligible. We need attribute no inductive fallacies to the disagreeing commentators.

We can see how material facts could underwrite Darwin's confidence in his theory if we presume that Darwin found his analysis to establish two facts:

1. It is possible that the variety of species arose from descent with modification through natural selection.

and

2. It is unlikely that any other admissible account can accommodate the origin of species.

These facts combined are sufficient to warrant acceptance of Darwin's theory. His account is possibly right; no others are; therefore his has to be right. No formal principle of induction is needed.

The first fact is demonstrated by the massive weight of Darwin's many examples. The second fact is essential, for, without it, merely establishing possibility is insufficient. Unfortunately establishing this second fact is more difficult. For the only other account given serious analysis in Darwin's volume is independent creation. He does cast significant doubt on independent creation. It is contradicted by many arbitrary facts in natural history for which a creator would have no evident purpose. Darwin even calls into doubt whether independent creation counts as an explanatory theory at all.

What is left open is the question of whether there are still other theories possible that may do as well or better than Darwin's. Of course it is hard for us to imagine what these still better theories might be. But that our imagination fails is poor proof that there are no such theories. Perhaps Darwin is expecting us to proceed from a background assumption that we have no reason to expect that *any* theory could do justice to the wealth of fact in natural history. So merely finding one is so extraordinary that we can cease our searching. Or perhaps we might suppose that Darwin poses a dilemma for us: either species descend from other pre-existing forms, or they did not. Darwin's theory and independent creation are, we are to suppose, the strongest version of each horn. Perhaps, when Darwin's theory bests independent creation, that is enough to establish that Darwin's theory is not just the better of the two but it is the best of all.

5. Lyell's *Principles of Geology*

Charles Darwin was influenced greatly by the uniformitarian geologist, Charles Lyell. Before Darwin left on his formative voyage on the *Beagle* in 1831, its captain, Robert Fitzroy, gave Darwin a copy of Volume I of Lyell's (1830) *Principles of Geology*. Subsequently in November 1832, Darwin received Volume II (1832) in his mail in Montevideo. The volumes had a profound impact on Darwin, who had been recruited as the voyage's naturalist to work in both geology and zoology.

For our purposes, what is striking in Lyell's *Principles of Geology* is that it provides a near perfect template for the argument that Darwin will later develop in his *Origin of Species*. Lyell's concern is to overturn earlier accounts of the origin of the earth's geological features. These earlier accounts supposed that modern features were formed by presently unknown geological processes typically of far greater violence than those now observed. These were the "catastrophist" theories, as Whewell soon called them. They correspond to Darwin's foil of independent creation, for both presume extraordinary occurrences in the past to explain present features: for Lyell, present geology; for Darwin, the diversity of species. Lyell sought to replace these catastrophist theories with a uniformitarian geology, such as defended by James Hutton before him. In it, present day geological feature are explained by very slow geological processes now in operation, while acting over a long time. Correspondingly, Darwin sought to explain the diversity of species by means of natural selection, which employed slow processes present now, acting over a long time.

The connection to explanation is in the title of Lyell's three volume work. It is

Principles of Geology, Being an Attempt to Explain the Former Changes of the Earth's Surface, by Reference to Causes Now in Operation

and the words "explain" and "explanation" appear throughout the text.[98] Lyell's overall argument is an inference to the best explanation. Present causes acting slowly over a long time are a better explanation of present geological features than past cataclysms and, presumably, the best explanation.

Since inference to the best explanation was not then a recognized mode of argumentation, we would expect that Lyell might provide some defense of it. How do we bridge the gap between an hypothesis that explains well and the truth of the hypothesis? Darwin was sensitive to the need to defend his method of argumentation and repeatedly introduced commentary in its defense. Lyell, however, gave no indication that he saw the need to defend the mode of argumentation. As far as I can see, the first two volumes of *Principles of Geology* contain no

[98] In Volume I, they appear together nearly 100 times.

methodological analysis, beyond chance remarks and colorful reprimands for the errors of past theorists. It is only in the first chapter of Volume III that Lyell gave a more extended defense of his methods.

There he summarized his approach as (1833, p.6)

> In our attempt to unravel these difficult questions, we shall adopt a different course, restricting ourselves to the known or possible operations of existing causes; feeling assured that we have not yet exhausted the resources which the study of the present course of nature may provide, and therefore that we are not authorized, in the infancy of our science, to recur to extraordinary agents. We shall adhere to this plan, not only on the grounds explained in the first volume, but because, as we have above stated, history informs us that this method had always put geologists on the road that leads to truth,—suggesting views which, although imperfect at first, have been found capable of improvement, until at last adopted by universal consent.

It was contrasted with the catastrophist foil (pp. 6-7):[99]

> On the other hand, the opposite method, that of speculating on a former distinct state of things, has led invariably to a multitude of contradictory systems, which have been overthrown one after the other,—which have been found quite incapable of modification,—and which are often required to be precisely reversed.

As with Darwin, the strength of Lyell's case rests ultimately on a massive compilation of illustrations of how presently acting causes could generate the geological features now observed. Conveniently for us, in this chapter Lyell selected three examples to illustrate the differences of the two approaches. The first concerned fossil shells and bones. The former view accounted for them as "fashioned into their present form by a plastic virtue, or some other mysterious agency" (p. 4). Lyell instead sought their origin in biological processes just like those in action today. The second concerned the origin of basalt and similar rocks. The former view attributed it to aqueous processes, while Lyell could point to igneous processes now in action that create such rocks. The third concerned the occurrence of fossil shells in rocks in high mountains. The former view sought some unusual process that might dry up oceans and drop their level. Lyell replaced it with processes that elevate land above an otherwise fixed sea level.

In all this, Lyell treated the uniformitarian view as little removed from providing an explanation of some process by directly observing its cause, as opposed to speculating on a novel

---

[99] The idea of employing just processes now acting is appealing in the abstract. However it can quickly run into trouble. The steady state cosmology of the mid-twentieth century was a uniformitarian cosmology that led its proponents to wild speculation, such as the continuous creation of matter. The big bang theory, its catastrophist competitor, won the day.

cause presently not in evidence. He complained of the catastrophists that "…they felt themselves at liberty to indulge their imaginations, in guessing at what *might* be, rather than in inquiring *what is*;…" (p.2, Lyell's emphasis). And then (p.2):

> It appeared to them more philosophical to speculate on the possibilities of the past, than patiently to explore the realities of the present, and having invented theories under the influence of such maxims, they were consistently unwilling to test their validity by the criterion of their accordance with the ordinary operations of nature.

Lyell's text becomes more polemical, heaping scorn on the catastrophists. (pp. 2-3)

> Never was there a dogma more calculated to foster indolence, and to blunt the keen edge of curiosity, than this assumption of the discordance between the former and the existing causes of change.

This stands is stark contrast with Darwin's cautious defense of his use of causes presently in operation. While we can see nature selecting favorable variations among living beings in processes now in operation, Darwin first offered a lengthy discussion of selection by domestic breeders to convince us of the potency of selection. Perhaps Lyell's task was less formidable. He needed only to establish that processes now in operation might eventually produce a mountain, not an eye.

How does this bear on the concerns of this chapter: the warranting of abductive inferences? In the formal approach, the fact that some hypothesis or theory explains what is observed is confirmatory in virtue of the special character of explanation. Unlike Darwin, Lyell saw no special explanatory relationship between his theory and the geological facts it accommodates that would require any circumspection. The theory, in Lyell's telling, does little more that instruct us merely to observe the causes directly.

The warrant for Lyell's argument for uniformitarianism is readily found in background facts, that is, materially. In analogy with the material warranting of Darwin's argument in *Origin of Species*, we can assume that Lyell seeks to establish two facts:

> 1. It is possible that present geological features arose over long time periods from causes now operating.

and

> 2. It is unlikely that any other admissible account can accommodate their origin.

These two facts are sufficient to warrant acceptance of Lyell's uniformitarianism. His theory is possibly correct; no others are; therefore his has to be correct.

The first fact is established by the wealth of examples in Lyell's account. The second proves a great deal easier to establish that the corresponding fact in Darwin's warrant. For Darwin's foil was specifically the thesis of independent creation; and that left open the possibility of many other theories excluded from explicit analysis. Lyell, however, has two cases

that are exhaustive. Either present geological features arose from causes now in operation; or they did not. The first case is Lyell's uniformitarianism. The second is a theory that must speculate on presently unknown causes or known causes but of presently unknown intensity.

Lyell has a direct and telling objection to theories of this second type: they are taking on an undischarged evidential debt. If fossil shells were formed by some plastic virtue or mysterious agency, then we are owed independent evidence that such virtues and agencies exist. If high mountains were thrown up suddenly by cataclysmic forces, we are again owed independent evidence that such forces existed. Lyell's theory takes on no corresponding evidential debt. We are assured of the existence of the causes he employs since they are in operation now. Perhaps his only evidential debt is that enough time has passed for these causes to produce the geological features we see now.

6. Thomson's Electron

J. J. Thomson's (1897) "Cathode Rays" marks a turning point in physics. Thomson identified the rays produced in a cathode ray tube as beams of negatively charged particles of a fixed charge to mass ratio. These particles would soon carry the name "electron" and would be the first fundamental particle of the menagerie of particles that would be discovered in the twentieth century.

Describing the achievement as a discovery makes it sound like a "look-see" event, such as the discovery that one has bats in one's attic. It was less that and more the identification by astute reasoning of the nature of a phenomenon long observed and probed. It was also the resolution of a debate between English and German physicists over the nature of cathode rays. Are these cathode rays beams of matter? Or are they waves in the ether? Thomson identified them as matter: particles charged with negative electricity. Lenard, Hertz and others identified them as waves in the ether.

For our purposes, the interesting point is that both sides employed abductive inferences. It was duel of abduction, won by Thomson. Below we will look at the abductive inference deployed on both sides. We shall see that they are fully controlled by background assumptions. Key to the arguments of both sides is an assumption of exhaustion: that the two alternatives they considered –-matter or waves--were exhaustive. For Thomson, the assumption was tacit. For Lenard it became explicit: finding trouble for both matter and waves posed, for Lenard, a troubling dilemma, resolved only by a new, third option.

Each side had to establish that their favored account fitted the experimental result and, preferably, did so very well. But that alone did not suffice. Each side also needed to demonstrate that the competing account was untenable. Each claims the other's account refuted by the

experiment. The assumption of exhaustion then did the critical work of allowing the step from the adequacy of each sides' account to its truth. The course of the debate was controlled by the assumptions of exhaustion,

6.1 Thomson: Cathode Rays are Charged Particles

Let us begin with the much-told story. Thomson's argument[100] in his (1897) "Cathode Rays" depended on the extensive series of experiments reported in his paper. In brief, Thomson shows that cathode rays are deflected by electric and magnetic fields in perfect agreement with the basic law of electrodynamics that we now know as the Lorentz force law. That the charges are negative is also affirmed by directing the rays at a metal vessel, which then becomes negatively charged. Perhaps the most powerful part of Thomson's argument is that the experiments with magnetic and with electric deflection both yield the same value for the characteristic mass to charge ratio m/e for the particles. This same ratio is recovered whatever the material of the cathode emitting the rays.[101]

There are many details here that could be pursued. Thomson's experiments were delicate and the detailed development of his case sophisticated. For our purposes, what matters is that Thomson's favored charged particle hypothesis fits his experimental results quite wonderfully well. He sums it up in a much-quoted passage (Thomson, 1897, p.302) as:

> As the cathode rays carry a charge of negative electricity, are deflected by an electrostatic force as if they were negatively electrified, and are acted on by a magnetic force in just the way in which this force would act on a negatively electrified body moving along the path of these rays, I can see no escape from the conclusion that they are charges of negative electricity carried by particles of matter.

---

[100] In his 1906 Nobel lecture, Thomson (1906) does use the words "argument" and "proof" to describe the case he makes (my emphasis):

> "The *arguments* in favour of the rays being negatively charged particles are primarily that they are deflected by a magnet in just the same way as moving, negatively charged electrified particles…. The next step in the *proof* that cathode rays are negatively charged particles was to show that when they are caught in a metal vessel they give up to it a charge of negative electricity."

[101] In Norton (2000, §3.2), I have described this part of Thomson's analysis as employing "overdetermination of constants," an argument strategy that has been employed elsewhere to good effect. I also note (§3.3) the overdetermination of constants by itself is not sufficient to rule out competitors, which is the issue the present text now turns to address.

Thomson does not use the word "explains" or "explanation" here. Unlike Darwin and Lyell, the words are barely used at all. However we can identify Thomson's overall argument as an inference to the best explanation.

The difficulty of Thomson's argument is that his summary establishes only that this particle theory fits wonderfully well. Nothing in his summary establishes that other accounts cannot do as well. One might think it excessive to demand anything more of Thomson, for there seems to be no gap at all in Thomson's argument. However there is a gap. In a few decades, with the rise of quantum theory, propagating electrons will turn out to be waves after all. They might not be waves in a nineteenth century ether. They are waves of quantized particles, so they have wavelike properties nonetheless. More important, these waves have exactly the properties that Thomson found so compelling: they carry negative charge and are deflected just as Thomson found by electric and magnetic fields.

6.2 Lenard: Cathode Rays are Waves

The explicit burden of establishing that no other account can do as well was carried by Thomson's arguments against the competing view. That view was that cathode rays are a form of radiation in some way akin to light or Röntgen rays (also called X-rays). The then prevalent theory represented such radiation as a wave propagating in the all-penetrating ether.

This wave account was defended by Philipp Lenard,[102] student and protégé of Heinrich Hertz, who had just died prematurely in 1894 at the age of 36. Lenard's (1894) poses the problem as one of deciding whether the rays are "processes in matter or processes in ether." These two possibilities represent the only two possibilities allowed by late nineteenth century physics. A discharge tube can contain ordinary matter and ether; there is no third possibility. So a process such as a cathode ray must be a process within one or both of these. Processes in matter, we soon learn, are akin to the propagation of sound, which is carried by the material substance of air. It is quite plausible that cathode rays are something comparable. The electric potential might ionize the gas molecules that are then driven as a ray through the tube by electrical attraction. Processes in ether are akin to light propagation, which is carried by the ether. If of this form, cathode rays would correspondingly be carried as waves in the ether. Any matter present, such as air, would act only as an interference and impede the wave.

This posing of the problem is critical to the further analysis, since it reduces the analysis to deciding between two cases. It is the key assumption of exhaustion. Lenard proposed to decide

[102] And alas soon to be a leading light of the anti-semitic, German science movement of the Nazi era.

between the two by means of an experiment in which cathode rays are propagated in a vacuum. He explained (1894, pp. 226-27):

> [it affords] the possibility of carrying out the very same fundamental experiments, that had decided for light and sound whether these latter are processes in matter or processes in ether.

Light can propagate in a fully evacuated space without obstruction since ether remains. Sound propagation is suppressed entirely, since its material carrier has been eliminated.

Lenard then reports the results of the experiment; they favor the ether process (1894, p. 248):

> Therefore cathode rays also propagate in spaces whose contained matter is only in that extreme dilution in which all known processes in it disappear. One cannot ascribe the mediation of the intensive processes observed to the remainder of the matter, which is more or less completely distant and without influence, but only to the ether, which we cannot remove from any space. If this is accepted, then our experiment on the nature of cathode rays decides that they are processes in the ether.

At first pass, this argument is an abduction: the best explanation of the propagation of cathode rays in a vacuum is that they are ether waves. A more careful analysis, however, shows that explanation as a primitive notion plays no essential role. It is really an eliminative argument. The rays are either material processes like sound or waves in the ether like light. They cannot be the first since they persist in a vacuum. Therefore, by elimination, they must be the second.

Conveniently for us, Lenard then reports others who share the ether process view, thereby giving a contemporary list of those whom Thomson (1897, p. 293) would later identify merely as "German physicists" in the opening of his celebrated "Cathode Rays" paper. They are Heinrich Hertz, Eilhard Wiedemann and Eugen Goldstein.

What comes in Lenard's next paper of the same year is still more interesting. The celebrated quote from Thomson's "Cathode Rays" paper above purports to show that cathode rays are beams of charged particles because they behave in just that way (e.g. "…are acted on by a magnetic force in just the way in which this force would act on a negatively electrified body…"). It is easy for us to read that now as compelling. We might ask, what explains that the rays behave *as if* they are streams of particles? It is that they *are* streams of particles! This is easy hindsight. In 1894, Lenard could dismiss just this argument. He began his second paper on cathode rays of 1894 by noting that cathode rays are deflected by magnetic fields, just as would beams of charged particles (1894a, p. 23):

> Here the behavior of cathode rays agrees with the behavior of a stream of massive, negatively charged particles, projected from the cathode.

Lenard then discounted this agreement as superficial:

> This agreement between cathode rays and radiating matter--which one finds again in other phenomena of radiation and which has even been seen by many physicists since Crookes' experiments to hold generally—can nonetheless only be superficial, if the result drawn earlier [footnote citation to Lenard, 1894], that cathode rays are processes in the ether, was justified.

Lenard's dismissal is not casual. The main point of his paper is to present experimental results that establish the dismissal. He proceeded to describe the experiments and their results (pp. 23-24):

> That the agreement is in fact only superficial seems now to me to be shown especially well in the following experiments, in which the agreement fails completely, when circumstances, which must be of the greatest influence on the speed of radiating matter, turns out to be completely without influence on the magnitude of the magnetic deflection of cathode rays.
>
> The experiments show that the magnitude of the magnetic deflection is not at all influenced by the medium in which the radiation is observed; rather the deflectability of one and same kind of cathode rays remains always immutably the same, in all gases, with all pressures, with each intensity of radiation and even then, if the latter [rays] have passed through a metal wall pushed in front;…

The inference against the particle account depends on the same analysis as Lenard's earlier paper. If cathode rays are streams of matter, then they cannot persist in a vacuum, where there is no matter, just as there can be no sound waves there. Since they depend so much on the matter present, we would expect changes in the matter present to change the amount of deflection of cathode rays by magnetic fields. Yet no such effect is found.

In short, Lenard had asserted three year's before Thomson's celebrated paper, that successful explanation of magnetic deflection by the particle theory is not enough. It is an insufficient basis for inferring to the particle theory that can be overruled by the failure of the theory to fit other experimental facts. Lenard claimed that it has been so overruled by his latest experiments. He then recalled another experiment by Hertz. It also precluded the deflected cathode rays merely being a beam of charged particles acted on directly by a magnet (1894a, p. 32):

> The deflection of cathode rays is, according to Hertz' experiments, not an effect of the magnet on the rays themselves, but an effect of the latter on the medium through which they radiate; the rays propagate differently in a magnetized medium than in a non-magnetized medium. For if the forces act between the magnet and the rays

> themselves, then the magnet must also be deflected by the cathode rays, if the magnet is made movable; this is not the case. [footnote[103]]

The basis of the experiment is elementary electromagnetism. If cathode rays are a beam of charged particles, then they behave electromagnetically much the same as a current in a wire. A current carrying wire creates magnetic effects. Oersted had found a magnetic needle deflected in the vicinity of such a wire. Correspondingly we should find magnetic effects in the vicinity of cathode rays. Yet when Hertz sought them, he found none. His delicately balanced magnet was undeflected.

With the failure of the particle theory now assured, Lenard turned to a brief elaboration of the ether theory. How is it that a magnetic field can deflect a cathode ray? The means, Lenard explained, is indirect. The magnetic field affects the ether and, indirectly through that effect, the cathode rays carried by the ether (1894a, pp. 32-33).

> The medium, however, whose magnetic alteration is shown through the curvature of the rays, is, as a result of our experiments, the ether itself. For the curvature is found to be fully independent of the nature and the density of any ponderable medium present; in particular, it was also observed in the highest vacuum. [footnote[104]]
>
> Therefore, through their curvature, cathode rays give an immediate indication that the state of the ether between magnetic poles is mutable, as is required by the theory of mediated action at a distance.

This is the ether-wave theorists' account of the magnetic deflection of cathode rays. It becomes the key target of Thomson's argument against the ether-wave theory.

6.3 Thomson: Cathode Rays are not Waves

Thomson's celebrated "Cathode Rays" paper of 1896 begins by posing the problem as a decision between two theories of the constitution of cathode rays (p.293): they are "some process in the aether" ("according to the almost unanimous opinion of German physicists"); or they are "wholly material…particles of matter charged with negative electricity." He continued:

> It would seem at first sight that it ought not to be difficult to discriminate between views so different, yet experience shows that this is not the case, as amongst the

---

[103] The footnote "Hertz, Wied. Ann. **19**. p. 799 f. and 805 f. 1883" is to those parts of Hertz (1883) where Hertz reports the negative result of this experiment. Hertz (1883, p. 807) also draws the same conclusion as Lenard: "the magnet acts on the medium, but cathode rays propagate differently in magnetized than in an unmagnetized medium."

[104] To Lenard (1894, pp. 244 and 246).

physicists who have most deeply studied the subject can be found supporters of either theory.

The electrified-particle theory has for purposes of research a great advantage over the aetherial theory, since it is definite and its consequences can be predicted; with the aetherial theory it is impossible to predict what will happen under any given circumstances, as on this theory we are dealing with hitherto unobserved phenomena in the aether, of whose laws we are ignorant.

Thomson's paper then proceeds to recount the well-known experiments that lead up to the conclusion already quoted earlier. Virtually the entirety of the paper and its argumentation are devoted to showing that the charged particle view fits the experiments. He addresses several objections from the ether wave theorists to the particle theory that can be answered experimentally.[105] However there is no sustained effort to show that the ether wave theory cannot perform just as well experimentally as the particle theory. His argument to this effect is so tersely stated as to be impossible to follow if read in isolation. He inserts within a sentence in the introductory paragraph (p. 293) of the ether process theory of cathode rays

"…in a uniform magnetic field their course is circular and not rectilinear—no phenomenon hitherto observed is analogous…"

The difficulty is not of an experimental character, but theoretical, and presumably that is why it was not elaborated in the heavily experimental "Cathode Rays" paper. Fortunately Thomson had already elaborated the point in his presidential address the previous year to the Sixty-Sixth Meeting of the British Association for the Advancement of Science. There he has expressed his skepticism (1896, p. 702)

… also I think very difficult to account for the magnetic deflection of the rays. Let us take the case of a uniform magnetic field: the experiments which have been made on the magnetic deflection of these rays seem to make it clear that in a magnetic field which is sensibly uniform, the path of these rays is curved; now if these rays were due to ether waves, the curvature of the path would show that the velocity of propagation of these waves varied from point to point of the path. That is, the velocity of propagation of these waves is not only affected by the magnetic field, it is affected differently at different parts of the field. But in a uniform field

[105] He shows experimentally that the electric charge is deflected with the rays. Hence they are not merely a distracting secondary effect--"no more to do with the cathode rays that a rifle-ball has with the flash when the rifle is fired." (p. 294) He corrects Hertz's experimental result that cathode rays are undeflected by electric fields by repeating the experiment more carefully. (p. 296).

> what is there to differentiate one part from another; so as to account for the variability of the velocity of wave propagation in such a field? The curvature of the path in a uniform field could not be accounted for by supposing that the velocity of this wave motion depended on the strength of the magnetic field, or that the magnetic field, by distorting the shape of the boundary of the negative dark space,[106] changed the direction of the wave front, and so produced a deflection of the rays.

Thomson here issues a quite fundamental challenge to the wave theorists. The widely recognized experimental fact of cathode rays is that they are deflected by magnetic fields. The standard mechanism through which waves are deflected is refraction, as manifested by light. When a light wave moves through a medium in which its speed becomes variable, the wave is bent. The amount of bending is recovered by the Huygens construction of elementary wave optics. The most familiar example is the bending of a light ray striking the surface of lens. The effect results fully from the difference of the speed of light in air and glass. It is faster in less dense air and slower in more dense glass.

Lenses alter the direction of light propagation abruptly. A gradual deflection arises with the phenomenon of mirages. Air closer to a heated desert surface is less dense than air at higher altitudes. So the speed of light is faster closer to the ground. The effect is that light grazing the desert surface is deflected upwards. Someone looking at the deflected light sees the blue of the sky but coming from the direction of the ground. The resulting illusion of water is a mirage.

Figure 1 shows how the bending arises. Light propagates from left to right. The wavefront AA' is vertical. Since the wavefront's speed is faster closer to the ground, the subsequent wavefront BB' has been turned upward.

---

106 In an incompletely evacuated cathode ray tube, there is a dark space in front of the cathode before the cathode rays strike the gas in the tube and trigger light emission. I have been unable to discern precisely Thomson's argument concerning it.

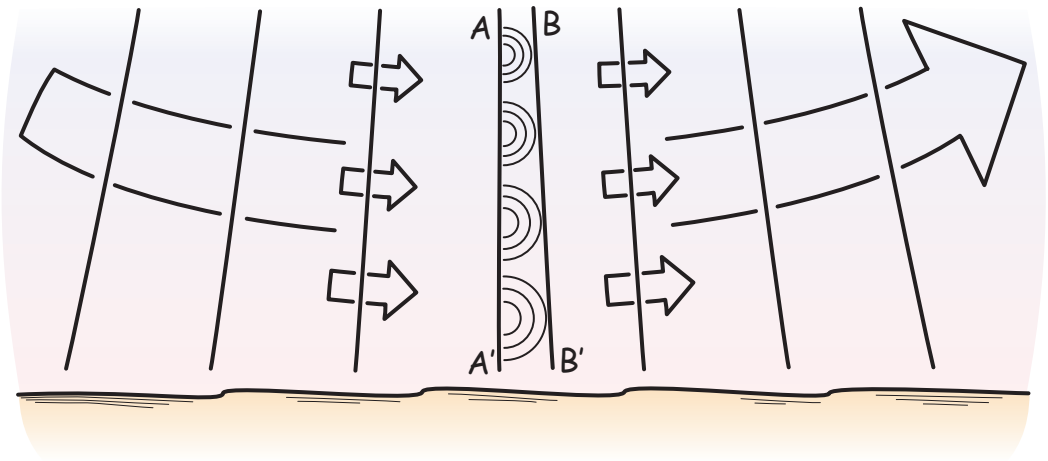


Figure 1. A Mirage: Refraction Bends Light

Thomson's point is that the refractive bending of waves depends essentially on differences in wave speed at different places. A uniform magnetic field, however, is the same everywhere. Hence, Thomson maintains, the effect it has on cathode ray wave propagation must be the same everywhere. There can be no differential alterations in the wave speed and thus no bending of the ray by diffraction. This conclusion would continue to hold even if we allow that the magnetic field might, in some circumstances, induce anisotropic speeds of propagation on the wave; that is, speeds that are different in different directions. Such anisotropy can arise for light propagation in anisotropic media. The corresponding anisotropy cannot arise here, however. The cathode rays are deflected in a plane perpendicular to the uniform magnetic field. The uniform magnetic field is isotropic in this plane.

The charged particle view of cathode rays has no trouble bending the rays. If the charges in the rays have the same initial velocity and start perpendicular to the direction of a uniform magnetic field, then the charges are deflected into a circular orbit in a plane perpendicular to the direction of the field, as shown in Figure 2. This is the "circular course" mentioned by Thomson (1897, p. 293).

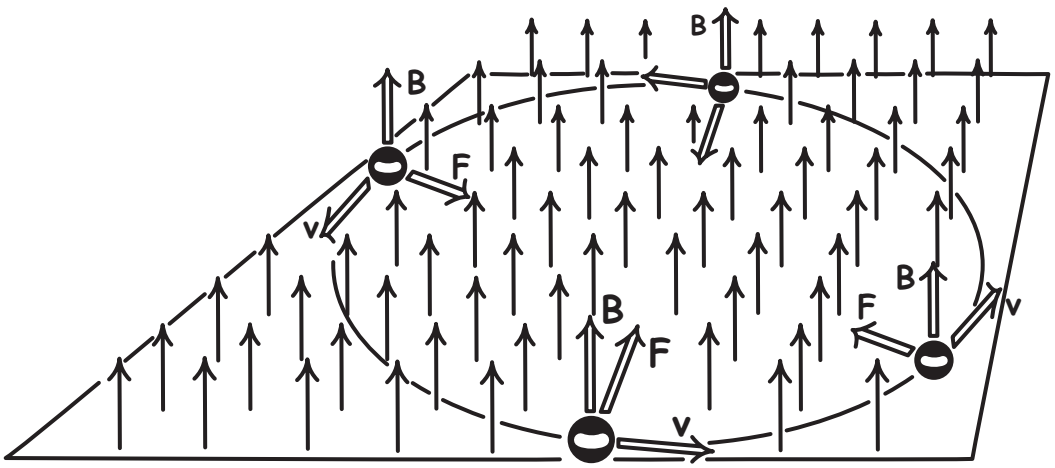


Figure 2. Moving Charge Deflected by a Uniform Magnetic Field

The electrodynamical details, for those who want them, are simple. The force **F** on a charge e of mass m moving at velocity **v** in a magnetic field **B** is

$$\mathbf{F} = m\mathbf{a} = e\,\mathbf{v}\times\mathbf{B}$$

where the force produces acceleration **a**. This acceleration is orthogonal to the direction of the **B** field,[107] so the trajectory will remain in a plane perpendicular to the direction of the **B** field. The acceleration **a** is also orthogonal to the velocity **v**,[108] which entails that the scalar speed $v = |\mathbf{v}|$ is constant.[109] Since the scalar acceleration *a* and scalar speed *v* remain the same in a uniform **B** field, the curvature of the trajectory must be the same everywhere; that is, it is a circle.

6.4 "The Dilemma of Accelerated Molecules and Ether Processes" Resolved

In 1906 Thomson was awarded a Nobel prize for "his theoretical and experimental investigations on the conduction of electricity by gases." The year before, Lenard was awarded a Nobel prize "for his work on cathode rays."[110] In his Nobel Prize lecture (1906), Lenard conceded to Thomson, or at least appeared to concede. The lecture is a boisterous history of his work on cathode rays. He describes the apparently irresolvable

---

[107] Since $\mathbf{a}.\mathbf{B} = (e/m)\,(\mathbf{v}\times\mathbf{B}).\mathbf{B} = (e/m)\,\mathbf{v}.(\mathbf{B}\times\mathbf{B}) = (e/m)\,\mathbf{v}.\mathbf{0} = 0$.

[108] Since $\mathbf{a}.\mathbf{v} = (e/m)\,(\mathbf{v}\times\mathbf{B}).\mathbf{v} = -(e/m)\,(\mathbf{B}\times\mathbf{v}).\mathbf{v} = -(e/m)\,\mathbf{B}.(\mathbf{v}\times\mathbf{v}) = -(e/m)\,\mathbf{B}.\mathbf{0} = 0$.

[109] Since $(d/dt)\,v^2 = 2\,\mathbf{v}.(d\mathbf{v}/dt) = 2\,\mathbf{v}.\mathbf{a} = 0$. To maintain a circular course, we must neglect energy lost by radiation, else v will decrease with energy loss.

[110] Nobel prize citations from
http://www.nobelprize.org/nobel_prizes/physics/laureates/1905/
http://www.nobelprize.org/nobel_prizes/physics/laureates/1906/

dilemma posed by cathode rays prior to Thomson's celebrated work of 1897(Lenard, 1906, p.18):

> For we knew already that the rays are processes in the ether and not material, so it had to appear as downright amazing, that nonetheless they mimicked accelerated, negatively electrified gas molecules so deceptively. Nothing known had led us out of this dilemma of accelerated molecules and ether processes;…

He then reported Thomson's experiments as decisive and announced the resolution of dilemma. (p. 19, Lenard's emphasis)

> The solution of the dilemma therefore was this: The rays are not accelerated, electrically charged molecules, but they are simply accelerated *electricity*. Something we had never believed we had seen: electricity without matter, electric charge without charged bodies. We have found that, therefore, in cathode rays, as already placed in our hands. We have, in some measure, discovered *electricity itself*.…

In short, Lenard is defending his long-standing denial that cathode rays are material processes. They are not matter, but pure electricity, an option not considered in the original analysis.

This was not Thomson's view. He did not offer his experiments as finally delivering "electricity itself." Rather the rays were matter, still, but in a new and very finely divided state (Thomson, 1897, p. 312):

> Thus on this view we have in the cathode rays matter in a new state, a state in which the subdivision of matter is carried very much further than in the ordinary gaseous state: a state in which all matter--that is, matter derived from different sources such as hydrogen, oxygen, &c.--is of one and the same kind; this matter being the substance from which all the chemical elements are built up.

6.5 Electrons are Waves After All

While the nature of cathode rays seemed secure in the wake of J. J. Thomson's celebrated experiments, the success was short-lived. With the coming of quantum mechanics, electrons were identified as having a dual wave- and particle-like character. The wavelike character of electrons was affirmed experimentally by Davisson and Germer (1927). They found that cathode rays, scattered off a crystal of nickel, produced diffraction patterns. The wavelengths of the associated waves conformed with the quantum formula for de Broglie waves.[111]

[111] The elder "J. J." Thomson's son ,"G. P." (George Paget), also conducted experiments of this type, affirming the wave character of electrons.

Thus Thomson's abduction arrived at the wrong conclusion. I state this *not* to impugn Thomson's abduction. It is as good as any. Rather my point is that his inference arrived at the wrong result, because it is dependent completely on background assumptions that proved to be incorrect. This can happen with any inductive inference, for they all depend on background assumptions. The material theory requires that dependence for all inductive inferences. The presence and importance of the background assumptions become quite visible, however, when we try to diagnose where the induction went astray.

In Thomson's case, the fatal intermediate conclusion was that a propagating wave could not *also* be deflected by a magnetic field in just the same way as a beam of charged particles. In quantum theory, neglecting spin, an electron can be represented by the same Hamiltonian as is used for an electron in classical physics. In the quantum case, this Hamiltonian is inserted into the Schrödinger equation to provide an account of an electron as a propagating wave. A standard theorem in quantum theory, Ehrenfest's theorem, assures us that the electron wave is deflected by electromagnetic field just as are classical electrons, as long as the wave packet of the quantum electron is confined to a small region in which the electromagnetic field does not appreciably change. Hence quantum electron waves will also be able to traverse Thomson's "circular course" in a uniform magnetic field.

The details of Ehrenfest's theorem for the electromagnetic case are straightforward but tedious. Working through them provides no special illumination.[112] A smaller observation gives a good sense of precisely which assumption ultimately brought grief to Thomson's abduction. He rejected the possibility that cathode ray waves could be deflected by a uniform magnetic field. The key assumption was that a magnetic field could only deflect the waves by the familiar mechanism of refraction, that is, by directly altering the velocity of propagation of the waves and having a different alteration in different parts of space. A uniform magnetic field could not do this, since its effects must be everywhere the same.

What Thomson had overlooked is that the magnetic field might couple to a propagating wave in other ways. Associated with each magnetic field $\mathbf{B}$ is a vector potential $\mathbf{A}$ by the relation $\mathbf{B} = \nabla \times \mathbf{A}$. The Schrödinger equation allows for the effects of magnetic fields on charged

---

[112] See Schiff (1968, pp. 177-79) for the derivation. The exact solution for the motion of charge in a uniform magnetic field is given in Landau and Lifshitz (1965, pp. 424-27), but it is unilluminating.

quantum particles by coupling the particles to the magnetic field through the vector potential.[113] The **A** field associated with a uniform **B** field is shown in Figure 3.[114]

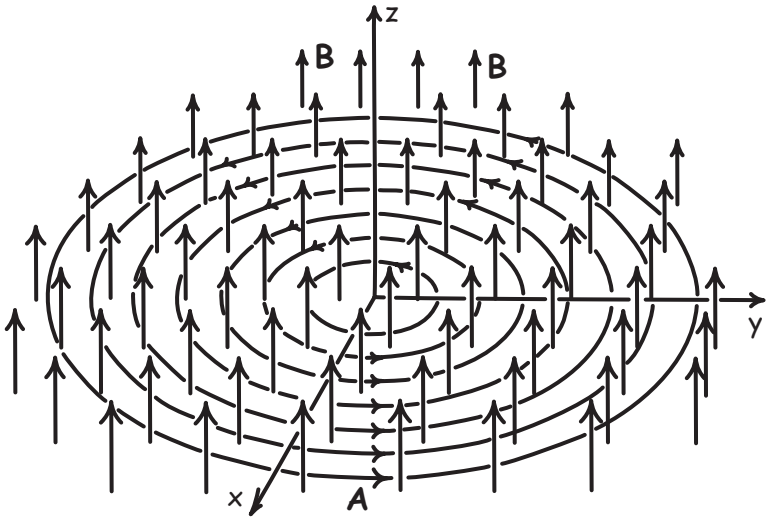


Figure 3. Vector Potential **A** Associated with Constant Magnetic Field **B**.

The integral lines associated with the vector field traces out a circle with the preferred direction of rotation in the figure. When a negatively charged quantum wave packet is coupled to a uniform magnetic field through this vector potential **A**, it will trace out circular trajectories with the same sense of rotation. Its speed, however, is unaltered by the **A** field.

The clue that such coupling is possible is present already in the classical analysis. For one might also ask how a uniform magnetic field can deflect a classical moving charge to the left or the right. If it is uniform, should not both directions be treated alike? They are not treated alike by the magnetic field, once a moving charge is present. A negative charge moving horizontally in an upward pointing magnetic field is deflected to the left, as shown in Figure 2. This derives from the magnetic field vector having what used to be called an "axial" character. That means that it changes sign under mirror reflection of space. (The cross product operator x and curl

---

[113] More precisely, using the "minimal coupling" prescription, the momentum operator $\mathbf{p} = -i(h/2\pi)\nabla$ in Schrödinger's equation is replaced by $\mathbf{p} - e\mathbf{A} = -i(h/2\pi)\nabla - e\mathbf{A}$.

[114] If we align the constant **B** field with the *z*-axis of a Cartesian coordinate system, so that $\mathbf{B} = (0, 0, B_Z)$, then a compatible vector potential is $\mathbf{A} = -(1/2)\ \mathbf{r}\times\mathbf{B} = (1/2)\ B_Z(-y, x, 0)$. Since **A** is determined up to a gauge transformation only, other representations are possible, such as $\mathbf{A} = B_Z(-y, 0, 0)$ in the "Landau gauge." The first however, displayed in Figure 2, preserves the rotational symmetry of the **B** field about the *z* axis and conveys the handedness in the field.

operator $\nabla\mathbf{x}$ has similar transformational properties.) The vector potential $\mathbf{A}$ encodes more clearly how the magnetic field is prepared to deflect charged, moving particles. This preferred sense of rotation of $\mathbf{A}$ will be replicated by the velocity $\mathbf{v}$ of the deflected charge: the velocity $\mathbf{v}$ is linearly related[115] to the $\mathbf{A}$ field by $\mathbf{v} = -2(e/m)\mathbf{A}$. For a negatively charged electron, e is a negative number. Therefore $\mathbf{v}$ and $\mathbf{A}$ agree in direction and relative magnitude.

7. Einstein and the Anomalous Perihelion of Mercury.

In November 1915, an exhausted Einstein was putting the finishing touches onto his general theory of relativity. It was the result of eight years of labor. The final three years had been tense. Einstein had settled upon and published an erroneous version of the theory in 1913. Over the next two years, he had alternated between confidence in the theory and despair over it until he finally found and resolved his errors. In the midst of this resolution, he also found that his theory accounted for a recalcitrant anomaly in planetary astronomy.

According to Newtonian gravitational theory, a planet orbits the sun in an elliptical orbit. The orbit is re-entrant. That means that, in each planetary year, the planet will trace out the same ellipse in space. This familiar results holds exactly only for a two body system of a very massive sun and a single planet. If other planets are present, their gravitational attraction will deflect the original planet's motion away from the re-entrant ellipse. In our solar system, these alterations are very slight and manifest as a very slow rotation of the ellipse of the planet's orbit. In the early twentieth century, careful calculations had accounted for nearly all these motions in the planets. The prominent exception was Mercury. The residual, unaccounted motion of the axis of its orbit was a rotation in the direction of the planet's motion. The planet's perihelion, the point of closest approach to the sun, had an unaccounted advance of roughly 40 seconds of arc per century.[116]

---

[115] For the classical particle, the scalar speed v satisfies $|e|vB_z = mv^2/R$, where R is the radius of curvature of the trajectory. Hence $v = (|e|/m)\, B_z\, R$. The two varying components v_x and v_y of the constant scalar speed v will oscillate harmonically as the charge orbits in a circle. If we locate the origin of the Cartesian coordinates at the center of this circle, we have $\mathbf{v} = (e/m)\, B_z\, (y, -x, 0)$, so that $\mathbf{v}$ is function of position in space. Then $\mathbf{v} = -2(e/m)\mathbf{A}$ follows. Different initial positions and velocities for the charge will locate the center of the orbit elsewhere. The appropriately matched vector potential is recovered by a gauge transformation of the original $\mathbf{A}$ field. To relocate the origin to $(x_0, y_0, 0)$, transform $\mathbf{A}$ to $\mathbf{A}' = \mathbf{A} - (1/2)\, B_z\, (-y_0, x_0, 0) = -(1/2)\, B_z(-(y-y_0), (x-x_0), 0)$. This is a gauge transformation since $\mathbf{B} = \nabla\mathbf{x}\mathbf{A} = \nabla\mathbf{x}\mathbf{A}'$.

[116] The earlier history of this problem is discussed in the chapter on simplicity.

In Einstein (1915), in the passage quoted at the start of this chapter, Einstein reported with delight that his new theory calls for a slight correction to the Newtonian motions that matches exactly this anomalous motion of Mercury. It provides, as the title of the paper asserts, an "Explanation of the Perihelion Motion of Mercury from the General Theory of Relativity."

For our purposes, three aspects of Einstein's claims are important and are developed in the subsections that follow.

7.1 Mere "Confirmation" not "Inference to…"

First, Einstein and subsequent commentators do not carry out a complete inference to the best explanation. They claim only, as Einstein writes (1915, p. 831), "an important confirmation" of the theory. Born's popularization of relativity from the 1920s (1922, p. 254) says: "it [Einstein's theory] is thus already confirmed in advance by Leverrier's calculation [of Mercury's motion]."[117] Pauli (1958, pp. 168-69), in his 1921 authoritative review article, is even more cautious. The question of Mercury arises as a "check by experiment" of consequences of Einstein's theory. The agreement of theory and observation constitutes "a great success."

All these affirmations noticeably fall short of an authorization to infer to the theory, as inference to the best explanation allows. The reason is not hard to find. There is no such authorization perceived in this result. The gap between the theory and observation is too great to be closed completely even by as striking a success as this.

Weyl (1921, p. 247) explains the evidential situation quite well in his celebrated *Space-Time-Matter*, after reviewing general relativity's success with Mercury and in two other astronomical tests.

> …the actual deviations from the old theory are exceedingly small in our field of observation. Those which are measureable have been confirmed up to now. The chief support of the theory is to be found less in that lent by observation hitherto than in its inherent logical consistency, in which it far transcends that of classical mechanics, and also in the fact that it solves the perplexing problem of gravitation

---

[117] The German is "Genau diesen Betrag aber fordert die Einsteinsche Theorie; sie ist daher durch Leverriers Rechnungen bereits im voraus bestätigt." Unfortunately the later English translation (Born, 1962, p. 348) mangles the German and translates this sentence as "But this is just the amount required by Einstein's theory. The confirmation of this result of Einstein's mechanics was therefore actually anticipated by Leverrier's calculation." That is, the translation mistakenly reports the predicted motion of Mercury confirmed, not the theory predicting it, as in the German.

> and of the relativity of motion at one stroke in a manner highly satisfying to our reason.

While we now have more observational and experimental support for general relativity, I believe Weyl's assessment still applies well today. The strongest support for the theory derives from our esthetic appreciation of the theory.

7.2 Preference for the Better Explanation

While the complete "inference to…" is absent, what is present in this example is a quite thorough implementation of the comparative step: the preference for the better explanation. This is embodied in two facts recognized in the literature. First, all other explanations of Mercury's anomalous motions on offer in the literature had been contradicted by the evidence. Second, other explanations might be possible. However the suggestion was these other explanation would likely take on undischarged evidential debt, by, for example, introducing parameters with arbitrarily set values. Einstein's explanation was distinctive in not requiring any arbitrary parameters.

When Einstein announced his successful explanation of Mercury's anomalous motion, it was very convenient that his colleague, the astronomer Erwin Freundlich, had just published an extensive survey of the problem of Mercury's anomalous motion. Einstein (1915, p. 831) cited Freundlich's account in a footnote to this announcement as support for the failure of Newton's theory to offer an explanation of Mercury's anomalous motion:

> E. Freundlich has recently written a noteworthy paper (Astr. Nachr. 4803, Bd. 201 June 1915) on the impossibility of satisfactorily explaining the anomalous motion of Mercury on the basis of the Newtonian theory.

Freundlich's paper listed four ways the astronomers had then tried to explain the anomalous motion. He concluded that none succeeded. That is (1915, p. 51):

> …in the explanation of the existing contradiction between theory and experiment, we have progressed no further than since the time of Newcomb.

His final, concluding sentence (p. 56) is:

> How the anomalies of these inner 4 planets really come about has unfortunately up to now not been answered thoroughly.

Freundlich cited Simon Newcomb, whose study (1895) of the motion of the four inner planets was then authoritative. Newcomb's (1895, Ch. VI) provided an extensive examination of various hypotheses advanced to explain the anomalous motion of Mercury and for smaller anomalies in the other inner planets. Freundlich then provided an update.

The first candidate was the supposition of as yet unknown planets between the sun and Venus. Freundlich deferred to Newcomb's (1895, pp. 112-115) analysis where he considered the possibility of a single planet or multiple planets in a ring. He was unable to find a suitable configuration that would accommodate the known anomalies. The celebrated but failed supposition of the nineteenth century, of a single new planet, Vulcan does not even bear mention by name. The possibility is dismissed by Newcomb with a casual (p. 115) "But I conceive that a planet of the adequate mass could not have remained so long undiscovered."

The second candidate was of a flattening of the sun, presumably as a result of its rotation. The deviations from sphericity would then lead to gravitational effects that could explain the anomalies. The possibility was ruled out, however, since the flattening would have to be much greater to get the desired effect than is compatible with observations of the sun.

The third candidate was a proposal by Asaph Hall (1894) that the force of gravity might not dilute with distance r as an inverse square $1/r^2$ but very slightly faster as $1/r^{2+\delta}$ where δ is a very small number. Newcomb reports that $\delta = 0.0000001574$ would suffice to create the anomalous advance of Mercury's perihelion. The proposal fails, Freundlich notes, since a value of δ sufficiently large to accommodate Mercury's anomalous motion produces effects in our moon's motions that are incompatible with observation and the then successful theory of Brown for the moon's motions.

The fourth candidate was a proposal by Seeliger. The zodiacal light is a halo of light around the sun. It is presumed due to some diffuse distribution of matter that extends as far as the orbit of Mars. The proposal was that the gravitational action of the matter in this halo might account for the anomalous motion of Mercury. The principal content of Freundlich's (1915) paper was to show that this possibility contradicted other evidence of the zodiacal light. His analysis was complicated. Merely finding the mean density of the postulated distribution was not enough. Non-uniformities made a difference. Matter within the orbit of Mercury would produce an advance in the planet's motion; and matter outside its orbit would retard it. Freundlich compared the sorts of densities of matter needed and their distribution with other possible properties of the zodiacal light, including how a distribution of massive dust might impede the motion of the planets, including the Earth. His final conclusion was these other properties forced a much smaller density of matter in the zodiacal light than needed to account for the anomalous motion of Mercury.

In sum, at the start of 1915, all concrete proposals for accounting for the anomalous motion of Mercury had been contradicted by further evidence. Freundlich's analysis leaves open the possibility that there might still be some as yet undiscovered account that explains the anomalous motion of Mercury. There proved to be one theory that could do this. Freundlich's paper was written shortly before Einstein perfected his theory and discovered that it accounted for the anomalous motion of Mercury. Might there be still others? Neither Einstein nor responsible commentators at that time asserted flatly that no other theory could accommodate the anomalous motion of Mercury. However they commonly pointed to a single feature of Einstein's explanation that they deemed of great significance.

Other accounts of the motion of Mercury had all required additional suppositions. If extra masses were invoked, their positions and distributions in space needed to be specified. If alterations to Newton's inverse square law of gravity were invoked, then the alterations would add extra parameters, such as Hall's δ above. Einstein's theory, however, required no such additional hypotheses or parameters. Einstein (1915) points to this at the outset with his remark that the explanation succeeds "…without having to posit any special hypotheses." Pauli (1958, p. 169) notes:

> Compared with Seeliger's explanation, Einstein's has at least the advantage that no arbitrary parameters are needed.

Born (1922, p. 254; 1962, p.348) also remarks:[118]

> This result is of extraordinary importance; for no new arbitrary constants enter into Einstein's formula…

Just how does this feature of Einstein's theory come to favor it? They do not say. However, among the ideas developed in this chapter, there is an obvious reading. The introduction of extra, arbitrary parameters or constants is the taking on of an evidential debt. One must eventually provide independent evidence for them, just as one must find independent evidence that there is a planet Vulcan perturbing the motion of Mercury. Until that is done, Einstein's explanation is better supported in the sense that it has no such undischarged evidential debt.[119]

Hence I take the repeated remark to suggest that any other explanation of the anomalous motion of Mercury is likely to need such extra arbitrary parameters and thus to be weaker than Einstein's. That is, we should not expect a serviceable competitor to Einstein's theory to emerge

---

[118] "Dieses Resultat is von ausserordentlichem Gewichte; denn in die Einsteinsche Formel gehen keine, neuen, willkürliche Konstanten ein…" This time, the English translation (Born, 1962, p. 348) is accurate.

[119] For an account that does not employ the notion of undischarged evidential debt, see Norton (2011).

sooner or perhaps even later. This oblique suggestion is far from a clearly asserted advance from the comparative Step 1. to the absolute Step 2., that is from a preference for the better to the inference to the best. It merely gestures in that direction.

7.3 Why Loveliness as an Explanatory Virtue is Overrated

This particular example enables us to mount an interesting test of a core motivation of inference to the best explanation. The idea is that successful explanations gain inductive support because there is something special in the explanatory relation itself. We saw above that Lipton identified explanatory virtues that would underwrite an inference to the *loveliest* explanation. Of all theories in modern physics, general relativity is distinctive in the praise it receives for its immense conceptual simplicity and scope. It is, by any measure, a *lovely* theory. So we might expect that the inductive support it accrues form its account of Mercury's motion would derive from this loveliness.

We can see quite quickly that loveliness has little to do with the support it accrues. That support depends almost entirely on the failure of competing theories to account for Mercury's anomalous motion. A simple thought experiment reveals just how little the loveliness matters. Imagine that, contrary to history, the nineteenth century astronomers did discover a new planet Vulcan in just the place expected from Mercury's anomalous motion. The discovery would be celebrated as a great triumph of Newtonian physics. It would be a replication of the great success of the discovery of Neptune on the basis of then anomalous motions in the planet Uranus.

In this thought experiment, the tables are turned. The Newtonian theory strains initially to explain the anomaly by taking on the evidential debt of a supposition of a hitherto unseen planet. The Newtonian theory is at a disadvantage. When independent, optical observation finds the planet, however, the evidential debt is discharged and the Newtonian theory prevails. General relativity, however, now finds itself in great difficulty. For the anomaly in Mercury's motion has disappeared, but general relativity still requires an additional advance of the perihelion of 43 seconds of arc per century, beyond what is predicted by the fullest Newtonian account. The observed motion of Mercury, in this fable, now threatens to refute general relativity.

Of course were this fable really to have happened, it is unlikely that this one misadventure would have overturned general relativity. The overall decision would come from a balancing of a greater body of evidence. Mercury's observed motion would weigh against the theory and the loveliness of its treatment of Mercury would have no inductive import at all.

There is a real coda to this fictional tale. In 1918, Hermann Weyl extended Einstein's general theory of relativity in a manner quite in keeping with loveliness of Einstein's original theory. Einstein's theory has incorporated gravity into the metrical structure of spacetime. Weyl now elaborated that structure slightly to allow it to incorporate electromagnetism as well.

Einstein was enthusiastic about the theory as piece of theory and praised it strongly to Weyl in correspondence. However Einstein also saw an empirical problem. According to Weyl's theory atomic emission spectra could not retain sharp lines, in contradiction with experience. The loveliness of the theory could not overcome the observational problem and Einstein opposed the theory.[120]

In sum, loveliness figures prominently in our thought about theories. But its importance is overrated. What matters more is the evidential failures of competitors and, if one's own theory suffers such a failure, loveliness cannot rescue it.

8. Cosmic Background Radiation

The discovery in the 1960s that our universe is permeated with a 2.7 degree Kelvin bath of thermal radiation seems tailor made for an abductive inference. For the thermal radiation is readily explained as the residue of the intense heat radiation of the big bang origin of the universe. The fit is so natural that I used it in the opening section of the last chapter to introduce and motivate the idea of inference to the best explanation. It looks like a safe example for philosophy of science textbooks. Hacking (2001, p. 16) gives it in a paragraph, headed *Inference to the Best Explanation*:

> Each of the arguments we've just look at is an *inference to a plausible explanation.*
>
> If one explanation is much more plausible than any other, it is an *inference to the best explanation.*
>
> Many pieces of reasoning in science are like that. Some philosophers think that whenever we reach a theoretical conclusion, we are arguing to the best explanation. For example, cosmology was changed radically around 1967, when the Big Bang theory of the universe became widely accepted. The Big Bang theory says that our universe came into existence with a gigantic "explosion" at a definite date in the past. Why did people reach this amazing conclusion? Because two radio astronomers discovered that a certain low "background radiation" seems to be uniformly distributed everywhere in space that can be checked with a radio telescope. The best explanation, then and now, is that this radiation is the result of a "Big Bang."

This compressed account makes the inference look all but instantaneous, much as we infer instantly that the slender cables just glimpsed explain the magician's levitation.

---

[120] For a brief account of this episode, see Norton (2000, pp. 153-54)

The reality of the example is more complicated in two ways.

8.1 The Thermal Character of the Radiation

First, the discovery of the cosmic background radiation is routinely attributed to work by Arno Penzias and Robert Wilson (1965) and announced in 1965. They found residual radiation with a cosmic source while measuring radio waves bounced off balloon satellites. While this is celebrated as the moment of discovery, merely finding cosmic radiation is not the inductively potent result. For charged matter is posited in all cosmological theories and such matter readily produces electromagnetic radiation. Without it, the stars cannot shine in the electromagnetic spectrum. To distinguish among the theories, a more distinctive property is needed. That distinctive property is that the radiation has a thermal character with a black body spectrum and, in this case, with the temperature of 2.7 degrees Kelvin, that is 2.7 degrees above absolute zero.

That there should be such thermal radiation was long suspected by cosmologists who worked with the idea of a "big bang" or, as they then preferred to call its radiative part, the "primeval fireball." They included the physics research group of Dicke, Peebles, Roll and Wilkinson, working at Princeton University, not far from Penzias and Wilson's Crawford Hill Laboratory in New Jersey. The group had begun its own efforts to detect the thermal radiation, only to find itself scooped by Penzias and Wilson's chance discovery.

Penzias and Wilson had measured the cosmic radiation at one wavelength only, 7.4 cm. While their results were compatible with black body radiation of a temperature 3.5K +/- 1.0K, it did not establish it. What was needed were measurements taken across a larger range of wavelengths or frequencies to show that the distribution of radiant energy across the range matched the quite precise functional form of the black body curve.

The early history is filled with collections of reports of measurements aiming at establishing this match. Weinberg's (1972) text includes a table (Table 15.1, p. 512) with reports of 31 measurements of various types. He still finds (pp. 516-517) that the discrimination between black body and gray body radiation rests entirely on one type of mountain top radiometer measurement; and that these are contradicted by rocket and balloon borne measurements.

This difficulty, in addition to the second concern below, allowed only a cautious celebration of the result. Weinberg (1972, p. 506) could give only a begrudging summary report: "It is widely, though not unanimously, believed, that the microwave radiation background discovered in 1965 is just this left-over radiation…"

The evidential difficulties were eventually resolved. The definitive results were delivered by NASA's COBE satellite. As an index of the completeness of resolution, we can note that Weinberg's (2008) text leads with the COBE results in the first paragraph of its Preface (p.v):

> November 1989 saw the launch of the Cosmic Background Explorer Satellite. Measurements with its spectrophotometer soon established the thermal nature of the cosmic microwave background and determined its temperature to three decimal places, a precision unprecedented in cosmology.

8.2 Competitors[121]

The second difficulty is that even a thermal spectrum is still not quite distinctive enough to be instantly diagnostic of a primeval fireball. The trouble is that a thermal spectrum arises whenever radiation comes to thermal equilibrium; and there may still be other ways that this spectrum can arise. It is too easily gained.

Once again, this difficulty permitted only measured statements of enthusiasm over the result. Partridge wrote a celebratory survey for the Spring 1969 issue of *American Scientist.* There the cosmic background radiation was offered as something a little less that definitive proof of the big bang, but merely a "new parameter" (1969, p. 39):

> The paucity of data in cosmology explains the excitement generated by the discovery of the cosmic microwave background, which we identify with the primeval fireball in which the Universe originated. The expansion of the Universe has now cooled the fireball to a few degrees Kelvin. Measurements of this isotropically distributed microwave radiation have given us a new parameter in cosmology, the temperature of the radiation field, and also one of the most accurate results of observational cosmology, a figure for the isotropy of the radiation field.

Big bang cosmology has the least difficulty in recovering the thermal spectrum. Even there, the recovery is indirect. In the very early universe, matter and radiation come to thermal equilibrium and a thermal spectrum is thus imprinted on the background radiation. However, as the universe expands, matter and radiation eventually decouple. This happens quite early, when the cosmos has cooled to around 3000K. The photons comprising the cosmic background radiation we measure have propagated to us, unimpeded, from this era. Their origins lie in a distant spherical shell surrounding us, the surface of last scattering. The trouble is that these photons have been underway for much of the history of the entire universe. During that time, their frequencies have been greatly reduced by the cosmological redshift that in turn derives from the expansion of space. Will the greatly red shifted distribution still be thermal? A short calculation and some reasonable assumptions, such as given in Weinberg (1972, pp. 506-507),

---

[121] The discussion here barely touches the range of alternative accounts of the cosmic background radiation that arose in the decades following Penzias and Wilson's measurements. For a survey, see Ćirković and Perović (2018).

show that the effect of the redshift is to preserve the thermal character of the radiation while merely reducing its temperature.

That big bang cosmology can eventually accommodate the thermal spectrum of the cosmic background radiation is not decisive. A long section in Partridge's (1969) survey ("B. But Is It the Primeval Fireball?") grapples with the question of whether the cosmic background radiation could arise by other means. Partridge reviews three other mechanisms. In one, short-lived proposal, Kaufman had sought the radiation in emissions from hot intergalactic plasma. Another due to Layzer, posited as the source dust grains heated during galaxy formation.

Partridge's longest analysis was given to proposals generated in the context of steady state cosmology, then the major competitor to big bang cosmology. This alternative cosmology proposed that the universe has maintained is present state on the large scale for all infinity of time. The universe now and the universe any time in the infinite past look much the same. Steady state cosmology was most directly threatened by the discovery of the cosmic background radiation. For background radiation could not be preserved in a steady state within a universe that has been expanding for infinite time. The cooling and diluting effect of the expansion would eradicate it.

Proponents of the steady state theory, Hoyle, Narlikar and Wickramasinghe, rose to the challenge and sought to account for the radiation within their theory in terms of the reradiation of starlight from interstellar grains. Partridge found severe difficulties for the proposal. Nonetheless, his assessment of the overall evidential situation was qualified to the point of awkwardness as "personal bias" (1969, p. 43):[122]

> Also, it is only fair for me to announce my personal bias in advance: I believe the fireball picture to be consistent with all the experimental data, and to be the simplest theoretical explanation of these data. In making this judgment, and in writing this section, I have kept in mind four questions. Can the suggested model for the background radiation explain its intensity? Can it explain the observed spectrum? Can it explain the isotropy of the radiation? And finally, does it survive the cutting edge of Ockham's razor: is it simple, useful, and not *ad hoc*?

For our purposes, the important point is that the assessment is comparative. It is restricted completely to *Step 1. Preference for the Better Explanation.*

In subsequent literature, the assessments remained comparative, but the comparisons quickly reduced to standard big bang cosmology versus just the flagging steady state theory.

---

[122] I have also given this quote at greater length since it is the only place in reading these early sources in which I found notions of explanation entering explicitly with the word "explanation."

Peebles' 1971 text, *Physical Cosmology*, retains Partridge's hesitancy. In a list of eight points of evidence for cosmology, the sixth reads (1971, p. 26):

> (6) The Universe may contain a Primeval Fireball, blackbody radiation left over from a time when the Universe was dense and hot (Chapter V). If this is substantiated by further measurements it will be direct evidence that the Universe really is expanding and growing less dense, in agreement with the Lemaître cosmology (but not the original Steady State model).

Twenty years later, doubt about the thermal character of the background radiation had gone. However Peebles still made the case for the evidential bearing of the measurements comparatively. The evidence favors big bang cosmology because no other account can accommodate it; and the only other account considered is its old nemesis, steady state cosmology. Peebles (1991, p. 19) wrote:

> The thermal form of the spectrum of this radiation is considered to be almost tangible evidence that the universe expanded from a state considerably denser than it is now, because it is exceedingly difficult to see any other way to make the spectrum so close to thermal. Consider for example the classical Steady State theory, in which the mean density of the universe is constant in time…

Peebles proceeded to dismantle the mechanism through which the steady state theorists sought to replicate the measured cosmic background radiation. Radiation, it is supposed, is created cosmically along with baryons in the steady state cosmology. Its spectrum is shifted to a thermal spectrum by absorption and reemission. This absorption corresponds to a certain degree of opacity of space. But the degree required directly contradicts the observed transparency of space.

The same, comparative assessment is repeated at greater detail in Peebles' later 1993 authoritative text, *Principles of Physical Cosmology*, pp. 203-206. He concluded for the absorptive mechanism (1993, p. 204):

> The point of this calculation is that if the universe were postulated to be opaque enough at radio wavelengths to have caused the radiation background to relax to the observed very nearly thermal spectrum of the CBR, space would be predicted to have been too opaque to have allowed the observations of distant radio sources…

Peebles then examined the character and density of dust needed for the relaxation mechanism with results once again unfavorable to the proposal.

This comparative assessment seems now to have acquired the status of the standard textbook formulation of the evidential import of the cosmic background radiation. Here it is in a more recent cosmology textbook: (Liddle, 2003, p. 80)

> The Hot Big Bang theory therefore gives a simple explanation of this crucial observation. In the Steady State theory, all radiation is supposed to originate in stars

> and so is at high frequency and is not a perfect black-body; one has to resort to a thermalizing mechanism such as whiskers of iron, which somehow managed to thermalize this into low-energy radiation in the recent past without preventing us from seeing distant objects. It has never been satisfactorily demonstrated that this can be achieved even allowing the *ad hoc* assumptions that the Steady State scenario requires.

8.4 Success Through Failure of the Competitors

The measurements of the cosmic background radiation do provide good evidence for big bang cosmology. This review brings into sharper focus how they do it. The accounts above of the success identify no special explanatory relation beyond mere accommodation. Big bang cosmology, with suitable auxiliary assumption, entails the existence of a thermal radiation background. Beyond this accommodation, there is no special explanatory coup through which we can make some philosopher's notion of explanation central to the evidential relation. Thermal radiation is something that can arise easily in any account that hosts energized charged matter and sufficient time for thermal equilibrium to be established. Nothing in the analysis provided by big bang cosmology indicates that it is the only theory that can accommodate the result.

Nonetheless, this exclusivity does turn out to be evidentially decisive. It is not established by examining how big bang cosmology explains the cosmic background radiation. Rather it is established by examining how competitors to big bang cosmology fail to accommodate the result. The decisive fact is not so much about big bang cosmology, but about its competitors. Big bang cosmology can accommodate the result, where no known competitor can. Big bang cosmology wins the day by default.

The explicit discussion of evidential import is restricted to this comparative result, fully within *Step 1. Preference for the Better Explanation.* of the present account. The second step, acceptance that the evidence supports big bang cosmology specifically and absolutely, is left tacit. That big bang cosmology bests its strongest competitor, the steady state theory, is stressed and, presumably, this victory is intended to lead us to believe that there is no better alternative possible.

In any case, over half a century after Penzias and Wilson's observation, the origin of the cosmic background radiation is no longer open to serious dispute in the cosmology literature: it is described without apology or qualification as a thermal residue of an early hot universe. Serious consideration is now given to the slight deviations from isotropy in the radiation, for they are now the key to understanding structure formation in cosmology.

9. Oxygen and Phlogiston

9.1 The Theories Compete

The establishment of Lavoisier's oxygen chemistry has been presented as a canonical instance of inference to the best explanation. A closer look will show that an intrinsic explanatory virtue had little to do with the establishment of the theory. Rather the decisive inferences of both Step 1 and Step 2 were warranted by a quite specific fact: that matter has weight.

Oxygen chemistry ascended in the late eighteenth century, when Lavoisier's oxygen theory competed with the phlogiston theory as the correct account of many chemical processes. Combustion illustrates the competition. The oxygen theory portrayed the combustion of a metal as its combination with oxygen from the air to form an oxide, then commonly called a "calx."

metal + oxygen → calx

The phlogiston theory took all metals to be a compound of a calx and phlogiston; and the combustion of a metal to be the decomposition of this compound into a calx and liberated phlogiston.

metal → calx + phlogiston

There was a close similarity of structure in oxygen and phlogiston chemistry. Just about any reaction accommodated by one was mirrored by a corresponding reaction in the other. To see how the reactions of each theory pair up, you merely need to think of phlogiston as a kind of "anti-oxygen." Then you can convert a reaction of oxygen chemistry into one of phlogiston chemistry and vice versa. In the phlogiston combustion reaction, for example, substitute anti-oxygen for phlogiston; and then move it from the right-hand product side to the left-hand reactant side, dropping the "anti" prefix. What results is the oxygen combustion reaction. Much of oxygen and phlogiston chemistry were mirror images of each other.

Thagard (1978) presents the triumph of oxygen theory as a canonical case of inference to the best explanation. He quotes his translation of a confident assertion by Lavoisier (pp. 77-78) in support:

> I have deduced all the explanations from a simple principle, that pure or vital air is composed of a principle particular to it, which forms its base, and which I have named the *oxygen principle*, combined with the matter of fire and heat. Once this principle was admitted, the main difficulties of chemistry appeared to dissipate and vanish, and all the phenomena were explained with an astonishing simplicity.

While we know that, in the long run, oxygen will win, the situation at the time of the debate was not so clear. Precisely because oxygen and phlogiston chemistry were, to a large measure,

intertranslatable, the two theories had considerable overlap in scope. It was not clear that oxygen's explanatory powers were greater. Thomas Kuhn made this fact a celebrated debating point in the question of the cumulativity of science, when he used it to illustrate what is now called "Kuhn loss" (1996, pp. 99-100):

> The much-maligned phlogiston theory, for example, gave order to a large number of physical and chemical phenomena. It explained why bodies burned--they were rich in phlogiston--and why metals had so many more properties in common than did their ores. The metals were all compounded from different elementary earths combined with phlogiston, and the latter, common to all metals, produced common properties. In addition, the phlogiston theory accounted for a number of reactions in which acids were formed by the combustion of substances like carbon and sulphur. Also, it explained the decrease of volume when combustion occurs in a confined volume of air the phlogiston released by combustion "spoils" the elasticity of the air that absorbed it, just as fire "spoils" the elasticity of a steel spring.

Whatever its other explanatory virtues, oxygen chemistry could not provide an explanation for the common properties of metals, as could phlogiston chemistry.

9.2 Weight and Levity

What turned the tide in oxygen's favor and formed the basis of Lavoisier's case for oxygen was weight. When a metal burned to form a calx, the calx weighed more, while the air above lost 1/6th of its volume; and when the calx was reduced back to metal, in the case of mercury calx, it lost weight and returned just the missing portion of air. These gains and losses of weight could be explained by the phlogiston theory if we assume that phlogiston had negative weight, that is, "levity" as opposed to "gravity." It now seems a curious assumption, but it saves the phenomena. When a metal forms a calx, it loses the levity of phlogiston. This is a loss of a negative weight. Taking away a negative has the effect of adding a positive. It results in a calx that weighs more that the metal.

Phlogiston chemistry fails if we deny the admissibility of levity and insist on the background fact that matter must have weight. John Herschel summarized the failure, writing a few decades later (1840, p. 301):

> So far as weight is concerned, it makes no difference whether a body having weight enters, or one having levity escapes; but there is this plain difference in a philosophical point of view, that oxygen is a real producible substance, and phlogiston is no such thing: the former is a *vera causa*, the latter an hypothetical being, introduced to account for what the other accounts for much better.

More picturesquely, Herschel characterized the question of weight as the crucial factor in deciding between the two: (p. 300; Herschel's emphasis)

> …of two possible roads the wrong was chosen; and a theory obtained universal credence on the credit of great names, ingenious views, and loose experiments, which is negatived, *in every instance*, by an appeal to the balance.

His language is reminiscent of Bacon's "crucial instances," which Bacon had described with an analogy to signposts directing us at branches in a road.[123]

Herschel's account leaves unsupported his conclusion that levity-bearing phlogiston cannot be a real substance. William Whewell (1847, pp. 409-11) lays out a more elaborate case. He too based the decision in favor of oxygen chemistry in this fact about matter. The levity of phlogiston was "rejected by all the sounder philosophers," he wrote, and "It is assumed, it appears, that all matter must be heavy…" He proceeded to attempt a quite general argument that deduces the heaviness of matter from the very idea of substance. One part of his argument returns to phlogiston:

> For if weight is not the criterion of the quantity of one element, phlogiston for instance, why is weight the criterion of the quantity of any other element? We may, by the same right, assume any other real or imaginary element have levity instead of gravity; or to have a peculiar intensity of gravity which makes its weight no index of its quantity.

We can now reassess just how the decision in favor of the oxygen theory was taken. While Lavoisier had boasted of the explanatory prowess of this oxygen theory, at the time of the decision there was little to choose between the explanatory capacities of oxygen and phlogiston chemistries. What was decisive, however, was a fact: matter has weight. That fact was compatible with oxygen chemistry but not with phlogiston chemistry, in so far as phlogiston was supposed to be material.[124]

Once again we see the two-step structure emerging for the inference. The first step is comparative between oxygen chemistry and the foil of phlogiston chemistry. The decision is not derived from some superior, intrinsic explanatory virtue in the favored oxygen explanation. Rather, the foil is rejected because of a logical incompatibility with a background fact: matter has weight. Oxygen chemistry thereby prevails.

This same fact mediates in the second step: that we should not just prefer oxygen chemistry over phlogiston chemistry, but that we should infer to and accept oxygen chemistry.

---

[123] A portrait of Francis Bacon is on the title page of Herschel's *Preliminary Discourse.*

[124] I pass over the rather great awkwardness for Lavoisier that he had also allowed caloric, the matter of heat, into his table of elements, even though no sensible weight for it had been found.

There are many component inferences and further factual assumptions required for the second inference. But the course of each component is unremarkable. The full accounting would need to look at many different chemical changes. Here is how one proceeds. When a metal calx transforms into a lesser weight of metal and a released gas, we read directly that this is a decomposition of the calx into its constituent metal and the gaseous component, oxygen. The further assumptions needed to make this inference from the observation a few instances to the generality would include: that matter is conserved, so that any weight lost must reappear in the matter of the gas; and that the calx is a pure substance all of whose samples have the same properties. Then the behavior of one sample can stand for all.

10. The Wave Theory of Light

Darwin (1876, p. 421) indicated the wave theory of light ("undulatory theory") as one established by the same abductive methods as he used in *Origin of Species*. Thagard (1978, pp. 77-78) includes it as one of his canonical scientific examples of inference to the best explanation. As a result, one might expect that it would be straightforward to reconstruct the abductive inference. Matters prove otherwise. The wave theory evolved slowly into its modern form, only gradually acquiring evidential support in a temporally extended process of great complexity. While the fuller evidential case cannot even be sketched here, we can see enough of it to know that it conforms to the pattern already seen. The two step character is present. The explanatory prowess of the wave theory was almost invariably compared with the foil of Newton's corpuscular theory, which gave it real competition. The latter was vanquished eventually either by its need take on undischarged evidential debt or by direct contradiction with experiment. The second step long remained fraught. At any moment, the explanatory achievements of the wave theory were threatened by new, as yet unexplained optical phenomena.

The complexity of the example derives from a pair of coupled circumstances.

First, *the* wave theory of light is a misnomer. There is a long history of theories that attribute wavelike properties to light, extending back to the seventeenth century in the work of Hooke and Huygens. However the theories adopt many forms as they develop, sometimes adapting to then current developments in the surrounding sciences. The earliest theories simply presumed light to be a propagation in some medium, akin to sound propagation in air. Later theories retracted, for sound waves are longitudinal rarefactions and compressions, whereas light waves proved to be an oscillation that was transverse to the direction of propagation. Ultimately, light was absorbed into electromagnetic theory as the propagation of a wavelike disturbance in the electromagnetic field.

Second, the behavior of light was examined carefully in many experiments. As a result, the range of experimental results to be accommodated by a theory of light was large and growing. They includes results on the speed and direction of light propagation, its decomposition into colors, reflection, refraction in media, colored bands in thin plates ("Newton's rings"), the polarization of light, stellar aberration, various interference patterns including fringes around shadows, double refraction in crystals, and more. The character of the wave motion attributed to light developed in concert with these developments.

The history of the establishment of the wave theory of light is a history of its competition with the Newtonian corpuscular theory, also known as the emission theory. The competition was quite real. In the seventeenth century, the wave theory was rudimentary. It was based, according to Huygens (1690, p. 11), on the supposition that light is "some motion impressed upon the matter which lies in the intervening space" and that the motion "is propagated, as that of sound, by surfaces and spherical waves." The explanatory successes of Huygen's theory are now well known. His constructions enable recovery of familiar processes of reflection and refraction.

10.1 Early Competition of Wave and Emission Theory

Huygen's theory faced considerable explanatory competition from Newton's corpuscular theory. The latter theory merely supposed that light consists of very small corpuscles, moving very quickly. The theory was ontologically frugal. Both posited the existence of matter. For the corpuscular theory, the matter posited just was the light seen. For the wave theory, vastly more matter needed to be supposed in the form a space-filling, all-pervading substance in which light would propagate as vibrations.

Newton's theory could deal quite effectively with these same phenomena as the wave theory. There, Newton's theory had advantages. Light propagates in straight lines. Wave propagations in media, such as sound, do not propagate linearly but follow tortuous pathways according to alterations in the medium and its motion. This problem, according to Shapiro (2002, p. 232) remained Newton's principal objection to the wave theory throughout his life. There were other explanatory advantages for the corpuscular theory. The equal angles of reflection of light matches perfectly with the behavior of bodies undergoing elastic collision. Newton had found that white light decomposes into rays of definite colors and that these rays were quite fixed in their color. It was not altered by reflection, refraction and other like processes. That constancy was easily accommodated into a corpuscular theory by assuming that the different colors correspond to different types of corpuscles with stable characters. It was less clear that mere vibrations in some unseen, all-pervading substance could provide the same stability.

10.2 The Emission Theory Weakens

The tide began to turn against the Newtonian theory with the work of Thomas Young in the early nineteenth century and then its development by Augustin Fresnel. They were able to account for many optical effects as arising from the constructive and destructive interference of light waves. Newton's theory could accommodate such effects to some extent. The most celebrated of these effects was "Newton's rings," that is, rings of light and dark that form in the small, intervening space when a lens sits on a flat sheet of glass. Newton's account was complicated, depending on "fits of easy transmission and reflection."

The details are too complex for recapitulation here. What is relevant, however, is William Whewell's (1858, p. 89) assessment of them in his *History of the Inductive Sciences*, written from the perspective of someone close to the episode. In spite of Newton's status as a national hero, Whewell was quite scornful of Newton's hypotheses (p. 89, Whewell's emphasis):

> The colors of thin plates. Now, how does Newton's theory explain these? By a new and special supposition;--that of *fits of easy transmission and reflection*: a supposition, which, though it truly expresses these facts, is not borne out by any other phenomena. But, passing over this, when we come to the peculiar laws of polarization in Iceland spar, how does Newton's meet this? Again by a special and new supposition;--that the rays of light have *sides*. Thus we find no fresh evidence in favor of the emission hypothesis springing out of the fresh demands made upon it.

In present terms, the problem was not that Newton's account was incompatible with experiment. Rather it required an undischarged evidential debt in the form of the hypotheses identified by Whewell.

One might imagine that the explanatory advantage of the wave theory was absolute by this time. However that was not so. It still did require a medium of unusual properties. Since light propagates in empty space, that medium—the luminiferous ether—must be all pervasive. It must be entirely unaffected when ordinary matter is evacuated from vessel, where such evacuation would completely suppress sound propagation. Tyndall (1873, pp. 47-48) could report as late as 1873 of the persistence of doubt over this assumption of the medium. He wrote of David Brewster (1781-1868), a celebrated pioneer in optical science:

> In one of my latest conversations with Sir David Brewster he said to me that his chief objection to the undulatory theory of light was that he could not think the Creator guilty of so clumsy a contrivance as the filling of space with ether in order to produce light.

10.3 Wave Theory Triumphs

Thus the competition proceeded. It is quite hard to locate simple cases of explanatory competition between the emission theory and a wave theory of light, suitable for a brief exposition here. Lloyd (1873, pp. 11-12) reports one such case. By the time of Lloyd's writing, it had been ascertained experimentally that the speed of light is the same everywhere in empty space, whatever the source of the light. Lloyd found it incredible that all the different processes that accelerate the corpuscles into propagating light should produce exactly the same speed. More puzzling is that they could retain that speed when the gravity of celestial objects would slow them down. He reported Laplace's computation that the gravity of a star 250 times as great as our sun, but of the same density, would stop the motion entirely. There was a desperate rescue possible:

> The suggestion of M. Arago seems to offer the only way of escaping the force of this objection. It may be supposed that the molecules of light are originally projected with different velocities, but that among these velocities there is but one which is adapted to our organs of vision, and which produces the sensation of light.

The constancy of the speed of light, however, followed naturally if light is a wave propagating in a medium. The speed depends only on the elasticity and density of the medium, which are assumed constant.

We see in this simple example that the wave theory accommodates the constancy of the speed of light fairly well. The accommodation is dependent on a special hypotheses, the uniformity of the medium. Since the constitution and nature of the medium remained uncertain, the wave theory account is not problem free. The emission theory, however, is in great trouble. Any reasonable mechanics of the era for corpuscles predicts many speeds. The fact that only one is observed is a refutation. The emission theory can be protected, but only by taking on a dubious hypotheses about our vision; that is, by taking on a quite significant evidential debt.

A decisive turning point came with experiments around 1850 that directly measured the speed of light in media. When light propagates from a less dense to a more dense medium, it is refracted towards the denser medium. This familiar effect is the basis of the functioning of optical lenses. It is explained quite differently by the wave and emission theories. The wave theory assumes that the speed of light in the denser medium is reduced and the angle of refraction is recovered by a Huygens construction. The emission theory, however, explains the refraction towards the denser medium by attractive forces that accelerate the light corpuscles into the denser medium. That is, the speed of light increases in a denser medium.

This stark difference of prediction was finally put to the test. The wave theory prediction was borne out. Crew, in 1900, reports the victory (1900, p. xii)

> It was in the year 1850 that Fizeau and Foucault measured directly the speed of light in air and water, and found the ratio of these speeds numerically equal to the ratio of their refractive indices. This experiment has sometimes been called the *experimentum crucis* of the wave theory; but with scant justice we venture to think, inasmuch as no great doctrine in physics can be said to rest upon any single fact, though modification may be demanded by a single fact.

Crew's caution was prudent. While this result may have ended the emission theory's prospects, the wave theory of 1850 still had obstacles to overcome. Its dependence on a medium of uncertain properties, the luminiferous ether, would fester and eventually become a focus when Einstein published his special theory of relativity in 1905.[125]

By this time, the wave theory of light was no longer an independent theory that would rise and fall according to new experimental results on light alone. Since the 1860s, light had been identified as a wave propagating in an electromagnetic field, so that the success or failure of the wave theory became intimately tied to that of electromagnetic theory. A fuller account of the final victory of the wave theory would have to include an account of the rise of electromagnetic theory upon which it came to depend.

By the turn of the century, the complex, lingering competition between emission and wave theories of light was reducible to a few brief sentences in the opening pages of a textbook. Walker (1904, pp. 1-2) summarizes it as:

> …the emission theory is lacking in simplicity, and overcrowded with hypotheses; moreover it contradicts the facts in an important particular, for it leads to the result that the propagational speed of light is greater in a dense medium, such as water, than it is in air, whereas direct experiments show that the reverse is the case.

This summary serves us quite well, for it encapsulates failures of the Newtonian foil in Step 1 of the abductive inference. It is defeated by the undischarged evidential debt of special hypotheses and by contradiction with experiment. The second step, the elevation of the wave theory from the better explanation to the best and the one to which we infer, is too complex to gloss here.

11. Conclusion

The standard philosophical account of this argument form tells us that we may infer to some hypothesis or theory because that hypothesis or theory displays some powerful and

[125] More relevantly, Einstein also published a startling result in 1905 concerning light. His light quantum hypothesis asserted that the energy of high frequency light was spatially localized into points; and that was quite reminiscent of Newton's tiny corpuscles.

distinctive explanatory prowess. This chapter has examined canonical examples in real science of inferences to the best explanation and finds something different. The favored theory or hypothesis does not gain favor because it implements some philosophically distinctive notion of explanation. The evidential successes are more successes of accommodation, albeit at times noteworthy. The real evidential challenge for proponents of the favored hypothesis or theory is to display the evidential failure of competitors. The favored theory or hypothesis does not so much prevail because of its own intrinsic virtue. It prevails by default because of the evidential failure of the competitors. The failures of the competitors are not explanatory failures. They are simpler and consist in two modes. Either the competitor is contradicted by the evidence; or its survival requires it to take an evidential debt, that is, to make suppositions for which there is insufficient evidential support.

As a result, it was possible in the last chapter (Section 7) to characterize these inferences in loose and general terms as "inference to the best explanation without explanation." The emphasis in the examples on comparison led the characterization to have two steps. The first and dominant step is comparative: one hypothesis or theory is favored over the competing foil. This step is clearly visible in the examples recounted here. The second step is logically very strong. It dispenses with comparisons: "favoring" is replaced by "inferring to." However it has little explicit presence in the examples. The step, if taken at all, is made tacitly. The competing foils are defeated and that is enough to let the victor ascend.

References

Born, Max (1922) *Die Relativitätstheorie Einsteins und ihren physikalischen Grundlagen*. 3rd ed. Berlin: Julius Springer.

Born, Max (1962) *Einstein's General Theory of Relativity.* Methuen, 1924; revised and enlarged, New York: Dover, 1962.

Ćirković, Milan M. and Perović, Slobodan (2018) "Alternative Explanations of the Cosmic Microwave Background: A Historical and an Epistemological Perspective," *Studies in History and Philosophy of Modern Physics*, **62**, pp. 1-18.

Crew, Henry (1900) *The Wave Theory of Light: Memoirs by Huygens, Young and Fresnel*. New York: American Book Co.

Darwin, Charles R. (1876). *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. 6th ed. London: John Murray.

Davisson, Clinton J. and Germer, L. H. (1927) "Diffraction of Electrons by a Crystal of Nickel," *Physical Review*, **6**, pp. 705-740.

Einstein, Albert (1915) “Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie,” *Preussische Akademie der Wissenschaften, Sitzungsberichte*, 1915 (part 2), pp. 831–839.

Einstein, Albert (2000) "'Nature in the Realization of the Simplest Conceivable Mathematical Ideas[1]: Einstein and the Canon of Mathematical Simplicity," *Studies in the History and Philosophy of Modern Physics*, **31**, pp.135-170.

Freundlich, Erwin (1915) “Über die Erklärung der Anomalien im Planeten-System durch die Gravitationswirkung interplanetaren Massen,” *Astronomische Nachrichten*, **201**, No. 4803, pp. 51-56.

Hacking, Ian (2001) *An Introduction to Probability and Inductive Logic*. Cambridge: Cambridge University Press.

Hall, Asaph (1894) “A Suggestion in the Theory of Mercury,” *The Astronomical Journal.* **14**, pp. 49-51.

Harper, William (2002) “Newton’s Argument for Universal Gravitation,” Ch.5 in I. B. Cohen and G. E. Smith, eds., *The Cambridge Companion to Newton.* Cambridge: Cambridge University Press.

Herschel, John (1840) *Preliminary Discourse on the Study of Natural Philosophy*. New edition, London: Longmans.

Hertz, Heinrich (1883) “Versuche über die Glimmentladung,” *Annalen der Physik,* **19**, pp. 782-816.

Huygens, Christian (1690) *Treatise on Light* in H. Crew (1900).

Kuhn, Thomas S. (1996) *The Structure of Scientific Revolutions*. 3rd ed. Chicago: University of Chicago Press.

Landau, Lev D. and Lifshitz, Evgeny M. (1965) *Quantum Mechanics: Non-Relativistic Theory.* 2nd ed. Trans. J. B. Sykes and J. S. Bell, Oxford: Pergamon.

Lenard, Philipp (1894) “Über Kathodenstrahlen in Gasen von astmosphärischen Druck; und im äussersten Vacuum,” *Annalen der Physik*, **51**, pp. 225-267.

Lenard, Philipp (1894a) “Über die magnetische Ablenkung der Kathodenstrahlen,” *Annalen der Physik,* **52**, pp. 23-33.

Lenard, Philipp (1906) *Über Kathodenstrahlen: Nobel-Vorlesung*. Leipzig: J. H. Barth.

Liddle, Andrew (2003) *An Introduction to Modern Cosmology*. 2nd ed. West Sussex: John Wiley & sons.

Lipton, Peter (2004) *Inference to the Best Explanation.* 2nd ed. London: Routledge.

Lloyd, Humphrey (1873) *Elementary Treatise on the Wave-Theory of Light.* London: Longmans, Green and Co.

Lyell, Charles (1830, 1832, 1833) *Principles of Geology*. Vol. 1, 2, 3. London: John Murray.

Newcomb, Simon (1895) *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*. Washington: Government Printing Office.

Norton, John D. (2000) "How We Know About Electrons," pp. 67- 97 in R. Nola and H. Sankey, eds., *After Popper, Kuhn and Feyerabend; Recent Issues in Theories of Scientific Method*. Dordrecht Kluwer.

Norton, John D. (2011)"History of Science and the Material Theory of Induction: Einstein's Quanta, Mercury's Perihelion." *European Journal for Philosophy of Science*. **1**, pp. 3-27.

Pauli, Wolfgang (1958) *Theory of Relativity*. Oxford: Pergamon.

Partridge R. Bruce (1969) "The Primeval Fireball Today," *American Scientist*, **57**, pp. 37-74.

Peebles, P. J. E. (1971) *Physical Cosmology*. Princeton: Princeton University Press.

Peebles, P. J. E. (1991) "The Emergence of Physical Cosmology," pp. 17- 30 in A. Blanchard, et al., eds., *Physical Cosmology*. Gif-sur_Yvette, France: Éditions Frontières, 1991.

Peebles, P. J. E. (1993) *Principles of Physical Cosmology*. Princeton: Princeton University Press.

Penzias, Arno and Wilson, Robert (1965). "A Measurement Of Excess Antenna Temperature At 4080 Mc/s," *Astrophysical Journal Letters*. **142**, pp. 419–421.

Shapiro, Alan E. (2002) "Newton's Optics and Atomism," pp. 227-255 In I. B. Cohen and G. E. Smith, eds., *The Cambridge Companion to Newton*. Cambridge: Cambridge University Press.

Schiff, Leonard I (1968) *Quantum Mechanics* 3rd ed. Tokyo: McGraw-Hill Kogakusha.

Thagard, Paul R. (1977) "Discussion: Darwin and Whewell," *Studies in History and Philosophy of Modern Physics*, **8**, pp. 353-56.

Thagard, Paul R. (1978) "The Best Explanation: Criteria for Theory Choice,"*Journal of Philosophy*, **75**, pp. 76-92.

Thomson, Joseph J. "Presidential Address." pp. 699-706 in Section A. Mathematical and Physical Sciences. *Report of the Sixty-Sixth Meeting of the British Association for the Advancement of Science Held at Liverpool in September 1896.* London: John Murray, Albemarle St.

Thomson, Joseph J. (1897) "Cathode Rays," *Philosophical Magazine,* **44**, pp. 293-316.

Thomson, Joseph J. (1906) "Carriers of Negative Electricity," Nobel Prize Lecture, December 11, 1906.

Tyndall, John (1873) *Six Lectures on Light: Delivered in America in 1872-1873*. London: Longmans, Green, and Co.

Walker, James (1904) *The Analytic Theory of Light*. Cambridge: Cambridge University Press.

Whewell, William (1847) *The Philosophy of the Inductive Sciences*. Vol. 1. London: John W. Parker.

Whewell, William (1858) *History of the Inductive Sciences from the Earliest to the Present Time*. Vol. II. 3rd ed. New York: D. Appleton and Co.

Weinberg, Steven (1972) *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. New York: John Wiley and Sons.

Weinberg, Steven (2008) *Cosmology*. Oxford: Oxford University Press.

Weyl, Hermann (1921) *Space-Time-Matter*. 4th ed. Trans. H. L. Brose. London: Methuen.

August 7, November 19, 2017
Minor typo corrections June 20, November 9, December 5, 2018
February 26, December 22, 2019
footnote added January 26, 2020

Chapter 10

Why Not Bayes

0. Prelude

A central proposition of this book is that there are no universal rules for inductive inference. The chapters so far have sought to argue for this proposition and to illustrate it by showing how several popular accounts of inductive inference fail to provide universally applicable rules. Many in an influential segment of the philosophy of science community will judge these efforts to be mistaken and futile. In their view, the problem has been solved, finally and irrevocably.

This segment of the community are "Bayesians" who work in what has come to be called "Bayesian epistemology." Its central idea is that issues of belief and inductive inference are to be treated solely by means of the probability calculus. The central structure is a conditional probability measure $P(A|B)$, the probability of proposition A against the background proposition B. The term "Bayesian" derives from an easily proven theorem in the probability calculus, Bayes' theorem:

$$P(H \mid E \,\&\, B) = \frac{P(E \mid H \,\&\, B)}{P(E \mid B)} P(H \mid B)$$

It provides the central engine for inference in Bayesian epistemology. The inference starts with some prior belief or inductive strength of support for an hypothesis H on the background B, $P(H|B)$. Learning evidence E leads the prior probability to be updated to the posterior probability $P(H|E\&B)$, which now incorporates the full import of the evidence E. This posterior probability is computed via Bayes' theorem using the auxiliary quantities, the "likelihood" $P(E|H\&B)$ and the "expectedness" $P(E|B) = P(E|H\&B)\ P(H|B) + P(E|\text{-}H\&B)\ P(\text{-}H|B)$.

This is the barest sketch of the core notions of the Bayesian approach. They are now so widely know as not to require further elaboration. There is much more to the general Bayesian approach and there are many variant forms. Recalcitrant cases that do not easily fit with the core notion in its bare form are treated by "imprecise probabilities." The imprecision derives from

replacing a single probability measure by a set of measures; or by replacing an additive measure by a superadditive measure. These are conceived as providing a generalized probabilistic analysis. Recently, Bayesian epistemology has been subsumed under the new heading of "formal epistemology." Its leading idea is that epistemic problems are to be addressed by formal and mathematical methods. There is little real change, however, as far as belief and inductive inference is concerned. Probability measures remain the principal instrument used to treat them.

For my purposes, the core commitment of this tradition resides in a single idea:

It's all probabilities.

Here just what it is to be probabilistic can be construed differently according to one's interpretive inclinations. However that general conception is taken to solve the essential problems addressed in this book. There are universal rules, this tradition holds. They are axioms of the probability calculus. Once that is recognized, all that remains are the finer details of determining just how they are to be applied to each problem. The big problem is solved.

The purpose of this and the following chapters is to explain why I am dissatisfied with this Bayesian solution.

1. Introduction

The case against the universality of probabilities will be made in this chapter in two parts. The first part will apply the general argument developed in Chapter 2 against the idea that any calculus, probabilistic or otherwise, can be universally applicable. Its core resides in the following:

> *Any logic of induction must restrict what happens in ways that go beyond logical consistency. Hence a logic of induction is applicable in some domain if the facts of that domain match the factual restrictions of the logic of induction. Since there is no universally applicable factual restriction, in general, different domains require different inductive logics.*

This argument will be developed more fully in Sections 3 and 4. It concludes that any calculus of induction must eventually reach a boundary to its domain of applicability beyond which it fails. I will argue in Section 5 that efforts to develop theories of imprecise probabilities are misplaced attempts to disguise these boundaries. They use an additive measure merely as adjuncts to simulate the non-additive inductive logic of a new domain. In foundational terms, they mislead by fostering the impression that "it's all probabilities" even when the logic simulated is inherently non-additive.

As a foil for further analysis, Section 6, 7 and 8 will present an extreme but simple example of such a non-additive inductive logic. It is the relation of "completely neutral support,"

which is derived from the principle of indifference and illustrated by von Mises' example of different mixtures of wine and water. Section 9 reviews the extent to which theories of imprecise probability can accommodate completely neutral support. In so far as they do not accommodate it, they are inadequate; in so far as they do, they are superfluous. Any success in this one case merely postpones the inevitable failures, I urge, that must arise when they seek to accommodate more exotic logics.

The second part of the case against the universality of probabilities is a general rebuttal of the many proofs offered in the literature as demonstrating the necessity of probabilities. These proofs come in different guises. One of the oldest and best known is Ramsey and de Finetti's Dutch book argument. It is used to infer than non-probabilistic distributions of belief are "incoherent," which is a form of irrationality. All such proofs must fail and they must fail in the same way, for this reason:

> *A proof of necessity of probabilities is a deductive argument whose premises must be at least as strong logically as the conclusion. Therefore the assumptions of the proof must already presuppose the necessity of probabilities or something logically stronger. Hence by dominance we are better off simply presupposing the necessity of probabilities at the outset and forgoing the proof.*

This general argument is developed in Sections 10 and 11. It is used to predict that a careful analysis of a proof of the necessity of probabilities triggers a regress of reasons. For the assumptions used in the proof will always be found to be improperly grounded. Attempts to provide proper grounding will require new assumptions that will then also prove to be improperly grounded.

The principal illustration of this circularity and the ensuing regress of reasons will be the recent efforts to vindicate probabilities by means of notions of accuracy and scoring rules. Since the analysis is quite extensive, it is postponed to the next chapter. To show that other attempts at vindication fail in the same way, two more examples are given briefer treatment in this chapter. Section 12 recalls the Dutch book argument. It identifies which assumptions already have the axioms of the probability calculus built into them and recounts the failure of attempts to remove the circularity. Section 13 repeats this analysis for a different approach developed by Cox (1961) and Jaynes (2003). In it, necessary conditions are identified for strengths of inductive support and from them the computational rules of the probability calculus are recovered by functional analysis. We shall see that the necessity of the conditions requires further grounding, triggering the now familiar regress of reasons. Conclusions in Section 14 suggest further directions of exploration.

As a preliminary, in Section 2, I will distinguish objective from subjective approaches and apologize to the reader for not always distinguishing them clearly as the chapter unfolds.

Finally before proceeding I would like to give Bayesian epistemology its due. My view is far from a complete dismissal of Bayesian epistemology. I view it in the same way as I view all other candidate logics of induction. Whether it applies in some domain is determined by the background facts of the domain. These background facts will also determine the variety of Bayesianism applicable. Stronger facts will authorize strict Bayesianism in which inductive support or, subjectively speaking, beliefs are measured by a single probability measure. Weaker facts will authorize a relaxed Bayesianism in which these supports are represented by sets of probability measures or upper and lower bounds. There are many domains in which varieties of Bayesian analysis are authorized and can be applied. In them, it provides a wonderful instrument.

It is formally precise where other accounts flounder. Arguments from analogy struggle to separate the strong from the weak analogies. Accounts that reward simplicity cannot provide a clear and unobjectionable notion of simplicity whose measure translates mechanically into inductive strength. In contrast, once the probability space is well-defined, the Bayesian analysis has no such trouble. Determining all its relations is reduced to well-defined computations in the probability space. When the system under investigation becomes very complicated, other approaches provide little guidance on how apparently conflicting evidence is to be combined. The fossil record is best explained by an old earth. The earth's cool temperature is best explained through Newton's law of cooling by a young earth. Using inference to the best explanation, to which do we infer? If they can pass the formidable hurdle of providing a well-defined probability space, Bayesians can answer the corresponding questions by mechanical computation. For all the information needed to trade off competing items of evidence lies within the conditional probabilities. Indeed, if any general question about belief or inductive support can be translated into a precise query in probability theory, it can be decided by a theorem that affirms or denies it.

With virtues as strong as these, it is all too appealing to hope that Bayesian analysis can be applied universally. When the inevitable problems arise, it is easy to dismiss them as the routine teething troubles of any infant who will outlive them and grow to boisterous maturity. Once that was a defensible attitude. As time passes and the problems remain unsolved, we can no longer afford to indulge the universal aspirations. If we are to understand what inductive inference is fundamentally, we need a different approach.

2. Objective and Subjective Bayesianism

My concern in this work are the objective relations of inductive support. Bayesians are also interested in these relations in so far as they expect them to be embraced by their analyses in one form or another. What complicates responding to Bayesian ambitions of universality is that Bayesianism is not a univocal doctrine. Rather it comes in many varieties.

A major division is between the objective Bayesians, such as Edwin Jaynes, and the subjective Bayesians, such as Bruno de Finetti. The objective Bayesians are distinguished by the claim that, in any epistemic situation, there is one correct probability distribution applicable. There is, in particular, one correct prior probability. In this regard, the project of objective Bayesians is closest to mine. I am comfortable regarding the objective Bayesians' conditional probability $P(H|E\&B)$ as an attempted expression of the objective strength of inductive support provided by evidence E in background B for hypothesis H. As I will remark briefly in the concluding section, the primary obstacle facing the objective Bayesians specifically is that rules needed to define this one correct prior are arbitrary.

Subjective Bayesians permit many probability distributions, constrained only by conformity with the axioms of the probability calculus. They characterize the freedom in our choice among the probability distributions as a free exercise of opinion. Thus, antecedent to the consideration of evidence, we are free to choose any prior probability distribution we like. Thus, at best, for a subjective Bayesian the conditional probability $P(H|E\&B)$ cannot simply express *the* strength of inductive support accrued to H since it has no unique value. Rather it is, at best, one of many possible mixes of evidential support and opinion. The hope is that eventually, in some longer term limit, the balance will move decisively toward objective support. There are also many attempts to derive measures from confirmatory support from the subjective probabilities.

My principal concern with subjective approach is that it demotes strengths of inductive support to a derived quantity. The primary quantity, the conditional probability, is a measure of belief or credence. Strengths of inductive support are to be recovered from them. That a notion of belief should be more primitive than a notion of inductive support has it the wrong way round. We wish to assess the strength of inductive support evolutionary theory derives from the fossil record or big bang cosmology derives from the cosmic background radiation. To make this assessment, we should not first have to determine our beliefs or credences. These strengths should not be dependent on our beliefs, else the objectivity of science is at risk. Worse, the project of assessing these strengths of inductive support from the beliefs has proven to be so troublesome that no univocal assessment is recoverable from the present literature in subjective Bayesianism. It was an approach that was risky to try and has not yet succeeded.

Since there is so much in common in the objective and subjective approaches to Bayesianism, it is impractical in what follows for me to keep them fully separated. There is often no need since argumentation concerning subjective probabilities can often be adapted to apply to objective probabilities; and conversely. Thus, in this chapter and the related chapters that follow, I will move freely between treating probabilities as objective relations of support and subjective credences. Please accept my apologies in advance for any confusions this may cause.

3. The Main Failing of Bayesianism

To repeat the disclaimer of the introduction, I have no quarrel with the application of Bayesian analysis in specific cases. There are many of successful and interesting cases that arise when the background facts provide the warrant needed. For example, a physical theory may supply the probabilities as the physical chances; and the particular conditions of the system may provide unambiguous prior probabilities.

My concern is the claim that Bayesian analysis is *universally* applicable to all systems, where it can supply the one, true logic of inductive inference. On the contrary, I will argue, there is a boundary beyond which Bayesian analysis fails. The existence of this limit is a corollary of the more general claim that there are no universal rules of inductive inference. For, without the boundary, Bayesian analysis would be providing a universal rule of inductive inference. The argument for this general claim was developed in Chapter 2. It is recapitulated here for the special case in which a mathematical calculus is used in one manner or another as a logic of inductive inference. The argument applies to all calculi used this way, not just to the probability calculus.

The key premise in the argument is that a calculus of inductive inference must place limitations on what can happen that is more restrictive than logical consistency.[126] Otherwise the calculus is not part of a system that implements inductive inferences. If the limitation goes beyond logical necessity, then it is by definition a contingent restriction, that is, one that may without logical contradiction be true or false. It follows immediately that there will be some conceivable domains that conform with the restriction and some that do not. Whether the calculus in question can be applied in some domain as a logic of inductive inference will depend on whether the requisite facts obtain. When they do, these facts warrant the use of this calculus in this domain.

This consideration applies directly to objective Bayesian approaches, for according to them, the probability calculus is the "logic of science." Or so proclaims Jaynes (2003) in the title of his treatise, *Probablity Theory: The Logic of Science.* The consideration also applies to

---

[126] That it must do this might not be immediately clear. The standard manipulations within the probability calculus, for example, are all deductive. Take ten independent tosses of a fair coin. If the probability of heads is 1/2 in each toss, then the probability that at least one heads appears is $1 - 1/2^{10} = 0.999$. So far all the reasoning has been deductive. The inductive component only enters when we employ an interpretive rule that tells us that outcomes of near unit probability are to be expected. Without some sort of interpretive rule like this, the probability is simply a mathematical quantity with no import for real things in the world.

subjective Bayesian approaches. While a probability measure for them mixes opinion and evidential warrant, the probability calculus does constrain someone to conform their beliefs with the evidence. The expectation is that conditionalization on a sufficiently rich body of evidence for some theory will lead a subjective Bayesian to mass the probability almost entirely on the true theory. Since such a body of evidence is finite but the theory's scope is infinite, this is a form of inductive inference. In this sense, a subjective Bayesian is implementing a scheme of inductive inference, although not as directly as an objective Bayesian.

We have now concluded that contingent facts warrant the applicability of some particular calculus of inductive inference in some domain. We might still hope that a single calculus is warranted universally. That would happen if it turns out that there is a single, contingent warranting fact that obtains in all domains in which we might conceivably practice inductive inference. Such a fact was pursued, in effect, in the nineteenth century under the guise of a search for a principle of the uniformity of nature. The search for such a universally applicable principle failed. As described in greater detail in Chapter 2, all candidates either proved empirically false or so hedged as to be vacuous.

The facts prevailing in some domain warrant the inductive logic applicable there. There is no single warranting fact common to them all. It follows that different inductive logics are warranted in different domains. This locality does not preclude one domain being very large. The success of probabilistic methods suggests that the domain or domains in which they are warranted as a logic of inductive inference are large. However every such domain is bounded and there are others beyond it in which a different logic is warranted. The logic warranted in these other domains may be governed by a calculus. But there may well be domains so irregular in their facts that no well-developed calculus can systematize whichever inductive inferences are warranted in them.

4. Probabilities without Warrants

The need for some sort of warrant for the use of probabilities becomes quite apparent if we consider cases in which there is no warrant. What results are striking inductive fallacies.

A simple example is provided by Van Inwagen's (1996, p. 95) question "Why is there anything at all?" He notes that there is only one possible world with nothing at all. There are infinitely many possible worlds with something, however, each differing in the configuration of something. Since we are assuming antecedently to have no basis for knowing what there is, if anything at all, we distribute our probabilities roughly uniformly over all the possible worlds. The result is that all the probability mass is attracted to the set of worlds in which there is

something. It follows that the probability of there being nothing is zero. It is, Van Inwagen concludes (p. 99), "as improbable as anything can be."

This conclusion is derived fallaciously. Not even the prodigious powers of the probability calculus can legitimately extract such a strong conclusion from premises so bereft of content. The fallacy derives directly from employing a probabilistic analysis in a context in which no background facts warrant the probabilities. This particular fallacy is unfortunately widespread. It is a version of what I have called elsewhere (Norton, 2010, §4) the "inductive disjunctive fallacy." More examples are described in (Norton, 2010, §4).

The "doomsday argument"[127] provides another illustration of a related fallacy. It uses only the evidence that our world has survived for t years. It asks after the probability that our world will end in $T > t$ years—"doom." Since our t can equally be any of the total T years of our world, the probability that our present world has survived for t years is $P(t|T) = 1/T$. The quantity we seek is the posterior probability $P(T|t)$, the probability that the world meets its doom after T years, given that it has survived for t years. An application of the ratio form of Bayes' theorem tells us that[128]

$$P(T_1 \mid t) / P(T_2 \mid t) = P(t \mid T_1) / P(t \mid T_2) = T_2 / T_1$$

Substitute $T_1 = t$ and $T_2 = 10t$, and we recover $P(T_1|t) / P(T_2|t) = 10$. This extraordinary conclusion tells us that doom is ten times more likely right now at t than is survival for another ten ages to $10t$.

Once again, the analysis delivers too much. The evidence is just that our world has survived t years. That is too thin an evidential basis for the strong conclusions drawn. They are not a reflection of what the evidence authorizes. They are merely artifacts of the use of an unwarranted inductive logic.[129] See Norton (2010, §6) for further analysis and for a proposal for a reduced inductive logic more appropriate to the problem.

---

[127] See Bostrom (2002a, chaps. 6–7) for an entry into the earlier literature on this argument.

[128] Assume equal prior probabilities $P(T_1) = P(T_2)$.

[129] Bostrom (2002, p. 57) seeks to warrant probabilistic analysis with his "self-sampling assumption": "One should reason as if one were a random sample from the set of all observers in one's reference class." Here "random sampling" implies equal probability of each sample drawn. Since it is an assumption without factual basis, it provides no warrant. Rather it enables us to identify and name the arbitrary posit that is the origin of the inductive fallacy.

5. Mapping the Boundaries: the Fate of Imprecise Probability

It is an interesting exercise to map out the boundaries for a probabilistic inductive logic. A simple axiom system, such as Kolmogorov's (1950) celebrated system, guides us to the boundaries.[130] The axioms have us posit a set of propositions[131] in which:

- non-negativity: we assign a non-negative, real valued probability $P(A)$ to proposition A;
- normalization: we assign a probability of unity $P(\Omega) = 1$ to the universal proposition (tautology) Ω; and
- additivity: $P(A \vee B) = P(A) + P(B)$ when proposition A and B contradict.

Inductive problems in which each of these fail are well-known. This chapter and chapters that follow will recount some of them. The "completely neutral support" relation below violates additivity, as do the indeterministic systems of a later chapter. Normalization and additivity are violated by the infinite lottery. That the structure is a real valued function is violated by the quantum inductive logic of a later chapter.

That some boundary has been reached is not controversial. What is controversial, at least in my mind, is how we should respond to it. I believe the correct response is to recognize that we have found the boundaries of probabilistic logic; that we should recognize that different logics prevail in the domains beyond it; and that we should begin the task of identifying them.

The standard response in the literature is different. It is to weaken the probability calculus until the generalized calculus has been weakened enough to encompass whatever troublesome counterexample has arisen. For example, as we shall see shortly, the field of imprecise probability encompasses systems that violate additivity by replacing a single probability measure with a set or them; or it may employ superadditive measures. This is the project of variety of approaches grouped under the heading of "imprecise probability."[132]

This approach is, in the end, an ill-fated attempt to preserve the core idea that "It's all probabilities." The domains covered are ones in which an inherently non-probabilistic inductive

---

[130] Kolmogorov's axioms are simpler since they specify an unconditional probability $P(A)$. Conditional probabilities are introduced through the definition $P(A|B) = P(A\&B)/P(A)$. This approach is more restrictive than providing more complicated axioms directly for conditional probabilities. But the simpler approach suffices for the present analysis since the more complicated approaches only move the boundaries slightly.

[131] More precisely, we posit a Boolean algebra of propositions, which is a set of propositions closed under finite or countable disjunction v, conjunction & and negation -.

[132] For an introduction to this literature, see Bradley (2016) and the resources on the website of the *Society for Imprecise Probability*, http://www.sipta.org.

logic is warranted. What imprecise probability does is to employ an additive calculus to simulate a non-additive logic. It thus preserves the illusion that an additive measure is somehow still the core of the logic. The better approach would simply have been to recognize that a qualitatively distinct logic is required and to map out its behavior as a distinct logic.

In any case, this general stratagem of extending the calculus offers only a temporary respite. For as long as the generalized calculus supports inductive inference, it must place restrictions on the system that go beyond mere logical consistency. These are contingent constraints that, recalling the argument of Section 3 above, will not obtain in all domains. Further investigation will reveal new boundaries and new domains beyond them, as we shall see shortly in Section 9 below.

This standard response is driving towards an unhappy ending. Each time a calculus is generalized to embrace new examples, it is weakened in the sense that it becomes less restrictive. As long as the generalized logic places some restrictions on systems beyond logical consistency, the domains in which it applies are limited. The need to generalize to embrace unanticipated counterexamples will continue. The process of generalization can only assuredly terminate when the logic places no restrictions beyond logical consistency on its domain. But then it has ceased to be an inductive logic.

6. The Principle of Indifference

To make these concerns more concrete, it will be helpful to develop the simplest case in which the boundaries of probabilistic analysis are breached. It is "completely neutral support" and arises when we have inductive support that is, in objective terms, maximally uninformative. In subjective terms, it corresponds to the case of complete ignorance. This case has been explored extensively in Norton (2008, 2010) as part of an investigation of the import of the principle of indifference. We will first recall the principle and its application; and then develop the notion of completely neutral support in the next section.

The form of the principle that I prefer is:

> *Principle of Indifference.* If one has no grounds for distinguishing several outcomes, then we should assign equal inductive support to them.

The principle in this form is a truism of evidence. It just reflects the requirement that discriminations in inductive support cannot be made arbitrarily.

The principle applies when we have indistinguishable outcomes. They are realized most securely through invariances, which are transformations that leave the relations of inductive support unchanged. Their use is familiar and unproblematic, initially. When we have a fair coin toss, our formal analysis is unchanged if we switch the labels on the sides of the coin. Whatever

ground we have for favoring heads would remain unchanged if reassign the label "heads" to the other side of the coin; and reassigned the label "tails" similarly. That is we have no grounds for distinguishing the outcome of either side. The principle of indifference requires us to assign equal support to each. If the relations of support are probabilistic, then each side is assigned equal probability.

In case of a coin toss, the invariance under this permutation of the labels is derived from background physical facts: the mechanical conditions of tossing are such that they favor both sides equally. One can also have a more epistemic version. The coin need not even be tossed. We might just imagine it as sitting somewhere, untouched, in a drawer. But since our information about the coin is so limited, we have no grounds for distinguishing whether it is heads up or tails up.

Invariances like these seem benign until we start to combine them. Then they yield the well-known paradoxes of indifference. An early and familiar example is due to Keynes (1921, Ch.4), who also named the principle of indifference. We ask after the country of a man who may be from

France, Ireland, Great Britain (1)

Since we have no grounds for discriminating among them, so we assign equal probability of 1/3 to each. However the disjunction of two outcomes (Ireland or Great Britain) is equivalent to just British Isles at the time Keynes first wrote. So we might equally ask if the man is from

France, British Isles (2)

Since again we have no grounds for discriminating, we assign equal probability of 1/2 to each. We have now arrived at contradictory assignments. For we have assigned both probability 1/3 and probability 1/2 to France as the man's country.

Examples like these are usually used to impugn the principle of indifference. That is a misdiagnosis. The principle is a truism of evidence and not readily discarded. It just says that the support for outcomes should not differ without a reason. What is overlooked in these efforts to impugn the principle is that the real cause of the trouble lies elsewhere. It is the presumption that relations of inductive support must always be probabilistic. These paradoxes of indifference are an early indication that they need not always be so.

Recently there have been several alterative interpretations of the import of the principle of indifference on representing the neutrality of support. Benétreau-Dupin (2015) draws on the existing ideas in imprecise probability and explores representing completely neutral support through sets of probability measures. Eva (manuscript) proposes a novel accommodation in which not all degrees of support are comparable.

7. Completely Neutral Support

7.1 Invariance under Redescription

The transformation from (1) to (2) is a "disjunctive coarsening" of the outcome space. Two outcomes are replaced by a single outcome, their disjunction. The reverse transformation is a "disjunctive refinement." If we conceive these operations as redescriptions of the outcomes, Keynes' example depends on a particular invariance:

> *Invariance under redescription.* In cases of completely neutral support, equality of inductive support over outcomes remains under disjunctive coarsening and refinement.

We arrive at the formal representation of completely neutral support by applying this invariance to an outcome space that has finitely many, mutually exclusive atoms $A_1, A_2, A_3, \ldots, A_n$, where an atom is the logically strongest proposition in the outcome space.[133] If our circumstance is maximally uninformative concerning these outcomes, then the principle of indifference enjoins us to assign equal support to each. Writing:

$[A|B]$ = inductive support accrued to proposition A from B.

we then infer:

$$[A_1|\Omega] = [A_2|\Omega] = [A_3|\Omega] = \ldots = [A_n|\Omega] = I \qquad (3)$$

where I represents the common inductive strength of support.

We can disjunctively coarsen the outcome space by replacing the first two propositions A_1 and A_2 by their disjunction, which we will write as $A_{1\text{v}2} = A_1 \text{ v } A_2$. The new outcome space is $A_{1\text{v}2}, A_3, \ldots, A_n$ and has only n-1 propositions. Proceeding as did Keynes, we remain maximally uniformed about these propositions, so we must assign equal support to each. That is,

$$[A_{1\text{v}2}|\Omega] = [A_3|\Omega] = \ldots = [A_n|\Omega] = I \qquad (4)$$

The same strength of support I must be used in both (3) and (4) since they have many common terms. For example, both include $[A_3|\Omega]$, so we can infer

$$[A_1|\Omega] = [A_2|\Omega] = [A_3|\Omega] = [A_{1\text{v}2}|\Omega] = [A_1 \text{ v } A_2|\Omega] = I$$

---

[133] As before, the outcomes space is a Boolean algebra of propositions whose universal proposition or tautology Ω is the disjunction of all the atoms $\Omega = A_1 \text{ v } A_2 \text{ v } \ldots \text{ v } A_n$. Proposition A_i is an atom just if any proposition A that entails it is either A_i itself or the contradiction.

Continuing by forming more disjunctive coarsenings of the original outcome space, it is easy to see that the support offered to any contingent disjunctions of atoms[134] is the same strength of support I:

$$[\text{any contingent disjunction of atoms} \mid \Omega] = I$$

However every contingent proposition in outcome space is equivalent to some disjunction of atoms. Thus we arrive at[135]

Completely neutral support (5)

$[\text{tautology } \Omega \mid \Omega] = 1$

$[\text{any contingent proposition} \mid \Omega] = I$

$[\text{contradiction} \mid \Omega] = 0$

The strengths "1" and "0" have been chosen for continuity with the familiar probabilistic case. Their arithmetic properties are not invoked.

This subsection argues for a unique characterization (5) of completely neutral support. It is tempting to block the argument at equation (4) by asserting that the principle of indifference should only be applied to the most refined outcome space, which is (3) in this case. I set aside the question of whether this knowledge is adequate to block the argument. For whether it is or not, the difficulty is easily escaped. It is presumed that we know that (3) is the case of maximum refinement. So let us consider the case in which we do not know that (3) is the maximum refinement; or that we know positively that there is no maximum refinement. Then the argument for completely neutral support goes through.

How might there be no maximum refinement? Such a case arises if the propositions represent ranges of some real-valued parameter x. E.g. A_m might correspond to $m \leq x < m+1$. Then we can disjunctively refine by replacing A_m by a disjunction $B \vee C$, where B corresponds to $m \leq x < m+1/3$ and C corresponds to $m+1/3 \leq x < m+1$. Since these intervals can be divided indefinitely, there is no most refined outcome space.[136]

---

[134] This requirement of contingency excludes the tautology Ω.

[135] Keeping distinct values for the tautology and contradiction presumes sufficient logical knowledge that we can discern them from the contingent propositions. One could also define a still more extreme case in which we are maximally ignorant of deductive relations amont the propositions, so that all the strengths are I.

[136] It is tempting to argue that the refinement must divide parameter values into uniform intervals. That requirement fails since what may be a uniform division for one scaling of the parameter will not be so for another. The wine and water example below illustrates the problem.

7.2 Invariance under Negation

The invariance under redescription of the last section is already well represented in the literature. There is a second, less familiar invariance that leads to the same result. To see it consider some proposition A about which you know nothing at all. How well supported is it? Now consider its negation, not-A. Is the negation any more or any less well supported? If there is any doubt about how to answer, imagine that the first question asked concerned the proposition not-A, which we initially labeled B. Is not-B any more or less supported than B?

Invariance under negation asserts that the two strengths of support are the same, simply because, by supposition, we have no basis for discriminating between them. Switching their labels makes no difference to the strengths of support.

> *Invariance under negation.* In cases of completely neutral support, the inductive support for a contingent[137] proposition and its negation are the same.

Let us implement this invariance in the outcome space with atoms $A_1, A_2, A_3, \ldots, A_n$. Any contingent proposition consists of a disjunction of some number of atoms from one to n-1. For example, the negation of A_1 is $A_2 \vee A_3 \vee \ldots \vee A_n$; and the negation of $A_2 \vee A_3 \vee \ldots \vee A_n$ is A_1; and so on for all other possible combinations. If we have a case of completely neutral support, we infer from invariance under negation that

$$[A_1 \,|\Omega] = [A_2 \vee A_3 \vee \ldots \vee A_n \,|\Omega] = I_1$$

$$[A_1 \vee A_2 \,|\Omega] = [A_3 \vee A_4 \vee \ldots \vee A_n \,|\Omega] = I_{1,2}$$

etc.

The strengths of support must be distinguished as $I_1, I_{1,2}, \ldots$ at this stage, since negation invariance, by itself, is not strong enough to force all the strengths to the same value. Their equality, however, can be recovered if we add the condition:

> *Monotonicity*. The strength of support of a proposition is no greater than[138] that of its consequences. If A entails B, then $[A|\Omega] \leq [B|\Omega]$.

Since A_1 entails $A_1 \vee A_2$ and $A_2 \vee A_3 \vee \ldots \vee A_n$ is entailed by $A_3 \vee A_4 \vee \ldots \vee A_n$, we have

$$I_1 = [A_1 \,|\Omega] \leq [A_1 \vee A_2 \,|\Omega] = I_{1,2}$$

---

[137] A stronger version arises if we cannot discern the tautology and contradiction from the contingent propositions. Then the restriction to contingent propositions can be dropped.

[138] That is, we assume that there is a partial order "<" defined over the strengths. It is transitive and antisymmetric. Monotonicity is widely assumed but is not necessary and one can conceive logics in which it fails. An example is the specific conditioning logic of Norton (2010a, §11.2).

$$I_1 = [A_2 \text{ v } A_3 \text{ v } \ldots \text{ v } A_n \,|\Omega] \geq [A_3 \text{ v } A_4 \text{ v } \ldots \text{ v } A_n \,|\Omega] \; = I_{1,2}$$

These two inequalities can only obtain if the two strengths are equal:

$$I_1 = \; I_{1,2} = I$$

Proceeding analogously for all the other cases, we recover the equality of the strengths of support with the single value I for all contingent propositions. The details of the recovery are straightforward but a little tedious and are given in Norton (2008, §6.3).

7.3 Invariance from Ignorance or Positive Warrant

The representation of completely neutral support has been developed assuming that the invariances prescribed can come about in some circumstance. It is tempting to invert the argumentation. Since the totality of these invariances is incompatible with a probabilistic treatment of inductive support, might we then infer that it is impossible for us ever to be in a position in which these invariances are realized? I have argued that sufficient ignorance will realize these invariances. However my real concern in this work is not ignorance but strengths of support warranted by background facts. We might well wonder what sort of facts could realize these invariances. What sort of coin tosses or die throws or other similar machines could yield probabilities such that $P(A_1) = P(A_2) = \ldots = P(A_n) = P(A_1 \text{ v } A_2) = P(A_1 \text{ v } A_2 \text{ v } A_3) = \ldots$ as required by (5)? Since these probability assignments violate the additivity axiom of the probability calculus, no probabilistic randomizer can realize them, precisely because its mechanism is probabilistic.

This inversion of the arguments succeeds only in so far as we are restricted to warrants for support arising from probabilistic randomizers. The chapter on "Indeterministic Physical Systems" describes many physical systems whose indeterminism is not probabilistic and which realizes the two invariances employed above. The inductive logic warranted for these systems conforms with completely neutral support (5). However, where here I have used the three values 0, I and 1, in the later chapter I relabel these values as *imp* ("impossible"), *poss* ("possible") and *nec* ("necessary") to reflect better they physical underpinnings of these new cases.

8. Von Mises' Wine and Water.

The argument for completely neutral support derives from either of the two invariances just stated. It is tempting to try to defeat them by calling upon asymmetries in propositions that are not respected by the invariances. For example, while there may be no finest disjunctive refinement, there are some that are more refined and some that are less so. All negations are not the same in their atom counts. The negation of a single atom proposition A_1 has a different atom

count from the negation of a disjunction of n-1 atoms, A_2v…v A_n. These are all asymmetries among the cases that are not reflected in the invariances. We may well ask how the invariances can be maintained with these asymmetries.

The uniform response to all concerns of this type is merely to reduce our knowledge still further, until the symmetries are restored. Then the invariances apply and completely neutral support is recoverable. The asymmetries depend on choosing a particular outcomes space to describe some system's behavior. A proposition may consist of one atom in one outcome space, but a disjunction of n-1 atoms in another, where both spaces describe the same system. These differences of atom counts are then immaterial to the invariance if we have no way to discern which of the outcome spaces is the "right one."

A version of von Mises' wine and water example illustrates this effect. A goblet contains a mixture of wine and water. All we know is that the ratio x of wine to water lies in the interval $0.5<x<2$. It follows that the ratio $y = 1/x$ of water to wine lies in the interval $2>y>0.5$. The variables x and y each allow the definition of outcome spaces that describe the same physical goblet. The first has atoms:[139]

$$A_1: 0.5<x<1 \quad A_2: 1<x<1.5 \quad A_3: 1.5<x<2$$

and the second has atoms

$$B_1: 0.5<y<1 \quad B_2: 1<y<1.5 \quad B_3: 1.5<y<2$$

The principle of indifference requires us to assign uniform support across the atoms:

$$[A_1|\Omega] = [A_2|\Omega] = [A_3|\Omega]$$

$$[B_1|\Omega] = [B_2|\Omega] = [B_3|\Omega]$$

One might try to adapt probabilities to these equalities, by setting

$$P(A_1|\Omega) = P(A_2|\Omega) = P(A_3|\Omega) = 1/3$$

$$P(B_1|\Omega) = P(B_2|\Omega) = P(B_3|\Omega) = 1/3$$

The adaptation fails since $A_1 = B_2$ v B_3 so that we end up with a contradiction:

$$1/3 = P(A_1|\Omega) = P(B_2 \text{v } B_3|\Omega) = P(B_2|\Omega) + P(B_3|\Omega) = 2/3$$

A similar contradiction follows from $B_1 = A_2$v A_3.

Figure 1 illustrates how the atoms in the two spaces are related:

[139] To avoid the need to juggle too many "<" and "≤," I employ the expedient assumption that x and y can never adopt exactly the values 0.5, 2/3, 1, 1.5 and 2.

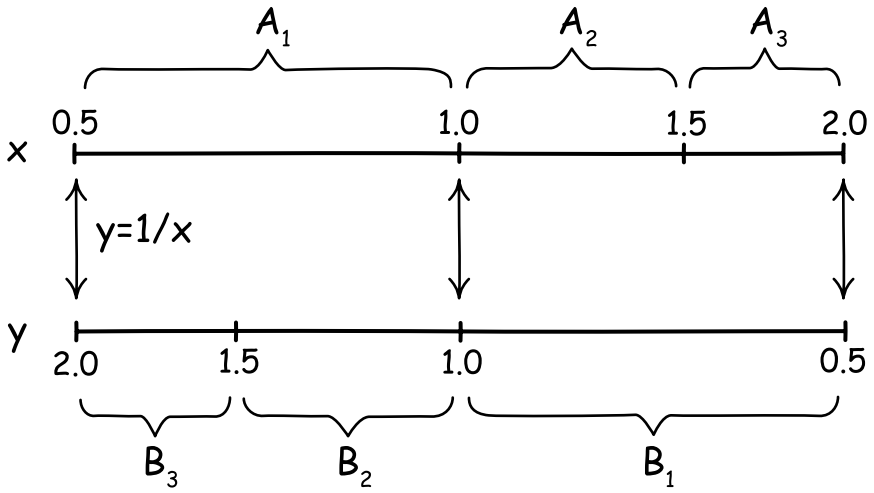


Figure 1. Relations between two outcome spaces for the wine and water example

Instead of trying to impose probabilities where they do not belong, we can apply the invariances to arrive at the completely neutral support (5). We start with a coarsened outcome space with two outcomes: the ratio of wine to water is either greater or less than 1. This space can be represented in two equivalent ways:

$$A_1, A_{2\text{v}3} = A_2 \text{ v } A_3 = B_1 \quad \text{or} \quad B_1, B_{2\text{v}3} = B_2 \text{ v } B_3 = A_1$$

Following the principle of indifference, we assign equal support to each outcome:

$$[A_1|\Omega] = [B_1|\Omega] = I$$

Implementing invariance under redescription, we disjunctively refine the outcome space and expect the equalities to be preserved. There are two ways to implement the disjunctive refinement. We can refine $B_1 = A_2 \text{ v } A_3$ and end up with

$$[A_1|\Omega] = [A_2|\Omega] = [A_3|\Omega] = I$$

Since $A_2 \text{ v } A_3 = B_1$ we have

$$[A_2|\Omega] = [A_3|\Omega] = [A_2 \text{v} A_3|\Omega] = I$$

Or we can refine $A_1 = B_2 \text{ v } B_3$ and end up with

$$[B_1|\Omega] = [B_2|\Omega] = [B_3|\Omega] = I$$

Since $B_2 \text{ v } B_3 = A_1$ we have

$$[B_2|\Omega] = [B_3|\Omega] = [B_2 \text{v} B_3|\Omega] = I$$

In these relations, we have now recovered much of completely neutral support (5) for the two outcome spaces. Continued application of the invariances recovers the remaining parts. For

example, applying either invariance to the disjunctive coarsening A_2, $A_{1v3} = A_1$ v A_3 returns $[A_2|\Omega] = [A_1 \text{ v } A_3|\Omega] = I$.

Returning to the concerns expressed at the start of this section, there is no sense in which one of the outcome spaces A_1, A_2, A_3 or B_1, B_2, B_3 is more refined than the other. The first represents a refinement of B_1 but not A_1; and the second a refinement of A_1 but not B_1. The perfect symmetry in all the formulae gives us no basis for preferring one over the other.

Negation invariance is also implemented automatically. A_1 is the negation of B_1 and conversely. They are both assigned the same support I. The earlier concern that negation invariance might be troubled by an asymmetry in atom counts is also not realized. In the A-outcome space, A_1 is comprised of one atom and B_1 is a disjunction of two atoms. In the B-outcome space, this is reversed: A_1 is a disjunction of two atoms and B_1 is comprised of one atom. Once again, the perfect symmetry in all the formulae gives us no basis for preferring one over the other.

9. Imprecise Probabilities Again

The literature in imprecise probability treats this case of completely neutral support or, as they characterize it often in subjective terms, of complete ignorance. In so far as these treatments are seeking to replicate formally the behavior of completely neutral support (5), they are doing the right thing. The invariances show that (5) is the correct representation for this case. There will be room to quibble about the treatments given in imprecise probability, as I will below, but these quibbles are minor in comparison to the major concern. It is that they are not doing the right thing, if this representation of complete ignorance is intended as part of the case for the universality of the particular scheme employed, now conceived as some kind of a generalized probability theory. For I have already described above in Section 4 how all such efforts are necessarily ill-fated. In so far as the generalized probability theory places restrictions that go beyond logical consistency, these restrictions are contingent and it is inevitable that there are systems that contradict these factual restrictions. The generalized probability theory must fail to apply to these systems, so that its aspirations for universality must fail.

Consider a popular approach to imprecise probability that employs sets of probability measures to represent credal states. Benétreau-Dupin (2015, §3) has given a careful accounting of them and their prospects in regard to complete neutrality of support. This approach depends on the assumption that probability measures can be defined for the system in question. Otherwise sets of the measures cannot be formed. Thus the approach fails when applied to systems with nonmeasurable outcomes, such as are investigated in the chapter "Uncountable Problems," for

these outcomes admit no probability measures. It also fails when applied to infinite dimensional outcome spaces, for they admit no non-trivial additive measure, even if the requirement of normalization to unity is dropped. We shall see such an example arising naturally in the chapter "Indeterministic Systems" in conjunction with an indeterminism in Newtonian cosmology.

There are lesser technical issues as well. As Benétreau-Dupin (2015, §3) recounts, it is unclear just which set of probability measures should represent completely neutral support. The natural choice is the set of all probability measures in the outcome space. However that set has the unappealing property for Bayesians that it is preserved under conditionalization so that inductive learning is precluded.

In my view, this representation is needlessly complicated since it assigns no definite probability value to a given outcome. Rather it assigns all values to each outcome and not just as an interval of values, but with each value part of one of infinitely many probability measures. There is no analog of the simple and adequate ignorance strength *I*. Worse, a set of probability measures violates invariance by negation. This arises because every probability measure is non-decreasing as we pass through chains of deductive consequences, such as

(Contradiction)
entails (A_1)
entails (A_1 v A_2)
entails (A_1 v A_2 v A_3)
entails …
entails (tautology)

Since the contradiction is assigned zero probability and the tautology unit probability, the probabilities of these outcomes must, at some point, be strictly increasing. That is, this strict increase endows the probability measures with a directedness from fewer to greater atom propositions. The operation of negation maps the strengths assigned to propositions with fewer atoms to those with more atoms; and conversely. That is, it flips the assignments of inductive strengths with respect to this direction. It follows that any assignments of strengths that is directed cannot be preserved by negation. Since all measures have such directedness, a set of measures cannot be preserved under negation. The invariance is violated. Norton (2007, §6) explores this failure as a failure of a duality required by the representation of completely neutral support. Benétreau-Dupin (2015, §3) has sketched a dissenting view.

Other popular approaches employ some form of superadditivity of a measure.[140] An early version arises in the theory of Shafer-Dempster belief functions. The "vacuous belief

[140] A measure is superadditive if the value assigned to a disjunction of mutually incompatible outcomes A v B is greater than the sum of the values assigned to A and B individually.

function" (Shafer, 1976, p.22) assigns unity to the tautology and zero to every other proposition, including the contradiction. This vacuous belief function does respect both invariances of Section 7 above. However it has the awkward feature of assigning the same value of zero also to the contradiction. That means that its individual values do not distinguish complete disbelief, which we must have in the contradiction, from complete ignorance, which is presumed for all the contingent propositions.

Walley's (1991) related approach represents a credence in each outcome by two numbers, a lower and an upper probability. So-called "vacuous upper and lower probabilities" (p. 92) assign a zero lower probability and a unit upper probability to all contingent propositions. "They seem to be," Walley writes, "the only reasonable models for 'complete ignorance'…" He continues to note that this representation accords with appropriate invariance properties: it is invariant under refinements and coarsenings of the outcome space. Walley's representation, considered in isolation from the rest of his system, is unobjectionable. It is equivalent to the representation of completely neutral support (5). Wherever the strength I appears in (5), for example, Walley has the functionally equivalent pair of upper and lower probabilities, 0,1. Unlike the Shafer-Dempster vacuous belief function, contingent propositions are distinguished from the contradiction, in that both lower and upper probabilities of zero are assigned to the contradiction.

While this particular representation of complete ignorance is successful, there are other problems with Walley's system, however. Notably, he derives his quantities in the de Finetti tradition as previsions associated with betting scenarios through which some sort of universality of applicability is suggested. My concerns over this betting approach will be addressed below.

More generally, more exotic problems in inductive inference will present continuing challenges to aspirations of universality for all systems of imprecise probability. We will see some in coming chapters. It is not clear now how these systems can accommodate the different sectors of the logic native to the infinite lottery. As detailed in a chapter to follow, the sectors are divided into finite sets of outcomes, infinite-co-infinite sets of outcomes and infinite-co-finite sets of outcomes, each with their own distinctive structures. If some form of probabilistic account is to be preserved, the three sectors would appear to need both infinitely small and infinitely large probabilities. Still more serious is the challenge of recovering the logic native to quantum systems, as sketched in the chapter "Quantum Inductive Inference," for the basic structure of that logic is not a real valued function but an operator in a Hilbert space.

As long as theories of imprecise probability implement an inductive logic, they will place contingent constraints on the domains to which they can apply. That means that we should always expect new systems to arise to which their inductive logic does not apply. The cycle of extension and counterexample can continue without end, unless the theory of imprecise logic is

so weakened by generalization that it places no factual restriction on the domains to which it applies. However then the theory has ceased to implement an inductive logic. It will implement only the requirement of logical consistency.

10. All Proofs of the Necessity of Probabilities are Circular

One of the more appealing aspects of the Bayesian approach is that its proponents have systematically taken on the burden of demonstrating that their approach is the uniquely correct one. The efforts at proof go back at least as far as the "Dutch book" arguments of Ramsey (1926) and de Finetti (1937) and their expansion by Savage (1954). Other approaches include the identification of necessary conditions and their consequences through representation theorems, such as developed by Cox (1961) and Jaynes (2003). More recent approaches, such as recounted in Pettigrew (2016), focus directly on the notion of accuracy as measured by scoring rules.

The literature is energetic. Existing approaches are subject to continuing amendment and expansion; and new approaches are offered. Optimists will see this as a proper and ever-improving response to a worthy problem of the first order. My reaction is more pessimistic. The ferment is the inevitable outcome when a literature sets itself an unattainable task. No proposal proves sustainable, but there is always the hope that a new approach might escape the problems that beset the last one.

That the goal is unattainable follows from the material approach to inductive inference. The proofs seek to establish the necessity of probabilities, that is, that objective degrees of inductive support or subjective degrees of belief must be probabilities. It has already been argued in Section 3 above that this is a contingent proposition. It may be true or false. It is not a necessary truth, demonstrable by pure logic alone.

It then follows directly that all proofs of the universal necessity of a probabilistic inductive logic, objective or subjective, must be circular. For all such proofs are logical deductions. They start with premises and from them deduce the conclusion sought. It is a basic fact of deductive logic that these premises must be at least as strong logically as the conclusion sought. Since the necessity of probabilities is not an a priori truth of logic, it follows that the premises of any demonstration of the necessity of probabilities must already contain exactly that necessity as contingent propositions, in one form or another.[141]

---

[141] While I know of no such efforts, one might seek to show the universal necessity of a probabilistic inductive logic through a demonstration that itself employs inductive inferences. These efforts would face a dilemma. If the inductive inferences used are not probabilistic, it is conceded at the outset that some inductive inferences are not probabilistic. If the inductive

Seen in this perspective, there is a simple procedure for undoing all purported proofs of the necessity of probabilities: one merely needs to explore the premises of the proof and uncover the disguised presumption of probabilities. No matter how natural and comfortable the proof's starting points, no matter how congenial and convincing they may appear initially, the proof will depend on contingent premises that presuppose precisely what is to be proved. One then sees that the proofs are, in the best case, no better than merely positing probabilities in the first place. In a worse case, the premises are logically stronger, so one must assume more than the necessity of probabilities in order to derive the necessity of probabilities. In that case one is better off positing probabilities directly in the first place.

There is a dominance argument implicit in these last observations. If our interest is to minimize risk of error in an attempt to vindicate probabilities, we are never better off using one of these proofs. That is, merely directly positing probabilities weakly dominates, in a game theoretic sense, any justification by a proof. The only possible gain from the proof is a psychological one: one might find the premises posited by the proof more intuitively congenial, even if they risk being logically stronger than the goal.

This way of approaching the proofs casts a different light on the activity of the vindicators of probabilities. For nearly a century, their efforts have produced a flourishing literature that never quite produces the final, definitive demonstration. Rather the community of vindicators finds itself forever dissatisfied with the latest vindication. Sometimes new avenues are explored. Sometimes, the dissatisfaction results in a regress: a quest for further demonstrations that would establish the premises of the most recent demonstration. This regress cannot end well, for each further demonstration merely faces the same challenge anew: it must find new, contingent premises from which to derive the old ones; and then it will have to justify these new premises. If the dissatisfaction is deep enough, it will seek another approach. This is a regress of reasons that cannot end in the proof sought.

We can now see that the continuing pain is not the result of some maddening inability of the vindicators to find just the right premises for their demonstrations. Rather it is the inevitable result of the awkward fact that there are no premises truly adequate to the task. The best a vindicator can do is to proceed from an assumption that probabilists will find intuitively appealing, since the assumption is equivalent to or logically stronger than the presumption of probabilities. That same assumption will appear arbitrary and even uncongenial to someone who is antecedently unconvinced of the necessity of probabilities.

---

inferences are probabilistic, then it must be shown that this particular probabilistic demonstration of the necessity of probabilities is not viciously circular.

The future of the vindication project is easy to predict. Like the circle squarers and angle trisectors of old, the vindicators will be trapped perpetually in the frustrating cycle of promising avenues; proofs that finally seem to succeed; and then the unhappy recognition that the latest proof falls just a little a short. If they persist, it must be so. The escape from the trap lies in the recognition that there is no necessity to probabilities.

Might one worry that this mode of objection is made too easily? If it works, might it not be able refute any demonstration of any proposition in philosophy? This worry is easily set aside, for this mode of objection applies only when a deductive proof is offered for a contingent proposition. Then it is cogent and should be applied.

For example, consider a theist who offers a deductive demonstration of the necessity of God's existence, such as Anselm's ontological argument. What reply can be given by a skeptic who holds the assertion of God's existence to be a contingent proposition? The skeptic would proceed as I have with the contingency of probabilities. Assuming its steps are valid, the skeptic would look at the premises of the theist's demonstration and expect to find contingent premises that are logically at least as strong as the necessity of God's existence. The theist would be untroubled by the display of these contingent premises. They would merely be a reformulation, possibly in logically stronger form, of what the theist already believes. The skeptic however would object that, precisely because of this, the demonstration is no demonstration at all, but is assuming what is to be proved.

11. Illustrations of Circularity

The last section established as a generality that all vindications of probabilities will prove circular and it predicted a manifestation of the failure in a regress of reasons. The exercise now is to find these circularities in the standard vindications. The literature in vindications is so large that it is impractical to cover it in any details. Therefore I have chosen to examine a recent, presently popular vindication in greater detail. The scoring rule or accuracy based vindication is driven by a single, intuitively appealing idea. It suggests that there is a unique way to distribute our beliefs such that we cannot improve their accuracy, whichever circumstance may prove the true one. That distribution is probabilistic. Treating this case adequately proved to be a lengthy undertaking and has required a separate chapter, which follows this one. That chapter illustrates how the presumption of probability resides in the particular choice of the scoring rule used to measures accuracy. That choice must be carefully fine-tuned, else the approach fails to return probabilities. We shall then see how efforts to protect the fine-tuning from suggestions of circularity trigger just the doomed regress predicted above.

Here we shall take a briefer look at two other attempts to vindicate probabilities. We shall see that each presumes the very thing sought.

12. The Dutch Book Argument

12.1 The Betting Scenario

The Dutch book argument or arguments, if we separate out various forms of them, derive initially from Ramsey (1926) and de Finetti (1937). They have been a mainstay of the subjective Bayesian approach for decades.[142] The argument begins with the assertion that beliefs must be manifested operationally. The method chosen is to offer agents various bets and to determine their beliefs from which bets they accept and refuse. The argument then takes on a normative[143] burden: if—and only if—the beliefs manifested do not conform with the probability calculus, then it is possible, the argument goes, to offer the agent a combination of bets that results in a sure loss. That combination is the "Dutch book." Beliefs that allow this sure loss are disparaged as incoherent and reflect the supposed irrationality of non-probabilistic beliefs.

The central structure of the argument is the wager offered. The stake S is the sum of money or some other valuable of similar type associated with the bet. How it is distributed is decided by the presently unknown truth or falsity of some proposition A. In a bet "on" A, the agent pays a price qS for the possibility of gaining $S>0$ if A turns out to be true. Otherwise, if A is false, the agent simply loses the price. In sum:

Proposition A is true.	Agent gains $S - qS$
Proposition A is false	Agent gains $-qS$

Table 1. Payoffs of a bet "on" A ($S>0$) and "against" A ($S<0$)

This arrangement is reversed for a bet "against" A. It is most simply implemented by selecting a negative S and using the same payoffs as in Table 1. The full analysis requires an additional assumption to which we will return shortly:

> *Existence of a fair bet.* For any proposition A, for each agent, there is a "fair" betting quotient q such that the agent is willing to accept either side of the bet: "on" A or "against" A. This betting quotient measures the agent's strength of belief in A.

---

[142] For recent surveys of a very extensive literature, see Hájek (2009) and Vineberg (2016).

[143] The normative element is essential. The system so narrowly constrains an agent's possible responses that it is a dismal means of ascertaining beliefs non-coersively.

The main result is that failing to conform the betting quotients to the axioms of the probability calculus allows a Dutch book to be made against the agent (Dutch book theorem); and that conforming the betting quotients to the axioms makes it impossible for the Dutch book to be made (converse Dutch book theorem). A simple illustration does not even require the notion of a fair bet. Avoidance of sure loss immediately precludes $q>1$. For if $q>1$, a bet on any proposition A leads to a loss of $S(1-q)<0$ if A turns out to be true; and a loss of $-qS<0$ if A turns out to be false.

12.2 The Dubious Presumption

How does this construction presume probabilities? The principal, tendentious presumption is laid out in plain sight at the very start. In requiring agents *always* to express their beliefs in terms of monetary bets, accepted or refused, it forces agents to represent their beliefs on a single numerical scale. The betting quotient q has to be a real number, else the payoffs $S-qS$ and $-qS$ cannot be formed. Since there is so much more detail to come, it is easy to treat this presumption as an unimportant preliminary and to skip past it. That is mistaken if one wants to understand which are the strongest assumptions underlying the Dutch book argument. Once proponents of the argument can get us to accept that beliefs are measurable on a real number scale, most of their work is done.

The arguments for this presumption are weak in relation to the strength of what it asserts. De Finetti (1937, p. 139) pretends the assumption is innocuous. It is, he says:

> …the trivial and obvious idea that the degree of probability attributed by an individual to a given event is revealed by the conditions under which he would be disposed to bet on that event.

Ramsey (1927, p. 166) recognized that this idea is not nearly so innocent. "It is a common view," he conceded "that belief and other psychological variables are not measurable…" He recognized that something stronger is needed and asserted that, without some measurement protocol, meaninglessness threatens (p. 167):

> …degree of a belief is just like a time interval; it has no precise meaning unless we specify more exactly how it is to be measured.

Here Ramsey echoes operationist sentiments of growing popularity in the 1920s. Bridgman (1927) was then writing his manifesto of operationism. It used special relativity, including its treatment of time, as his motivating example (Ch.1). Perhaps Ramsey's remark on time alluded to this example. At the same time in psychology, behaviorists were urging the elimination of invisible thoughts and ideas in favor of observable behaviors.

Nearly a century later, operationism and behaviorism have long fallen from favor. They proved unable to deliver accounts that match the richness of complex physical theories and

mental content. The deepest problem with operationism was its core assertion. It is, in Bridgman's (1927, p.5, his emphasis) words:

> In general, we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations.*

It is a false assertion. Concepts are not synonymous with the operations that measure them. Time is not the ticking of a clock; or length the laying out of a ruler; or mass the extension of a spring in a weighing scale; or electric current the deflection of a needle in an ammeter. Correspondingly, belief is not the behavior of accepting or refusing bets. To reflect a widespread objection[144] to the Dutch book approach: beliefs have cognitive goals concerning learning the truth; betting behaviors have pragmatic goals in the maximizing of one's fortune. Supposing otherwise risks oversimplifications comparable to those that doomed operationist analyses elsewhere. Here we might imagine the (possibly fictional) enlightened Buddhist who has no material desires. Such a figure would, under these operationist strictures, be incapable of holding beliefs.

While concepts are not operations, there is still some value in asking how something might be measured, as long as we do not infer too hastily to meaninglessness if the operations prove elusive. What operations might measure strength of belief? Here we face the awkward realization that only one operation has been proposed: measurement through monetary bets accepted or refused. Why must we accept just this? Why is money the measure of belief?

The answer is that nothing forces the acceptance in general. However there are quite specific circumstances in which we are forced to make money the measure of belief. The most obvious arises if we are wagering in a casino or racetrack. Others arise if we are buying or selling insurance. We must determine what premium is appropriate now as insurance against some uncertain, future harm whose gravity will be measured monetarily.[145] In the financial futures market, one can buy a contract that allows purchase of some asset at a fixed price at some later date. For example, an airline, fearing an increase in jet fuel prices, might buy a contract that enables purchase of jet fuel at present prices, but at a later date. Whether the later purchase will be made depends on the unknown of whether jet fuel prices will rise or fall. Thus pricing the contract requires the same sort of judgments over uncertainties as insurance and wagering.

Something close to a fair bet is also realized in these circumstances. In casinos and racetracks, every wager bought is sold by someone. In insurance, every policy, bought by someone, is sold by someone else. In the futures market, every contract bought by one trader is

---

[144] Recounted in Weirich (2010, p. 246).

[145] Starting in 1931, de Finetti had worked for an insurance company. Might this explain why he found it trivial that monetarily rewarded betting behavior measures belief?

sold by another. These transactions would comprise fair bet if we neglect a small spread between the buying and selling prices and the house's small margin in casino gambling.

Viewed materially, if we are in any of these circumstances, then pragmatic goals will force us to reason inductively as the framework of the Dutch book argument requires. The facts of the circumstances warrant the resulting logic. It provides a nice illustration of how the material theory of induction is applied. When we move to other circumstances, however, when facts of this type are missing, nothing warrants an inductive logic in which strengths of support must conform to defensive gambling strategies. Viewed material, Dutch book argumentation fails to establish the universal rationality of probabilistic inference precisely because the factual presumptions of the Dutch book scenarios do not hold universally.

Hájek (2008) has proposed an amusing device that we can use to underscore the dependence of the logic on the background conditions. If an agent has incoherent betting quotients, a benevolent bookie can offer the agent a combination of bets that assures a gain, a "Czech book." If, for example, the agent's betting quotient q is greater than one for some proposition A, the bookie would offer the agent a bet against A. Since $S<0$, the agent will make a gain $S(1-q)>0$ if A turns out to be true; and a loss of $-qS>0$ if A turns out to be false. That is, if one finds oneself in the clutches of a benevolent bookie, coherent betting quotients would be the only thing preventing the benevolent bookie providing you an assured gain!

12.3 The Rationality of Refusing to Bet

A sharper expression of the coercive presumption of the betting scenario is provided by a common response to it: it can be an expression of rationality for agents simply to refuse to bet. In the abstract, this refusal may seem like a crafty evasion. That it need not be is easier to see if we consider how a bookie might seek to force a Dutch book on someone whose beliefs conform with completely neutral support (5). Consider the case of three mutually exclusive outcomes A_1, A_2, A_3, such as arises in the wine and water problem. An agent whose credences conform with (5) would judge all of the following to be equally supported:

$$A_1, A_2, A_3, \text{not-}A_1 = A_2 \text{v} A_3, \text{not-}A_2 = A_1 \text{v} A_3, \text{not-}A_3 = A_1 \text{v} A_2$$

Assume *per impossibile* that there is some bet on A_1 that is acceptable to the agent as fair. Since the agent regards not-A_1 as equally supported, the agent will accept as fair a bet on not-A_1 with the same payoffs. Since a bet "on" A_1 is just the same as a bet "against" not-A_1, one can see that the two net payoffs, $S(1-q)$ and $-qS$ must be numerically equal, but different in sign.[146] For

[146] Suppose that the bet "on" A_1 pays a net of $X>0$, if A_1 is true, and $Y<0$, if A_1 is false. Then the bet "against" A_1 pays a net of $-X<0$, if A_1 is true, and $-Y>0$, if A_1 is false. The bet with the

simplicity, assume that the bet "on" A_1 pays a net of 1, if A_1 is true, and -1, if A_1 is false. The agent will judge similar bets fair for A_2 and A_3.

The three bets on A_1, A_2 and A_3 combined form the Dutch book in Table 2:

	Bet on A_1 pays:	Bet on A_2 pays:	Bet on A_3 pays:	Net payoff
A_1 is true.	+1	-1	-1	-1
A_2 is true.	-1	+1	-1	-1
A_3 is true.	-1	-1	+1	-1

Table 2. Dutch Book for an Agent with Beliefs conforming with Completely Neutral Support

What we cannot conclude from this Dutch book is that the assignments of support by the agent are irrational. They were determined as the only assignments compatible with the invariances of the system in question. If this Dutch book impugns the rationality of these assignments, then all we can conclude is that no rational treatment of systems like von Mises' wine and water is possible.

The obvious alternative is to recognize that someone who harbors assignments of support like those of (5) should not accept bets in accord with the rules specified in the Dutch book gambling scenario. For such an agent's credences are in conflict with the assumptions of the scenario. The irrationality lies not in the assignment of beliefs but in the indiscriminate acceptance of bets devised using those rules. Here I concur fully with the assessment of Bacchus et al. (1990, pp. 504-505) who argue:

> …that an agent ought not to accept a set of wagers according to which she loses come what may, if she would prefer not to lose, is a matter of deductive logic and not of propriety of belief.

12.4 Circularity in the Notion of a Fair Bet

Consider again a key assumption in the Dutch book argument: for any proposition A, an agent can find a fair bet with payoffs comprising those of Table 1; and the associated betting quotient q is the strength of belief in A. This may seem like a benign preliminary before the real work of assembling a Dutch book begins. However it is not. That assumption in effect already has the axioms of the probability calculus built into it and an excursion in repeated betting shows it.

---

same stakes "on" not-A_1 pays $X>0$, if not-A_1 is true, and $-Y>0$, if not-A_1 is false. These last two bets can only be same if, $X=-Y$.

To see it, consider some set of atomic propositions $A_1, A_2, \ldots, A_n$ and their Boolean combinations over which an agent distributes belief. Imagine that there are repeated scenarios in which there is a like set of propositions over which the agent distributes the same beliefs. Call the corresponding propositions of the form A_1 in each scenario "like propositions"; and so on for the remaining $A_2, \ldots, A_n$.

The obvious example is provided by the two propositions that a tossed coin shows heads (A_1) or that it shows tails (A_2). The repeated scenarios are then just independent tossing of many coins. For another example, we might consider the propositions that someone named in a telephone directory was born on Monday (A_1), or born on Tuesday (A_2), or born on some Boolean combination of days, such as (not-Monday and not-Friday) = (not-A_1 & not-A_5.) We create scenarios with identical beliefs over like propositions by scanning down a list of names in a telephone directory and asking after the birthday of each person named.

Since the agent has the same belief in the truth of each of the like propositions in the corresponding sets, the agent can execute the same bet on each like proposition. That is the agent's betting quotient for proposition A_i in each scenario is the same value q_i for what the agent judges to be a fair bet with the same fixed stake S_i in each case.[147] Assuming that there are W_i wins and N-W$_i$ losses among N bets, we find for the bets on proposition A_i:

$$\text{total payoff}_i = W_i(1-q_i)S_i - (N-W_i)q_iS_i = (W_i - Nq_i)S_i$$

From this we compute the average payoff per wager in terms of the frequency $r_i = W_i/N$ with which the propositions turn out to be true:

$$\text{average payoff}_i = (W_i - Nq_i)S_i/\text{N} = (r_i - q_i)S_i$$

To proceed we need to separate two cases. These frequencies r_i may or may not stabilize to definite limiting values as N grows indefinitely large. In the first case, we can define the limiting frequency as

$$p_i = \lim_{N\to\infty} r_i$$

It would be natural to identify this limiting frequency p_i with the probability of truth among the propositions A_i, for, if there is such a probability, the law of large numbers assures us that, with probability one, it will be revealed as this limit. However to arrive at the results that interest us, we do not need to do this. We can simply treat the p_i as parameters that have the specific property of importance here. Since they are derived from relative frequencies, they conform with the axioms of the probability calculus. That is, they are non-negative, additive for mutually

[147] What follows is an analysis concerning the atomic propositions. An analogous analysis can be applied to the propositions that are Boolean combinations of them.

exclusive outcomes and normalize to unity. For example, the limiting frequencies of truth p_i among the repetitions of the atomic propositions A_i always sum to unity:

$$p_1 + p_2 + \ldots + p_n = 1.$$

We can now see that the following two propositions are equivalent where the same set of betting quotients q_i is indicated in each proposition:

(a) There are fair betting quotients q_i such the agent fares equally well by making all the bets over A_i "on" bets with $S_i>0$; or by making all the bets over A_i "against" bets with $S_i<0$.

(b) There are betting quotients q_i that equal the limiting frequency of truth p_i among the propositions A_i (so that these betting quotients conform with the axioms of the probability calculus).

To infer from (a) to (b), note that "fares equally well" means that "on" and "against" betting yields the same results concerning payoffs. It follows that the limiting average payoff must be unaltered when we merely change the sign of S_i from positive to negative, where

$$\text{Limiting average payoff}_i = (p_i\text{-}q_i)S_i$$

If we interpret the parameters p_i as probabilities, this limiting average payoff is just the expected payoff per bet. Now, the limiting average payoff is linear in S_i. So it can only remain unchanged under an alternation of the sign of S_i if it is zero. That is $(p_i\text{-}q_i)S_i = 0$. It follows immediately that $p_i = q_i$. That is, we have inferred (b). The reverse inference from (b) to (a) follows by taking the steps of the inference in reverse order: $p_i = q_i$ entails that both total and average payoffs are zero, so that bets "on" and "against" are equally attractive.

In the second case, there is no stable limit to the frequencies r_i as N grows indefinitely large. This is an uncommon case, but it can arise. We shall see it, for example, arising for outcomes of drawings from an infinite lottery in the chapter "Infinite Lottery Machines." It is the case that is unfavorable to probabilities and thus we might not expect that the assumption of the existence of fair betting quotients might still drive the quotients toward conformity with the axioms of the probability calculus. However they still do so in the following sense.

Since the frequencies r_i have no limiting value, there is no unique value for them unless we specify the specific number of repetitions N. Once it is specified, fairness of the bet on proposition A_i is implemented if the agent can pick a betting quotient q_i that matches the actual frequency of truth r_i among the set of like propositions A_i in the N repetitions. For then the average payoffs are the same for both "on" and "against" bets. Any other value of q_i will favor either the bets "on" or "against" the like propositions according to whether $q_i>r_i$ or $q_i<r_i$.

As the number of repetitions varies, the particular target set of frequencies of truth r_i will vary. But what will not vary is that the target set for the betting quotients q_i is always a set of frequencies. Frequencies obey the axioms of the probability calculus, but with the added restriction that they are rational number valued. Thus we have weaker analogs of the equivalent propositions (a) and (b) for the same set of betting quotients q_i in each proposition:

(a') For some fixed set or repetitions N, there are fair betting quotients q_i such the agent fares equally well by making all the bets over A_i "on" bets with $S_i>0$; or by making all the bets over A_i "against" bets with $S_i<0$.

(b') For some fixed set or repetitions N, there are betting quotients q_i that equals the frequency of truth r_i among the propositions A_i (so that these betting quotients conform with the axioms of the probability calculus).

The proof of the equivalence of (a') and (b') is analogous.

In sum, for both cases, the assumption that there are fair betting quotients in the context of repeated betting scenarios is equivalent to assuming that these betting quotient behave like frequencies, that is, they conform with the axioms of the probability calculus. Thus one should not think that the assumption of fair betting quotients is an innocent background assumption. It does not merely provide a context in which Dutch book argumentation can prove that credences must conform with the axioms of the probability calculus. Rather conformity with those axioms is already tacitly presumed by them. All the Dutch book argumentation does is to make that conformity visible.

This outcome may be untroubling to someone who already believes that credences must be probabilistic. Why be troubled with a demonstration that just clarifies the probabilist's commitments? If, however, you are someone like me who does not believe that credences must be probabilistic, you will find this result damning. What was supposed to be a demonstration of the incoherence of non-probabilistic beliefs turns out to be an exercise in circularity. Probabilities are demonstrable merely because they were introduced covertly in an assumption of the argument at the outset.

12.5 The Regresses Begin

The prediction of the general analyses in earlier sections is that recognition of weaknesses in an attempted proof of probabilities leads to a regress. One form is a successive weakening of what is sought to be proved. This form of regress is well underway for Dutch book arguments. For it has been long recognized in the literature that the assumption of fairness is arbitrary and can be discarded without compromise to the rationality of the enterprise. This

recognition is at least a half century old, extending as far back as Smith (1961). It is the basis of the analysis of Walley's (1991) treatise *Statistical Reasoning with Imprecise Probabilities.*

Following Walley (1991, p. 28), it may be quite prudent for an agent to refuse to admit any bet over some proposition A as fair. Rather, the agent may be willing to accept a bet "on" A with a maximum betting quotient of q_{lower} and be willing to accept a bet "against" A with a minimum betting quotient of q_{upper}. If the two are equal, then they comprise a fair bet over A. If $q_{lower} < q_{upper}$, then the agent is more cautious in the agent's betting behavior. No Dutch book can be made against such an agent. The agent's belief in A is no longer a single probability, but an interval bounded by a lower probability equal to q_{lower} and an upper probability equal to q_{upper}.

That betting quotients $q_{lower} < q_{upper}$ betoken caution becomes most evident in the extreme case in which $q_{lower} = 0$ and $q_{upper} = 1$. In this extreme case, the agent is willing only to accept individual bets for which no loss is possible.[148] The interval of probabilities is maximally large, bounded by 0 and 1. For this reason, Walley (1991, p. 66) associates this state with vacuity or maximum ignorance.

To support this discarding of the necessary existence of fair bets, Walley (1991, p. 3) decries what he calls "the Bayesian dogma of precision"—"that uncertainty should always be measured by a single (additive) probability measure…" He continues:

> For example, de Finetti assumes that for each event of interest, there is some betting rate that you regard as fair, in the sense that you are willing to accept either side of a bet on the event at that rate. This fair betting rate is your personal probability for the event. More generally, we take your lower probability to be the maximum rate at which you are prepared to bet on the event, and your upper probability to be the minimum rate at which you are prepared to bet against the event. It is not irrational for you to assess an upper probability that is strictly greater than your lower probability. Indeed, you ought to do so when you have little information on which to base your assessments. In that case we say that your beliefs about the event are indeterminate, and that (for you) the event has imprecise probability.

---

[148] The agent is willing to make a bet "on" A only if the bet pays a net of S-0.S = S>0, if A is true; and with a payoff of -0.S=0 if A is false. The agent is willing to accept a bet "against" A only if the bet pays S – 1.S=0 if A is true; and with a payoff of -1S = -S > 0 if A is false, since S<0 for an "against" bet.

This work is motivated by the ideas that the dogma of precision is mistaken, and that imprecise probabilities are needed in statistical reasoning and decision.

There are two ways to understand the import of this relaxation of the conditions of the betting scenarios. The correct way, in my view, is merely to regard the various betting scenarios envisaged as circumstances that may or may not arise in different domains. There is no necessity for their implementation everywhere. All we can say is that *if* an agent is in a circumstance in which the assumptions of the scenario are realized, *then* reasoning inductively according to the prescribed system is their best course. Such circumstances arise, as noted above, in the insurance and futures market. Indeed the slight spread between the buying and selling prices in both cases suggests that Walley's imprecise logic is the appropriate one.

The incorrect way to understand the import of this relaxation is to think of it as a successful purging from the Dutch book analysis of an unwarranted element, the necessary existence of a fair bet, so that the analysis that remains is universally applicable. That would merely be to replace the dogma of precision by the dogma of imprecision. For there is no necessity in the presumption of strict upper and lower limits on the betting quotients or even that having beliefs requires their operational manifestation in betting behavior. With this understanding, we have taken the first step in the regressive weakening of what is sought to be proved, as described in Sections 4 and 9 above.

The other sort of regress arises when we retain what is sought to be proved, but seek to strengthen the grounds used in the proof. This is how I see Savage's (1954) decision theoretic proof of probabilities. It seeks, as does the Dutch book argument, to infer from an agent's preferences to the beliefs that must conform with them and thereby show them necessarily to be probabilistic. Savage does acknowledge (p. 4) inspiration from de Finetti's (1937) work. Now de Finetti simply posits certain betting behaviors and his posits are, as we have seen above, quite susceptible to challenge. I read Savage's analysis as an attempt to provide a more secure grounding for this approach.

Savage's full theory is based on seven postulates. Since they entail the same contingent result that beliefs are probabilities, they must contain that contingency in one form or another. Once again, careful scrutiny should reveal its presence. Because of the complexity of Savage's system, detailed analysis is precluded here. However we can discern the direction of the analysis by considering just the first postulate. It asserts, in effect, (p. 18) that the relation of preference over acts is a total order, which means that it is antisymmetric and transitive.

Consider the transitivity of preference. It says that if you strictly prefer A to B and B to C, then you must prefer A to C. (Antisymmetry precludes you also strictly preferring C to A.) This form of transitivity is important in the system. It provides an order that, when filtered through the other postulates, orders strengths of beliefs and eventually enables them to be real valued.

Savage provides no argument to preclude intransitivity when the postulate is introduced. He merely announces that "the definition of preference suggests …" (p.18). If you are antecedently disposed towards probabilities, it is quite comfortable to accept the suggestion and let the reasoning lead to the result you expect. However if you are not so disposed, you will have seen no good reason in the account that precludes intransitive preferences. Say I prefer eating apple pie to cherry pie and cherry pie to apricot pie. Aside from unsupported declaration in the definition of preference or in notions of rationality, nothing precludes me from preferring apricot pie to apple pie. But that would be an intransitive set of preferences.

It was soon recognized that more was needed if this prohibition on intransitivity was to be sustained. That is, the regress of reasons continues. The instrument that sustains it came to be known as the "money pump" argument, which appears in the literature as early as Davidson et al. (1955, pp. 145-46). Assume an agent harbors intransitive preferences, captured compactly by the obvious notation: $A>B, B>C, C>A$. Presumably the agent could be induced to trade a C to gain a B, while paying some small price, such as \$1; and a B to gain an A, for \$1; and an A to gain a C for \$1. The net effect is that the agent has paid \$3 to be returned to the original C. This, it is supposed, makes the intransitive preferences "irrational."

Once again we have an argument that can only be convincing to someone who already believes that there is some irrationality in intransitive preferences. Someone who does not believe this will have no trouble seeing that the irrationality is not in the intransitivity of the preferences. Rather it lies in the agent engaging in free trading with a second commodity (money) over which the agent's preferences are transitive. That trading behavior is dangerous and should be avoided; and that is all the money pump argument shows. Maher (1993, p. 36) puts it well:

> This is such a simple and vivid argument that it is a pity it is fallacious. But fallacious it is. The fallacy lies in a careless analysis of sequential choice. The argument assumes that someone with intransitive preferences will make each choice without any thought about what future options will be available, yet this is not in general a rational way to proceed. For example, …"

13. Necessary Conditions

As another briefer illustration of the inevitable failure of the proofs of necessity of probabilities described in Section 10 above, consider the approach taken by Cox (1961) and Jaynes (2003). The general approach is both elegant and appealing. Necessary conditions are laid down for a structure called "$\mathbf{i}|\mathbf{h}$" (Cox) or "$A|B$" (Jaynes) which represents the strength of

support of the first (**i** or *A*) afforded by the second (**h** or *B*). From them, by some simple but powerful functional analysis, the computational rules of the probability calculus are derived.

Precisely because these computational rules can be derived, the assumptions used must be logically at least as strong as them. Since the conclusion is contingent, so also are the assumptions. Any hope that the assumptions might somehow be self-evident will fail under scrutiny. Sustaining the proof will then trigger a regress of reasons each of which fails in the sense that the new reasons themselves need further support. We shall see this regress begin with Cox first positing the necessary conditions with short justifications. The inadequacy of the justifications becomes clear. Jaynes then intervenes and provides a stronger justification; and then sometimes when that stronger justification proves inadequate, yet another is provided, but still without arriving at a satisfactory end point.

There are three necessities: that the strengths are real values and what Jaynes calls the sum and product rules. We shall look at each in turn.

First, Cox (1961, p. 1) introduces the idea that the strengths are real valued with some rather casual remarks about their measurability. As an analogy he mentions the measurability of the pitch of a stairway.[149] Jaynes is rightly not satisfied with such a casual development. He makes the requirement explicit as his first desideratum (p. 17)

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

Jaynes first seeks to establish that it is satisfied by means of his parable of a robot (pp. 8-9) who will compute with the degrees. He then asserts (p. 17):

> 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.

Of course this is incorrect. A robot can represent and compute with all sorts of magnitudes and relational structures. To presume otherwise suggests willful ignorance, if someone has a minimal understanding of computers. To establish that the magnitudes treated are real numbers, we must assume a quite extensive list of specific properties including a transitive order ("greater than") and universal comparability under this order of all the magnitudes.

Tribus (1969, Ch. 1) develops a similar account that includes the parable of the robot. He reports (p. 6) drawing on Cox (1961) and unpublished course notes by Jaynes. His remark (p. 13) is:

> The only general way in which objects may be compared with one another is to assign to the objects a real number. The real number system provides the only scale of universal comparability.

---

[149] Cox's (pp. 29-34) later remarks on measurement pertain not to whether the strengths have real valued magnitudes, but whether they can be assessed with precision.

It is easy to see that one might let this pass if one already believes that the strengths of support must be probabilities. Otherwise it is baffling that such a claim could be made.

The regress of reasons continues. Jaynes presumably recognized the weakness of the robotic justification and included an Appendix (pp. 656-59) designed specifically to strengthen it. He notes that if an order on the strengths is transitive and universal, then, in the case of a finite outcome space, real valued degrees can be adapted to it. He proceeded to argue rather ineffectively for both transitivity and universality. Counterexamples to transitivity can be readily constructed, as in Norton (2007a, pp. 149-50). Keynes (1921, Ch. III) had long before realized that we must take seriously the possibility of incomparable degrees. More troublesome is that transitivity and universal comparability are insufficient to assure real values strengths.

Cox's second necessity is expressed as (p. 3):

> The probability of an inference on given evidence determines the probability of its contradictory on the same evidence.

Cox's justification is brief. He gives a few simple examples (p. 2) and announces that "in this all schools can agree." Jaynes' treatment is similarly hasty and incomplete. He declares (p. 30) "The plausibility that A is false must depend in some way on the plausibility that it is true." He proceeds immediately to conclude the much stronger result that there must be a functional relation of dependence between $A|B$ and not-$A|B$ and even that "common sense requires [the function] to be a continuous monotonic decreasing function…"

Once again, all these suppositions can pass without objection if one already has the goal of additivity of strengths in mind. That is, one might imagine that these necessities are simply reduced descriptions of the rule in the probability calculus that $P(A|B) + P(\text{not-}A|B) = 1$. If one is not antecedently committed to this rule or something like it, these necessities will appear as unfounded stipulations. One need only consider superadditive measures to find all the conditions laid down by Cox and Jaynes violated.

Cox's third necessity is (p. 4):

> The probability on given evidence that both of two inferences are true is determined by their separate probabilities, one on the given evidence, the other on this evidence with the additional assumption that the first inference is true

Its content is more easily grasped if we give it in symbolic form, as does Jaynes (p. 25). The support for the conjunction of A and B on the evidence C, $(AB|C)$ is some function F of two other strengths $(B|C)$ and $(A|BC)$:

$$(AB|C) = F[\ (B|C)\ ,\ (A|BC)\ ] \qquad (6)$$

To someone remote from probability theory, this functional stipulation will appear quite arbitrary. To probabilists, it immediately calls to mind the product rule for forming conjunctions

$$P(AB|C) = P(B|C)\ .\ P(A|BC)$$

So, for them, it can pass as reasonable and even natural. Both Cox and Jaynes seek to establish this functional dependence by recalling informal sequences of inferences. If we are to infer to (A and B) from C, we might first establish from the truth of C that B is true. Then we would establish from the truth of (C and B) that A is truth and so also that the conjunction (A and B) is true. (For later reference, represent this as "$C \rightarrow B \rightarrow A \rightarrow AB$.") This sequence is one way that we might proceed deductively. Cox and Jaynes then declare that the functional dependence (6) follows since it mimics the same order of steps.

The inference is quite dubious. Indeed one of the lessons of twentieth century philosophy of science was that transferring properties of deductive inference over to inductive inference regularly produces incorrect rules. For example, if C deductively entails each of A and B separately, then C also deductively entails their conjunction. However the corresponding rule for induction fails. C may strongly support each of A and B separately, but actually refute their conjunction.

Presumably because they recognize the inadequacy of the arguments for (6) so far, Jaynes and Tribus (1969, Ch. 1) embark on a more elaborate demonstration. Its basic supposition is that $(AB|C)$ must be a function of some or all of the following four strengths only:

$$(A|C), (B|C), (B|AC), (A|BC).$$

They then argue that the only possibility is (6) or its equivalent form under relabeling, $(AB|C) = F[(A|C) , (B|AC)]$. Locally the argumentation is quite cogent. For example, $(AB|C)$ cannot depend functionally on just $(A|C)$ and $(B|C)$. For each of A and B may be strongly confirmed by C, but C may either confirm or even refute (A and B).

However this next attempt to buttress the functional dependence of (6) fails. For the assumptions used are still far too strong and likely only unobjectionable if one already accepts the final result. The lacunae are both narrow and broad. Narrowly, consider the details of the functional dependencies. They infer that $(AB|C)$ can depend functionally on $(A|C)$ and $(B|AC)$; or it can depend functionally on $(B|C)$ or $(A|BC)$. These dependencies are analogous to the two deductive pathways "$C \rightarrow B \rightarrow A \rightarrow AB$" and "$C \rightarrow A \rightarrow B \rightarrow AB$." Since either individually suffices in the deductive case, Jaynes seems to presume that either will also suffice in the inductive case. That certainly does not follow since the analogies between deduction and induction are fragile. They have not ruled out the case that both inductive pathways must enter into the functional dependence, which would mean that $(AB|C)$ is still a function of all four strengths listed.[150]

[150] Tribus (1969, pp. 16-17) has an argument against this possibility that appears flawed. He seems to argue that it is ruled out since $(B|AC)$ becomes ill-defined when C is not-A. But this sort of difficulty is routinely overcome by allowing that in some special cases the function is ill-defined.

Taking a broader, synoptic view, the most obvious lacuna is the assumption that ($AB|C$) must be a function of the four strengths listed. They might be related, but must the relationship be functional? Might there not be a more complicated relationship? Perhaps there is one that involves some auxiliary quantity and the strengths ($A|C$) are derived from them? Or might there simply be no definite relation at all? This last possibility would then mimic the situation with superadditive measures. They decouple the values of ($A|C$) and (not-$A|C$), so that there is no functional relation between them. Each value of ($A|C$) may be compatible with many (not-$A|C$); and conversely.

14. Conclusion

The approach taken in this chapter pursues the two lines of criticism indicated. There are more grounds for hesitation over probabilities. My (Norton, 2011) reviews many of these. Perhaps the best known and most intractable of these problems is the problem of the priors. It arises from the need for a Bayesian analysis always to provide some prior probability, $P(H|B)$, antecedent to the consideration of evidence. The very fact that they must be provided in this way introduces an arbitrariness into the analysis that has been the bane or all forms of Bayesianism. Objective Bayesians try to find good reasons for picking a particular prior, such as through Jaynes' ill-chosen[151] maximum entropy principle. Subjective Bayesians try to avoid the problem by demoting the prior probability to mere opinion, which can be freely chosen. Theirs has proven to be a poor bargain since, once one allows opinion to be mingled with evidential warrant, they prove virtually impossible to separate.

The necessity for prior probabilities is a form of incompleteness of the inductive logic: those priors always supply inductive content that is beyond the reach of the evidence to be considered subsequently. One might imagine that it is a problem peculiar to the probability calculus so that the best escape is to find another calculus free of the problem. In recent work, I have shown that this escape fails. The sort of incompleteness that troubles the probability calculus must arise in a large class of calculi of induction that include all those we would

[151] The principle tells us to distribute our prior probabilities as uniformly as the external constraints allow. Thus it is an extended form of the principle of indifference. If there are no constraints other than conformity with the probability calculus, maximizing entropy reduces to choosing the uniform probability distribution required by the original principle of indifference. This principle, as we have seen in Sections 6, 7 and 8, is an insecure basis for reasoning within the probability calculus since it rapidly produces results that contradict the calculus.

reasonably entertain. An informal development of this result is provided below in the Chapter "Incompleteness of All Calculi of Inductive Inference."

What I have sought to establish in this chapter is that the probability calculus does not supply a universally applicable logic of inductive inference. The emphasis here is on the universal applicability. I do not doubt the utility of Bayesian analysis in specific domains in which background facts positively warrant it. My hope is that, if they have it, Bayesians can relinquish tacit commitments to the idea that "It's all probabilities." and to the notion that this idea solves the foundational problems of inductive inference. For then we can address these foundational problems anew and, it is to be hoped, find better solutions. Readers, of course, will know that I offer the material theory of induction as my solution to the foundational problem of the nature of inductive inference.

If we are loosening tacit Bayesian commitments, there is a second one that can be relaxed profitably. The general view seems to be that the probability calculus must be accepted or rejected as a whole. In contrast, I have urged in that we can be more selective. Norton (2007a) provides an axiomatization of the probability calculus, using familiar techniques. Its novelty is that it was designed explicitly to identify qualitative properties of support relations that may be employed selectively. The most important result is that there are two components in these qualitative properties, that the two are readily separated and that they can be deployed individually as circumstances demand.

The first is a property I call "addition" which captures the additivity of the calculus. It resides in a reciprocal relation between the support accorded to a proposition and to its negation. Addition is an appropriate property when degrees of support span from positive to negative support. It should be dropped, however, if neutral support is to be represented.

The second property, "Bayes property," provides the probability calculus with the updating dynamics characteristic of Bayesian analysis. It depends on a particular mode of updating in which the import of evidence is simply to refute disjunctive parts of the hypothesis logically incompatible with the evidence; and then to redistribute support uniformly.

Many of the successes of Bayesian analysis can be traced back to these properties. Since conditions may favor the use of one but not the other, their utility can only be increased if we decide to employ them separately, for then they can be used more widely. For example, the completely neutral support described in this chapter contradicts additivity, but it is compatible with the Bayes property. Thus an extension of the theory of completely neutral support will permit updating by a Bayesian-style dynamics.

References

Bacchus, Fahiem; H. E. Kyburg, Jr., Henry E.; and M. Thalos, Miriam (1990) "Against Conditionalization," *Synthese*, **85**, pp. 475-506.

Benétreau-Dupin, Yann (2015) "The Bayesian Who Knew Too Much," *Synthese*, **192**, pp. 1527-42.

Bostrom, Nick (2002) *Anthropic Bias: Observation Selection Effects in Science and Philosophy.* New York: Routledge.

Bradley, Seamus (2016) "Imprecise Probabilities," *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2016/entries/imprecise-probabilities/

Bridgman, Percy W. (1927) *The Logic of Modern Physics*. New York: MacMillan.

Cox, Richard T. (1961) *The Algebra of Probable Inference*. Baltimore: The Johns Hopkins Press.

De Finetti, Bruno (1937) "Foresight: Its Logical Laws, Its Subjective Sources," pp. 134-74 in S. Kotz and N. L. Johnson, eds., *Breakthroughs in Statistics. Volume 1. Foundations and Basic Theory*. New York: Springer Verlag, 1992.

Davidson, Donald; McKinsey, J. C. C.; and Suppes, Patrick (1955) "Outlines of a Formal Theory of Value, I," *Philosophy of Science*, **22**, pp. 140-160.

Eva, Benjamin (manuscript) "Principles of Indifference."

Hájek, Alan (2008) "Arguments for–or against–Probabilism?" *British Journal for the Philosophy of Science*, pp. 793–819.

Hájek, Alan (2009) "Dutch Book Arguments," pp. 173-195 in Paul Anand, Prasanta K. Pattanaik, Clemens Puppe, eds., *The Handbook of Rational and Social Choice: an Overview of New Foundations and Applications*. Oxford : Oxford University Press.

Jaynes, Edwin T. (2003) *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

Keynes, John Maynard (1921), A Treatise of Probability. London: Macmillan; reprinted, New York: AMS, 1979.

Kolmogorov, Andrey (1950) *Foundations of the Theory of Probability*. Chelsea: New York.

Maher, Patrick (1993) *Betting on Theories*. Cambridge: Cambridge University Press.

Norton John D. (2007) "Disbelief and the Dual of Belief," *International Studies in the Philosophy of Science,* **21**, pp. 231-52.

Norton, John D. (2007a) "Probability Disassembled," *British Journal for the Philosophy of Science*, **58**, pp. 141-171.

Norton, John D. (2008) "Ignorance and Indifference," *Philosophy of Science*, **75**, pp. 45-68.

Norton, John D. (2010) "Cosmic Confusions: Not Supporting versus Supporting Not,"*Philosophy of Science*, **77**, pp. 501-523.

Norton, John D. (2010a) "Deductively Definable Logics of Induction." *Journal of Philosophical Logic*. **39**, pp. 617-654.

Norton, John D. (2011) "Challenges to Bayesian Confirmation Theory," *Philosophy of Statistics, Vol. 7: Handbook of the Philosophy of Science*. Prasanta S. Bandyopadhyay and Malcolm R. Forster (eds.) Elsevier

Pettigrew, Richard (2016) *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Ramsey, Frank P. (1926) "Truth and Probability," in *The Foundations of Mathematics and other Logical Essays*, Ch. VII, pp.156-198, ed. R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company, 1931.

Savage, Leonard J. (1954) *The Foundations of Statistics*. John Wiley & sons. Revised ed., New York: Dover, 1972.

Shafer, Glenn (1976) *A Mathematical Theory of Evidence.* Princeton, NJ: Princeton University Press.

Smith, Cedric A. B. (1961) "Consistency in Statistical Inference and Decision," *Journal of the Royal Statistical Society. Series B*, **23**, pp. 1-37.

Tribus, Myron (1969) *Rational Descriptions, Decisions and Designs*. New York: Pergamon.

Van Inwagen, Peter (1996) "Why Is There Anything at All?" *Proceedings of the Aristotelian Society* **70** (suppl.), pp. 95–120.

von Mises, Richard (1957), *Probability, Truth and Statistics*. London: George Allen & Unwin.

Vineberg, Susan (2016) "Dutch Book Arguments," *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2016/entries/dutch-book/.

Walley, Peter (1991) *Statistical Reasoning with Imprecise Probabilities.* London: Chapman and Hall.

Weirich, Paul (2011) "The Bayesian Decision-Theoretic Approach to Statistics," pp. 233-61 in Prasanta S. Bandyopadhyay and Malcolm R. Forster, eds., *Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.* Amsterdam: Elsevier.

Field to initialize equations:

May 12, 16; June 15, 2017
Preamble removed August 7, 2017
Minor typo corrections June 20, 2018

Chapter 11

Circularity in the Scoring Rule Vindication of Probabilities[152]

1. Introduction

The last chapter argued that all proofs of the necessity of probabilities fail. They are deductive arguments for a contingent conclusion, that probabilities must be used to represent inductive degrees of support or subjective degrees of belief. Thus the proofs must employ premises that are deductively at least as strong as or even stronger than the conclusion sought, the necessity of probabilities. It follows that any proof of the necessity of probabilities can be undone merely by examining the premises of the proof and revealing the presence of the necessity of probability, in whatever congenial disguise is used to hide it. Moreover the last chapter predicted that any program of demonstration of the necessity of probabilities will be trapped forever in a cycle of near misses, corrections and renewed attempts, none of which ever succeed completely, for the program's goal is unattainable.

The present chapter offers an extended illustration of these conclusions through the recent literature that seeks to demonstrate the necessity of probabilities by means of considerations of accuracy alone, where accuracy here means quantifiable closeness to the truth. That closeness is in turn measured by numerical scoring rules, which will become the major focus of what follows. If these scoring rule vindications succeed, they have the potential of displacing the decision theoretic approaches, for the scoring rule approach has no need to envisage elaborate scenarios with agents adapting beliefs to decisions that maximize utilities. Credences are chosen simply by the criterion of accuracy. The approach depends on an appealing dominance argument: if our credences are not probabilistic, then they will always be dominated by probabilistic credences in the sense that, whatever may be the case, we improve accuracy by shifting from the non-probabilistic credences to the probabilistic credences.

The discussion below will proceed within the framework routinely employed by the scoring rule literature. Its suppositions include:

[152] I thank Joshua Fry, Lee Elkin and Richard Pettigrew for helpful discussion.

• credences in any two propositions are always comparable;

• the relation of comparison can be captured by a real-valued degrees in the interval 0 to 1.

Each of these and others like it also require justification; and attempts to justify them would in turn face just the same issues of circularity developed here.

The focus of attention in the analysis below will be the particular scoring rule employed to measure the accuracy of credences. We shall see that almost every slight change in the rule undoes the demonstration; and almost every larger change leads to a wide variety of alternative results. This fact shows that it is not the general notion of accuracy that drives the proof, for accuracy alone gives very little. Rather everything depends on the delicate selection of an accuracy measure tailored to give the desired result. Here is the circularity. It is in this delicate fine-tuning that the probabilistic credences are presumed in disguised form.

The response has been a flourishing of attempts to make the choice of the fine-tuned scoring rule seem necessary or inevitable or perhaps just natural. We find a regress of reasons that never quite terminates in success; or a proliferation of alternatives, each of which is replaced by another, without apparent end. This endless, frustrating dynamic is just what was predicted by the general argument against all proofs of the necessity of probabilities.

The exploration here of scoring rule approach will necessarily be partial. The literature on the topic is so large that a mere chapter can only scratch the surface. The goal is not to review every demonstration. Rather it is to display by example how the regress and proliferation of reasons comes about in this specific instance. In case after case, we shall see that plausible assumptions that initially appear independent of the assumption of the necessity of probabilities actually contain the assumption in covert form. An ardent vindicator will, no doubt, have further demonstrations that I have not discussed and may urge these as finally resolving all difficulties. I can only respond with some confidence as I would to a circle squarer or angle trisector: these further demonstrations would in turn succumb under scrutiny. For if they are to succeed, they must employ premises logically at least as strong as the conclusion sought.

The accuracy driven demonstration of the necessity of probabilities draws on a much larger literature in meteorology, economics and subjective Bayesianism that uses scoring rules for other purposes. These other uses will be sketched in Sections 2 and 3 below. They include the elicitation of true but secret probabilities from subjects who, we are to suppose, might otherwise not reveal them. In that context, the adaptation of scoring rules specifically to probabilities is benign, since these uses presume explicitly that credences are probabilistic. Use of these adapted rules in the newer context of the vindication of probabilities ceases to be benign for there we are no longer allowed to presume that all credences are probabilities: the circularity of vindication lies precisely in that adaptation.

The original form of the accuracy driven demonstration of the necessity of probabilities will be developed in Section 4. It employs a quadratic Brier scoring rule. This rule, we shall see, so favors probabilities that it rewards subjects with non-probabilistic credences for lying that their credences are probabilities. In Section 5, we shall see that the success of this original accuracy driven vindication depends on selection of exactly the Brier scoring rule and not any other in its neighborhood. When we replace the power of 2 in the Brier score formula by a more general exponent *n,* the slightest change in the exponent--a shift from 2 to 2.01 or to 1.99--is enough to undo the proof. Section 6 will reflect on how little in the original proof comes from the mere idea of accuracy, as opposed to the careful choice of scoring rule. Section 7 will review attempts to justify the restricted choice of scoring rule.

Sections 8 will describe the "strictly proper" scoring rules that have been introduced into the larger literature with a different purpose. They are a generalization of the Brier scoring rule, contrived to preserve its key property of favoring probabilities. Hence, as we see in Section 9, the success of strictly proper scoring rules in the dominance proof is to be expected. However that contrived favoring of probabilities is precisely how the proof can covertly assume probabilities at the outset. Section 10 will review the inevitable failure of attempts to justify independently the restriction to strictly proper scoring rules in the dominance analysis. Section 11 will remind us once again of the pitfalls of "natural" criteria. Section 12 has a short conclusion.

2 Origins in Frequencies

The present literature in scoring rules has origins in considerations of frequencies. Identifying them proves important in understanding what otherwise looks like arbitrariness in the systems now used.

In 1950, meteorologist and statistician Glenn Brier addressed a vexing problem in systems used to track the reliability of meteorologists' weather forecasts. The systems were leading meteorologists to deliver something other than their best forecasts in efforts to improve their ratings. They would, as Brier (1951, p.10) put it, be " 'hedging' or 'playing the system.'" For example, as Brier and Allen (1951, p. 843) note, if a temperature forecast must be given as a single number, the forecaster may choose to report different temperatures according to the statistic that would be used to measure the forecaster's reliability. If it was measured by a count of how many predictions proved exactly right, the best strategy is to report the most probable temperature. If reliability is measured by mean absolute error, then the best strategy is to report the median temperature. If reliability is measured by the root-mean-square error, then the mean

temperature is best. The forecaster's best judgment has been overshadowed by a concern for the performance measure.

Brier's solution was to propose an assessment system that would not reward efforts to play the system: the forecasts are given as probabilities and a "verification score"—later call the "Brier score"—is computed according to scheme in which higher scores represent poorer performance. If there are n possible, mutually exclusive weather condition, the forecaster predicts them with probabilities $x_1, \ldots, x_n$. The best forecasts are to be given the lowest scores. So, if condition i does not occur, a term in x_i^2 is added to the score. The higher is the probability x_i, the more defective the prediction and thus the worse, that is, the higher the score. Correspondingly, if condition k arises, a larger associated probability x_k should contribute less to the score. This is achieved by adding a term $(1- x_k)^2$ to the score. The final score P is recovered by averaging this sum over the N possible occasions over which the forecaster is scored.

Write x_{ik} for the probability predicted on occasion i for condition k. The actual outcomes are encoded in the matrix E_{ik}, where $E_{ik}=1$ encodes occurrence on occasion i of condition k; and $E_{ik}=0$ encodes its failure to occur. The "verification score" Brier proposed is

$$P = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n}\left(E_{ik} - x_{ik}\right)^2 \qquad (1)$$

At first, the choice of a reward $(1-x)^2$ for correct predictions and a punishment of x^2 seems arbitrary. One might imagine that almost any decreasing or increasing functions of x, respectively, would serve equally well. That turns out not to be the case, for this score has an important property shared by relatively few other scores, as we shall see in Appendix B below. The property appears in the case of N occurrences of some circumstance for which the same probability forecast x_k for condition k is appropriate for each occurrence. The frequency f_k of the k-th condition among the N occurrences is given by $f_k = \sum_{i=1,N} E_{ik}/N$. For this case, Brier (1951, p.2) described the key property:[153]

> It is also easy to show that if $[f_1, \ldots, f_n]$ are the relative frequencies that the event occurred in classes $1, 2, \ldots, [n]$, then the minimum score that can be obtained by forecasting the same thing on every occasion is when
>
> $$[x_{ik} = f_n]$$

In this special case, Brier's verification score reduces to

[153] The square brackets indicate minor changes from Brier's notation to mine.

$$P = f_1 (1 - x_1)^2 + f_2 x_1^2 + f_3 x_1^2 + \ldots + f_n x_1^2$$
$$+ f_1 x_2^2 + f_2 (1 - x_2)^2 + f_3 x_2^2 + \ldots + f_n x_2^2$$
$$+ \ldots$$
$$+ f_1 x_n^2 + f_2 (1 - x_n)^2 + f_3 x_n^2 + \ldots + f_n (1- x_n)^2 \qquad (2)$$

The optimal (minimum) score arises when the derivative of P with respect to each of the $x_1, \ldots, x_n$ vanishes: $dP/dx_1 = \ldots = dP/dx_n = 0$. An easy calculation shows the minimum occurs when:

$$x_i = f_i \qquad \text{for } i = 1, \ldots, n \qquad (3)$$

Brier predicted the effect of the use of this score on a forecaster (1950, p.2)

> A little experience with the use of the score *P* will soon convince him that he is fooling nobody but himself if he thinks he can beat the verification system by putting down only zeros and unities when his forecasting skill does not justify such statements of extreme confidence. And in the complete absence of any forecasting skill he is encouraged to predict the climatological probabilities instead of categorically forecasting the most frequent class on every occasion.

Two features of Brier's verification score are noteworthy. First, Brier assumed at the outset that the forecasters' predictions, both private and public, are probabilities. There are no weights that do not normalize to unity and thus need correction to bring them into conformity with the probability calculus. Second the score is designed to ensure that forecasters' probabilities are well calibrated in the sense that they are given the best scores when their forecast probabilities for the conditions match the frequencies of the conditions. In this calibration, the probabilities are calibrated to the *short-term* frequencies in *N* occurrences. These are not long-term, infinite limit frequencies, but the actual frequencies in a run of *N* occurrences, where *N* may be quite small.

3 Eliciting Credences

Brier used his score as a way of matching weather forecasts with short-term frequencies. At around the same time as Brier's work, a second literature sprang up in which the same devices were used for a different purpose. (See, for example, McCarthy, 1956; De Finetti, 1965; Savage, 1971; and de Finetti, 1974, Ch.5.) The literature addressed a subject who harbored certain credences or subjective probabilities and the task was to elicit those credences. The means was to assign a score to probabilities announced by these subjects. The Brier score is most commonly used, but not exclusively so. For example, Brier's score formula (2) is used but its terms are interpreted differently. The quantities x_i are the subject's announced probabilities and the

quantities f_i are the subject's true beliefs. Replacing frequencies f_i by probabilities p_i, we have a penalty function:

$$\begin{aligned} P = {} & p_1 (1 - x_1)^2 + p_2 x_1^2 + p_3 x_1^2 + \ldots + p_n x_1^2 \\ & + p_1 x_2^2 + p_2 (1 - x_2)^2 + p_3 x_2^2 + \ldots + p_n x_2^2 \\ & + \ldots \\ & + p_1 x_n^2 + p_2 x_n^2 + p_3 x_n^2 + \ldots + p_n (1 - x_n)^2 \end{aligned} \qquad (2a)$$

If the Brier score is a penalty that the subject seeks to minimize, the analog of (3) above shows that the subject does best by announcing the subject's true beliefs.

The literature presents different scenarios to motivate an interest in what otherwise seems an arcane scenario of dissembling subjects who may not announce their true subjective probabilities. Murphy (1956, p. 654) imagines a forecaster and a client. The client uses the penalty as a way to "keep the forecaster honest," where the quote marks are Murphy's. De Finetti (1965, §3; 1974, §5.5) is more detailed. He imagines scenarios in which an expert makes a probabilistic recommendation. A geologist, for example, may announce probabilities on the success of drilling an oil well at a particular site. We interest the geologist "*in giving an honest answer; in expressing his deep felt belief*"[154] by associating the score with the fee to be paid to the geologist on completion of the drilling. In another scenario, probabilistic bets are made on the outcome of sporting events and the payoff tied to the score. Finally, it is proposed that answers to multiple choice exam questions be given as probabilities and that the final score be computed as a Brier score.

For our purposes, however, minimizing the Brier score works *too* well. Our concern includes credences that may not be probabilities. Imagine that the true credences p_i of the subject are not probabilities. They are just a set of numbers $p_1, \ldots, p_n$ that do not sum to unity. The minimum of the penalty function P of (2a) occurs when the reported values $x_1, \ldots, x_n$ are not the true credences $p_1, \ldots, p_n$ but the true credences normalized to unity.

To see this, note that the minimum of (2a) with respect to varying x_i arises when we have $dP/dx_1 = \ldots = dP/dx_n = 0$. Thus we have:

$$\begin{aligned} 0 = dP/dx_1 &= -2 p_1 (1 - x_1) + 2 p_2 x_1 + 2 p_3 x_1 + \ldots + 2 p_n x_1 \\ &= -2 p_1 + 2 x_1 (p_1 + p_2 + p_3 + \ldots + p_n) \end{aligned}$$

and similar conditions for the remaining $x_2, \ldots, x_n$. Rearranging we have

$$x_i = p_i / (p_1 + p_2 + p_3 + \ldots + p_n) \qquad \text{for } i = 1, \ldots, n \qquad (3a)$$

The credences reported are the true credences renormalized, so they sum to unity.

---

[154] De Finetti (1974, p. 193; emphasis in original).

Thus, elicitation of true credences by means of a Brier score rewards subjects for lying and saying that their credences are probabilities, when they are not. This is an indication that the scoring method is biased towards probabilities, for it rewards a shift to probabilities, even when they are not the quantities sought.

4. The Dominance Argument

What is distinctive about this last literature is that, first, the elicitation is governed by pragmatic factors. The students' score best on an exam or the geologist will be paid the most if they reveal their true probabilistic credences. Second, the primary focus is the eliciting of credences, already assumed to be probabilities. It is not offered as a way of demonstrating that one's credences must be probabilities.[155]

A more recent development of this literature sought to alter both features. (See for example, Rosenkrantz, 1981, 2.2; Joyce, 1989, 2009; Pettigrew, 2016.) It produced an argument for the necessity of probabilities that is presently enjoying considerable popularity. The core idea is that credences should be distributed not on pragmatic grounds but in a way that optimizes the accuracy of the credences. The main result is that the accuracy of a non-probabilistic credence can always be improved by switching to probabilistic credences, no matter which outcome obtains

The simplest instantiation of the argument employs a Brier score. We have n mutually exclusive outcomes $E_1, \ldots, E_r$, over which credences $x_1, \ldots, x_r$, are distributed. All credences here and henceforth are restricted to the interval [0,1]. The original Brier score (1) or (2), (2a) is broken up into r component loss functions L_i, $i = 1, \ldots, r$, according to which of outcome $E_1, \ldots, E_r$ obtains:

$$\begin{aligned} L_1 &= (1 - x_1)^2 + x_2^2 + x_3^2 + \ldots + x_r^2 \\ L_2 &= x_1^2 + (1 - x_2)^2 + x_3^2 + \ldots + x_r^2 \\ &\ldots \\ L_r &= x_1^2 + x_2^2 + x_3^2 + \ldots + (1 - x_r)^2 \end{aligned} \qquad (4)$$

Greatest accuracy is achieved by minimizing these scores. Hence it is natural to characterize the quantities as "losses" to be minimized; and to think of an increasing loss score as a measure of increasing inaccuracy.

---

[155] For completeness, the devices needed are present. They are just not emphasized. The essential step of the dominance argument is mentioned in passing in the captions to Figure 1 and 2 of De Finetti (1965, p. 92) and Figure 5.3 of De Finetti (1974, p. 189).

The association of loss with inaccuracy derives from the loss generating functions used. That is, each loss function L_k, associated with outcome E_k obtaining, is a sum of r terms:

$$g_1(x_i) = (1 - x_i)^2 \quad \text{when } i = k$$
$$g_0(x_i) = x_i^2 \quad \text{when } i \neq k \tag{5}$$

Generating function $g_1(x_i)$ assures that a larger x_i makes a smaller contribution to the loss, for the case in which E_i obtains. Generating function $g_0(x_i)$ assures that a larger x_i makes a larger contribution to the loss in all the remaining cases.

With these loss functions (4), no matter which of $E_1, \ldots, E_r$ will obtain, we always improve accuracy by replacing a non-probabilistic credence with a probabilistic credence. The argument is seen graphically in the simplest case of two outcomes E_1, E_2, with credences x_1, x_2. Figure 1 shows the space of credences with individual points $\langle x_1, x_2 \rangle$, where both credences are restricted to values in [0,1]. On the left, the figure shows curves of constant loss L_1. They are circular arcs, centered on the corner point, $\langle x_1, x_2 \rangle = \langle 1,0 \rangle$, On the right, the figure shows the corresponding curves of constant loss L_2. The diagonal dashed line represents those credences conforming with the additivity of the probability calculus. That is, $x_1 + x_2 = 1$.

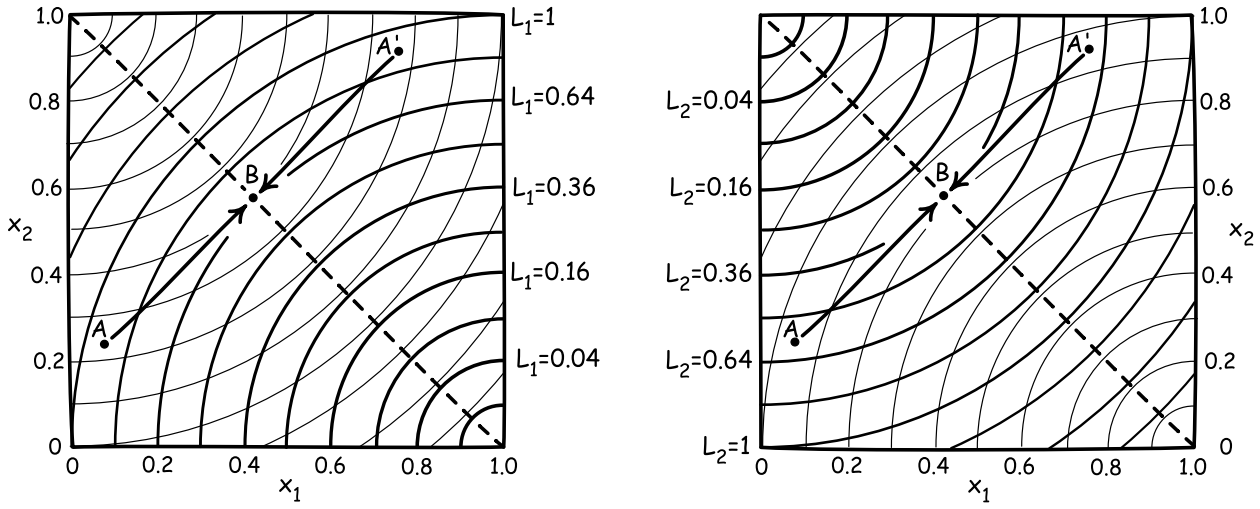


Figure 1. Dominance of probabilistic credences using a Brier score

Pick any point in the space not on this diagonal, such as point A. It represents credences that violate the additivity axiom of the probability calculus. If we move along line AB, perpendicular to the diagonal, to the point B on the probabilistic diagonal, we replace the non-probabilistic credences at A with the probabilistic credences at B. We see in the figure on the left, that replacing credences at A by those at B reduces the loss L_1. The same is true if we approach

probabilistic credence B from a corresponding non-probabilistic credence A', on the other side of the diagonal. That is, among all credences on the line AA', the probabilistic credence at B has the lowest loss L_1. That is, it is the most accurate among them if E_1 occurs. The same lines AB, $A'B$ are shown on the right. Once again, among all credences on the line AA', the probabilistic credence at B has the lowest loss L_2. It is the most accurate among them if E_2 occurs. That means that whichever of E_1 or E_2 occur, the probabilistic credence at B is the most accurate among all credences on the line AA'. Probabilistic credence B dominates: we achieve greater accuracy by replacing any non-probabilistic credence in AA' with a probabilistic credence B.

In both cases, what is key is the concavity of the curves[156] of constant loss towards the direction of smaller loss. Thus moving towards the diagonal of probabilistic credences moves us to credences of smaller loss.

The result generalizes to the case or r outcomes, $E_1, \ldots, E_r$. The easy way to see it is to identify a differential condition that expresses the dominance. In the case of two outcomes E_1 or E_2, each probabilistic credence $\langle x_1, x_2\rangle$ on the diagonal $x_1 + x_2 = 1$ dominates a set of non-probabilistic credences $\{\langle x_1+\mathrm{k}, x_2+\mathrm{k}\rangle\}$ where k can have any value, both positive and negative, that generates points within the space. Each such set forms a line, such as AA' of Figure 1, that is perpendicular to the diagonal of probabilistic credences and will intersect it at one dominating point. For the case of L_1 and L_2 restricted just to the set $\{\langle x_1+k, x_2+k\rangle\}$, the dominating point satisfies:

$$\frac{dL_1}{dk} = \frac{dL_2}{dk} = 0$$

We now give the same analysis for the case of r outcomes, $E_1, \ldots, E_r$. The hypersurface in the space of $x_1, x_2, \ldots, x_r$, corresponding to probabilistic credences is

$$x_1 + x_2 + \ldots + x_r = 1$$

Each such point $\langle x_1, x_2, \ldots, x_r\rangle$ dominates points in the set $\{\langle x_1+k, x_2+k, \ldots, x_r+k\rangle\}$, where k is both positive and negative as before. The dominating point will satisfy an extension of the differential condition above:

---

[156] To preclude confusion, "concavity" here simply reports that the curves of constant L_1 are geometrically concave towards the point that represents certainty of E_1's occurrence. The same property is described in Section 7 below, by standard convention, as the "convexity" of the function L_1. This usage presumably reflects geometrical convexity in the direction of increasing L_1.

$$\frac{dL_1}{dk} = \frac{dL_2}{dk} = \ldots = \frac{dL_r}{dk} = 0 \qquad (6)$$

To find the dominating point, we start with some point $< x_1, x_2, \ldots, x_r >$ in the set that is not necessarily the dominating point and seek the value of k that satisfies condition (6). L_1 expressed as a function of k is

$$L_1(k) = (1 - x_1 - k)^2 + (x_2 + k)^2 + (x_3 + k)^2 + \ldots + (x_r + k)^2$$

A short computation shows that the condition (6) for L_1 is satisfied when

$$k = (1 - (x_1 + x_2 + \ldots + x_r))/r$$

and, by the obvious symmetry in the formulae, the same value of k leads to satisfaction of condition (6) for the remaining loss functions.[157]

Thus the dominating point in the set has credences

$$X_i = x_i + (1 - (x_1 + x_2 + \ldots + x_r))/r$$

For $i = 1, \ldots, r$. It is easy to confirm that these dominating credences satisfy the additivity condition

$$X_1 + X_2 + \ldots + X_r = 1$$

That is, the dominating credence point $< X_1, X_2, \ldots, X_r >$ is probabilistic.

5. The Problem: Sensitivity to the Scoring Rule Chosen

The analysis as laid out in the last section shows a dominance argument that appears at once elegant and compelling. This impression fades, however, when we realize that the dominance of probabilistic credences depends delicately on the scoring rule or inaccuracy measure chosen. Most scoring rules do not return the dominance of probabilities. Even rules that differ minutely from the Brier score are enough to undo the dominance.

To see this, replace the power of 2 used in the Brier score with a different exponent n. That is, the generating functions for what I shall call the "n-power" scoring rule are now

$$g_1(x_i) = (1 - x_i)^n \quad \text{when } i = k$$
$$g_0(x_i) = x_i^n \quad \text{when } i \neq k \qquad (5a)$$

[157] Based on geometric intuitions, the tacit assumption above was that the set of points $\{< x_1+k, x_2+k, \ldots, x_r+k >\}$ is dominated by a single point. This assumption is now vindicated, since a single value of k produces a unique optimum for all loss functions. For completenss, the second derivative of all loss functions with respect to k is everywhere positive, so the optima computed are true minima.

where, as before, outcome E_k is the one that obtains.

For $n>0$, these will lead to what are, intuitively, accuracy measures. The function $g_1(x_i)$ is strictly decreasing, so it rewards a higher credence x_i in the result that obtains with a smaller loss. The function $g_0(x_i)$ is strictly increasing, so it punishes a higher credence in a result that does not obtain with a greater loss. The loss functions become

$$\begin{aligned} L_1 &= (1 - x_1)^n + x_2^n + x_3^n + \ldots + x_r^n \\ L_2 &= x_1^n + (1 - x_2)^n + x_3^n + \ldots + x_r^n \\ &\ldots \\ L_r &= x_1^n + x_2^n + x_3^n + \ldots + (1 - x_r)^n \end{aligned} \qquad (4a)$$

Among all values of $n>0$, the only value that supports the dominance of probabilistic credences is $n=2$. The slightest deviation from it undoes the dominance. Choosing different values of n allows us a generate results of considerable variety, as we shall now see.

5.1 Scoring Rules with $n>1$

We begin exploring the dominance relations by considering loss functions with $n>1$. They exhibit dominance relations qualitatively similar to those of the Brier score. Their curves of constant loss are concave towards the region of lower loss, so that dominating points in the space arise in the same way, qualitatively, as in the case of the Brier score. However the credences that dominate are not probabilistic. Loss functions with $1<n<2$ lead to superadditive credences. Loss functions with $n>2$ lead to subadditive credences.

To recall the definitions: if credences $x(A)$ and $x(B)$ for mutually exclusive outcomes A and B are subadditive, then the credence $x(A \vee B)$ elicited for their disjunction satisfies $x(A \vee B) < x(A) + x(B)$. If the credences are superadditive then we have for this last case that $x(A \vee B) > x(A) + x(B)$. In the analysis that follows, we will identify sub and super additive behavior in relation to the credence in the full outcome set to which credence 1 is assigned:

$$x_1 + x_2 + \ldots + x_r > 1 \qquad \text{(subadditive)}$$

$$x_1 + x_2 + \ldots + x_r < 1 \qquad \text{(superadditive)}$$

To see with least effort how these deviations from additivity arise, we calculate the dominating credence for the "diagonal" set of points:

$$\{< x_1, x_2, \ldots, x_r>: x_1 = x_2 = \ldots = x_r = x, 0 \le x \le 1\} \qquad (7)$$

This is just the diagonal that runs from the origin <0,0,…,0> to <1,1,…,1> of the r-dimensional hypercubic space. The dominating point in the set is identified once again by condition (6). In this set, each loss function is the same function of x:

$$L_1 = L_2 = \ldots = L_1 = L(x) = (1 - x)^n + (r\text{-}1)\, x^n$$

A short calculation that sets $dL/dx=0$ in accord with condition (6) shows that the minimum loss for all the loss functions occurs when[158]

$$x_{dom} = \frac{(1/r)^{1/(n-1)}}{(1/r)^{1/(n-1)} + (1-1/r)^{1/(n-1)}} = \frac{(1/r)^{1/(n-1)}}{(1/r)^{1/(n-1)} + (r-1)^{1/(n-1)}(1/r)^{1/(n-1)}} \tag{8}$$

That is, $< x_1, x_2, \ldots, x_r > = < x_{dom}, x_{dom}, \ldots, x_{dom} >$ dominates this diagonal set as the point of smallest loss.

To conform with the probability calculus, the r credences of this dominating point must be $x_{dom} = 1/r$, so that their sum for the r outcomes, (r x $1/r$), equals unity. This will happen only in two cases. First is the case of $r=2$, that is, of two outcomes only. Then $(r\text{-}1)^{1/(n-1)} = (1)^{1/(n-1)} = 1$ and we have, for all n, that

$$x_1 = x_2 = x_{dom} = 1/2$$

Second is the case of the Brier score, $n=2$. For then $1/(n\text{-}1) = 1$, so that $(r\text{-}1)^{1/(n-1)} = (r\text{-}1)$; and we have for the dominating point

$$x_1 = x_2 = \ldots = x_r = x_{dom} = 1/r$$

In all other cases, additivity fails.

For $r>2$ and $n>2$, the exponent in (8) satisfies $0 < 1/(n\text{-}1) < 1$ and we have

$$(r\text{-}1)^{1/(n-1)} < (r\text{-}1)$$

It follows from (8) that:

$$x_{dom} > \frac{(1/r)^{1/(n-1)}}{(1/r)^{1/(n-1)} + (r-1)(1/r)^{1/(n-1)}} = \frac{(1/r)^{1/(n-1)}}{r \cdot (1/r)^{1/(n-1)}} = \frac{1}{r}$$

This entails that the r credences x_{dom} sum to greater than unity (subadditivity):

$$x_1 + x_2 + \ldots + x_r = r x_{dom} > 1$$

For $r>2$ and $1<n<2$, the exponent in (8) satisfies $1/(n\text{-}1) > 1$ and we have

$$(r\text{-}1)^{1/(n-1)} > (r\text{-}1)$$

By analogous reasoning to the previous case, the r credences x_{dom} sum to less than unity (superadditivity):

$$x_1 + x_2 + \ldots + x_r = r x_{dom} < 1$$

The failure of additivity arises with the slightest deviation from the Brier score exponent 2. That is, the dominance argument fails to returns probabilities if the exponent is 2.01 or 1.99. In those cases, the deviations from additivity of the dominating credences will be small. The

[158] For $n>1$, the second derivative $d^2L/dx^2 > 0$, everywhere, so the turning point is a minimum.

deviations can be made as large as we please simply by selecting suitably large or small values of n.

For example, for r=28 and n=4, we find x_{dom} = 1/4. Then the credences sum to

$$x_1 + x_2 + \ldots + x_{28} = 28(1/4) = 7$$

If we set r=11 and n = 11/10, we find $x_{dom} \approx 10^{-10}$. Then the credences sum to

$$x_1 + x_2 + \ldots + x_{11} \approx 11\text{x}10^{-10}$$

A more general sense of the range of possibilities is provided by a plot in Figure 2 of the sum S = $r.x_{dom}$ against n, for various values of r>2. Additivity is respected just when S=1. This arises only when n=2. All the curves intersect at S=1, n=2.

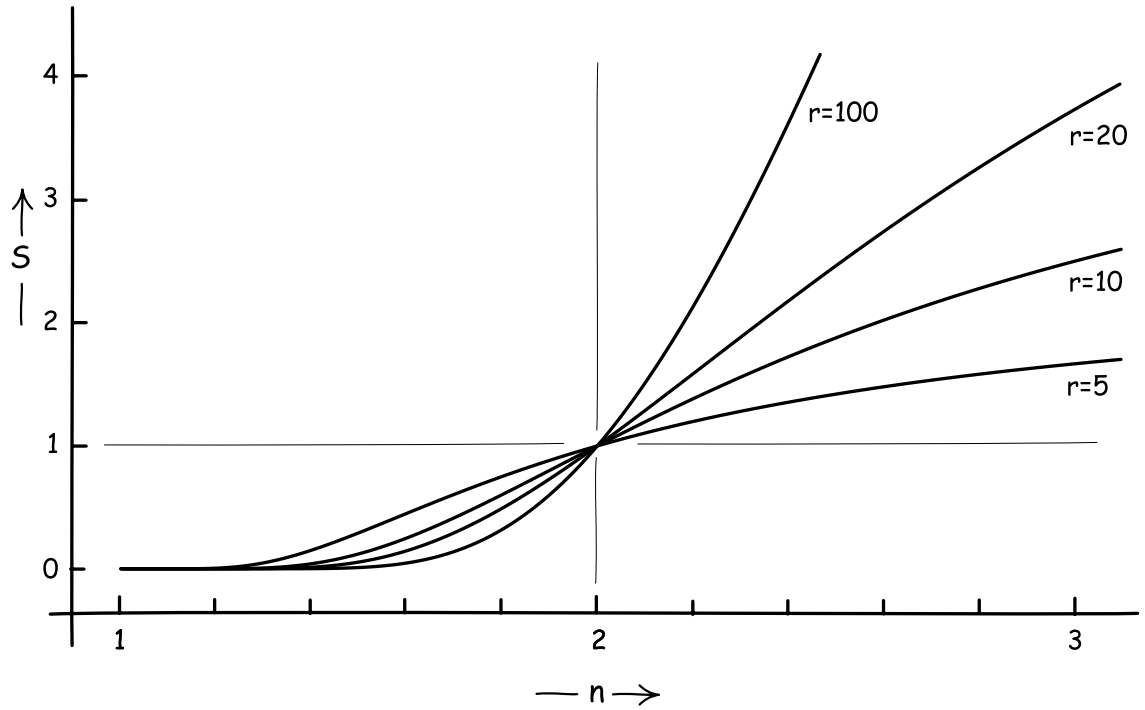


Figure 2. Failure of additivity for n-power scoring rules

These results are a special case of the general result demonstrated in Appendix A. That is, for n>1, the dominating points in the space of r credences $x_1, x_2, \ldots, x_r$ lie on an r-1 dimensional hypersurface in the space of credences, satisfying:

$$1 = \frac{x_1^{n-1}}{\left[x_1^{n-1} + (1-x_1)^{n-1}\right]} + \ldots + \frac{x_i^{n-1}}{\left[x_i^{n-1} + (1-x_i)^{n-1}\right]} + \ldots + \frac{x_r^{n-1}}{\left[x_r^{n-1} + (1-x_r)^{n-1}\right]} \qquad (9)$$

For r>2, this surface coincides with the surface of additive probabilities

$$1 = x_1 + x_2 + \ldots + x_r$$

only when n=2. Otherwise, for n>2, the surface lies above this additivity surface and the credences are subadditive. For n<2, the surface lies below this additivity surface and the credences are superadditive.[159]

5.2 Scoring Rules with 0<n<1

We now consider the case of loss functions (4a) with exponent n satisfying 0<n<1. This case exhibits behavior that is qualitatively different from the case of n>1. For now the surfaces of constant loss are convex towards the direction of smaller loss. That inclines credences to move to extreme values to secure smaller losses. This effect can be seen in the case of two outcomes, r=2, and a square root loss function, n=1/2. Then we have two loss functions:

$$L_1 = \sqrt{1 - x_1} + \sqrt{x_2}$$

$$L_2 = \sqrt{x_1} + \sqrt{1 - x_2}$$

Curves of constant loss are plotted in Figure 3. Those for loss L_1 are on the left; and those for loss L_2 are on the right. Probabilistic credences satisfying $x_1 + x_2 = 1$ lie on the dashed diagonal.

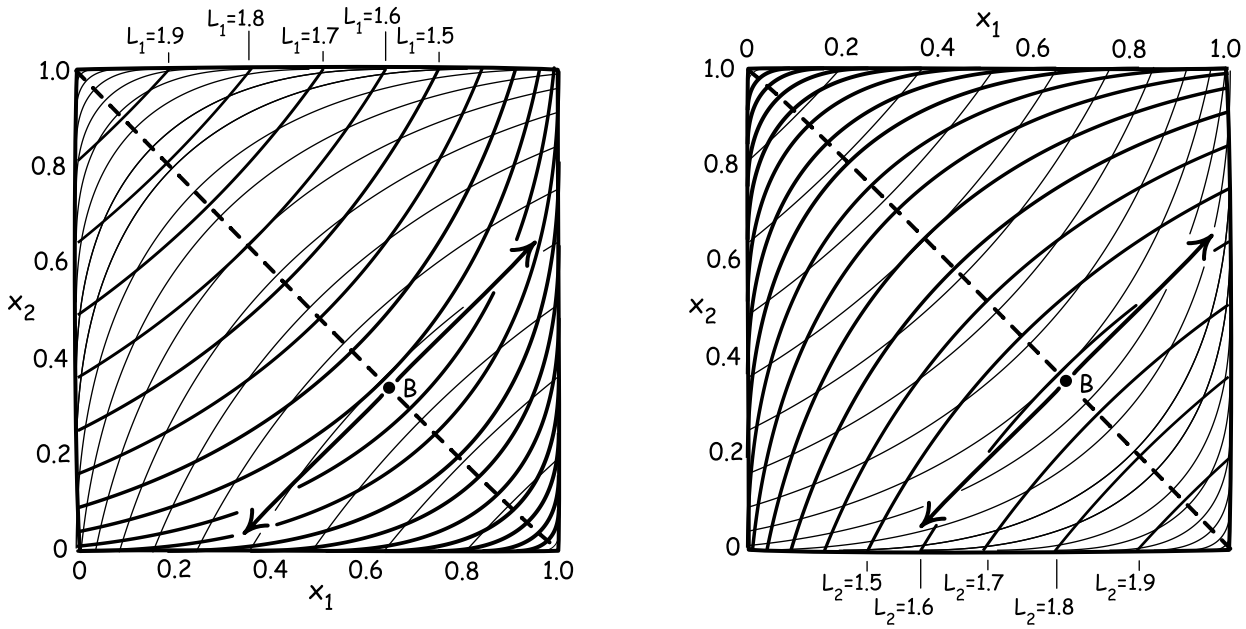


Figure 3. Dominance of extremes with n = 1/2

[159] Equation (8) picks out a point on this surface. It is recoved by substituting $x_1 = \ldots = x_r = x$ into (12) and solving for x.

Repeating the analysis of Figure 1, we find in this case that moving credences away from this diagonal decreases both loss functions L_1 and L_2 and thus increases accuracy. An arbitrarily chosen additive credence at B is dominated by non-additive credences to which we arrive by following the arrows towards the extremes. Most striking is that the additive credences at $x_1 = x_2 = 0.5$, are dominated by the credences $x_1 = x_2 = 0$; and $x_1 = x_2 = 1$.

This striking behavior of dominance of probabilistic credences by both subadditive and superadditive credences is an artifact of having just two outcomes, r=2. For the case of more that two outcomes, the dominating credences all have lower values and are superadditive. This is easy to see in the case of the diagonal set (7). All the loss functions for it are the same for the case of n=1/2:

$$L_1 = L_2 = \ldots = L_r = L(x) = \sqrt{1-x} + (r-1)\sqrt{x}$$

More generally, for all $0<n<1$, the loss functions are

$$L_1 = L_2 = \ldots = L_r = L(x) = (1-x)^n + (r-1)x^n$$

For all these cases, the loss functions has a dominating minimum at the origin only:

$$x_1 = x_2 = \ldots\ x_r = x = 0$$

where $L = 1$.[160] When $x_1 = x_2 = \ldots\ x_r = x = 1$, $L = r$-1, which is greater than one for r>2.

5.3 Scoring Rules with n=1

The final case uses the absolute norm. That is, the generating functions are now[161]

$$g_1(x_i) = (1 - x_i) \text{ when } i = k$$
$$g_0(x_i) = x_i \qquad \text{when } i \neq k \tag{5b}$$

where, as before, E_k is the outcome that obtains. In the case of two outcomes, this scoring rule exhibits qualitatively different behavior again. The two loss functions are

$$L_1 = (1\text{-}x_1) + x_2 = 1 - (x_1 - x_2)$$
$$L_2 = x_1 + (1\text{-}x_2) = 1 + (x_1 - x_2)$$

The curves of constant loss for both are the same

$$x_1 - x_2 = \text{constant}$$

---

[160] Write, $L(x,n) = (1\text{-}x)^n + (r\text{-}1)\ x^n$. We have $L(0,n) = 1$. Also $L(x,1) = 1+(r\text{-}2)x > 1$, for all x>0, r>2. But $L(x,n) > L(x,1)$, for all $0<n<1$ and x>0, since then $(1\text{-}x)^n > (1\text{-}x)$ and $x^n > x$.

[161] This case is often presented as the absolute norm, writing $g_1(x_i) = |1 - x_i|$. Since $0 \leq x_i \leq 1$, the absolute operator |.| is superfluous.

They differ only in the values assigned to the curves. Since $L_2 = 2 - L_1$, the curves differ in the direction of increasing loss. These curves are plotted in Figure 4, with curves of constant L_1 on the left; and curves of constant L_2 on the right.

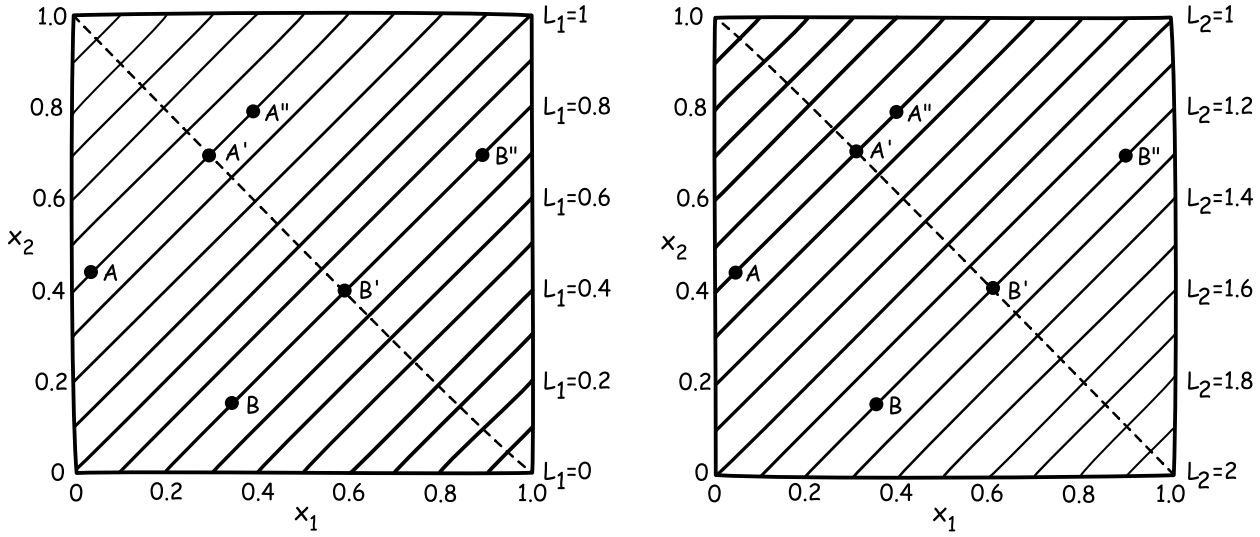


Figure 4. Degeneracy of dominance with n=1

In this degenerate case, dominance fails, since both loss functions are constant along the curves shown. Thus, as far as the accuracy measure is concerned, all the credences A, A', A'', … are equally accurate; and all the credences B, B', B'', … are equally accurate.

This degeneracy is not specific to the absolute norm n=1, but is recoverable in the case of two outcomes, $r = 2$. For example, take generating functions

$$g_1(x_i) = 1 - h(x_i) \quad \text{when } i = k$$
$$g_0(x_i) = h(x_i) \quad \text{when } i \neq k \tag{5d}$$

where, as before, outcome E_k is the one that obtains. Then, as above, curves of constant loss for both L_1 and L_2 are the same:

$$h(x_1) - h(x_2) = \text{constant}$$

Instead of a dominance relation, we find all credences on each of the curves to have the same loss L_1 and L_2 and thus to be equally accurate. We can take many increasing functions for $h(x)$, such as $h(x) = x^2$. For this case, these curves are hyperbolas, with an asymptote of $x_1 = x_2$.

This degeneracy of the absolute norm rule does not persist when we move to more than two outcomes, r>2. Then, smaller valued credences dominate. The loss functions are

$$L_1 = 1 - x_1 + x_2 + x_3 + \ldots + x_r$$
$$L_2 = x_1 + 1 - x_2 + x_3 + \ldots + x_r$$
$$\ldots$$
$$L_r = x_1 + x_2 + x_3 + \ldots + 1 - x_r$$

For the diagonal set of credences (7), all the loss functions are equal

$$L_1 = L_2 = \ldots = L_1 = L(x) = 1 + (r\text{-}2)x$$

The dominating credence is

$$x_1 = x_2 = \ldots = x_r = x = 0$$

More generally, uniformly reducing credences in such a way that we remain within the space $0< x_i < 1$ ($i = 1, \ldots, r$), uniformly decreases all the loss functions and thus increases accuracy. For example, we start at $\mathbf{x} = < x_1, x_2, \ldots, x_r>$ in this space and move to a new point:

$$\mathbf{x} - \boldsymbol{\varepsilon} = < x_1\text{-}\varepsilon, x_2\text{-}\varepsilon, \ldots, x_r\text{-}\varepsilon>$$

for some increment $\varepsilon>0$ sufficiently small to keep us in the space. Then we have for all $i = 1, \ldots r$,

$$L_i(\mathbf{x} - \boldsymbol{\varepsilon}) = L_i(\mathbf{x}) - (r\text{-}2)\varepsilon$$

Thus the credence $\mathbf{x}$ is dominated by the uniformly smaller credence $\mathbf{x} - \boldsymbol{\varepsilon}$. We can continue descending to smaller credences until we finally strike the origin $\mathbf{x} = \mathbf{0}$ or end up on one of the two dimensional edges of the hypercubic space (in which case the above degeneracy replaces the dominance relations).

6. Accuracy Gives Very Little

In sum, the above exploration shows that the accuracy dominance of probabilistic credences is fragile. It depends critically on choosing exactly the right scoring rule. The Brier score belongs to a larger family of power rule scores (4a) and (5a), characterized by the exponent n. The case of $n=2$ is the only case among them that returns the dominance of probabilistic credences. Other values of n give widely varying results. For $n>2$, the dominating credences are subadditive. For $1<n<2$, the dominating credences are superadditive. Scoring rules with $0<n\leq 1$, generally exhibit dominance by the lower values of credence in the space. Cases of equal credence, such as the probabilistic $x_i=1/r$, ($i = 1, \ldots, r$) are dominated by all zero credences $x_1 = x_2 = \ldots x_r = 0$, for example. We also saw anomalous cases of dominance by small and large credences and failures of dominance, in favor of equality of accuracy over some sets of credences.

If one is not antecedently committed to probabilistic credences, there is nothing especially troublesome in these results. We learn from them that a requirement of accuracy does not have univocal import. It must balance rewards for credence in the outcome that obtains with punishments for credences in those that do not. There are, it turns out, many ways to effect this balance. There is no obviously right way to do it.

Some rules, such as those with $n>1$, encourage prudence and direct credences towards intermediate values, while generally still not favoring probabilities. Others (such as $n=1/2$, $r=2$) effect the balance so that rashness is rewarded. All unit credences dominate in the equal credence case, since the reward for assigning unit credence to the outcome that obtains exceeds the punishment for assigning unit credence to the outcome that does not obtain. Still other rules encourage timidity. For them, assigning all zero credences is most accurate since the reward for a higher credence on the outcome that obtains is overwhelmed by the punishment for higher credences in outcomes that do not obtain.

These are widely varying results and we should accept them. To do otherwise and select among them for those we prefer, is simply to invalidate the whole accuracy-based method. We would not be using the method to inform our understanding and correct our prejudices. We would be using our prejudices to overturn what our method tells us.

7. Attempts to Justify the Choice of Scoring Rule

If one is antecedently committed to probabilistic credences, matters look very different. These results are troublesome. One has to find some way to impugn virtually all the accuracy measures employed in favor of the very few that return the desired result. In effect, one must work backwards from the probabilistic result desired to a condition that will deliver it. When the working backwards is done well, the resulting conditions will be congenial to those who already conceive credences as probabilities. To others, however, they will appear arbitrary.

Rosenkrantz (1981, 2.2) is an early attempt to justify the Brier score independently within the context of the dominance based vindication of probabilities. He noted that, when it is used for elicitation of credences, the Brier score has the property that a subject with non-probabilistic credences minimizes the loss by reporting credences that are *proportional* to the "true probabilities." This, he calls "absolutely non-distorting." Rosenkrantz conjectures but does not show that the Brier score is uniquely selected by this property, supplemented by other, weaker properties. The analysis seems hasty, since all strictly proper scoring rules (to be discussed below) share this property. Moreover the property does not seem praiseworthy, since it is just the result reported above in Section 3 that a Brier score elicitation rewards subjects for lying about

their non-probabilistic credences by rescaling them to probabilities with a constant multiplicative factor.

Joyce's (1998) proposal for restricting scoring rules is more definite and more confident. His "main theorem" (pp. 587-588) shows that probabilistic credences dominate if we use a scoring rule that satisfies six conditions that he names:

Structure, Extensionality, Normality, Dominance, Weak Convexity, and Symmetry

None of these conditions is a logical necessity. Each is merely natural for probabilists. Each introduces into the proof a contingent presupposition congenial to probabilists. As a result, each contributes to the circularity. Lest the analysis grow too lengthy, we consider only two of the strongest conditions, weak convexity and symmetry.

If two credences $\mathbf{c}$ and $\mathbf{c'}$ have the same score on some outcome, then Weak Convexity requires that the score assigned to their midpoint, $(\mathbf{c}+\mathbf{c'})/2$ is strictly less, unless $\mathbf{c} = \mathbf{c'}$. Considered abstractly, the requirement seems natural enough. "Weak Convexity is motivated by the intuition that extremism in the pursuit of accuracy is no virtue," Joyce (p. 596) assures us. However weak convexity is violated by power scoring rules with $0<n<1$. As we saw above in Section 6, that does not make them defective, but just different ways of balancing the rewards for true beliefs and punishments for false beliefs. To preclude them is not to learn from what accuracy measures tell us, but to tell accuracy measures what they should be doing to accord with our other notions. It is part of the artificial adjustment of the premises needed if the demonstration is to yield the predetermined result, the necessity of probabilities.

Weak convexity alone, however, does not restrict power scoring rules with $n>1$. The further restriction needed in the main theorem is "Symmetry." If two credences $\mathbf{c}$ and $\mathbf{c'}$ have the same score on some outcome i, then the distribution of scores over the intermediate credences is symmetric in the sense that, for any $0\leq\lambda\leq1$

$$L_i(\lambda\mathbf{c} + (1-\lambda)\mathbf{c'}) = L_i((1-\lambda)\mathbf{c} + \lambda\mathbf{c'})$$

This condition does pick out just the quadratic Brier score from all n-power scoring rules as required.[162] Thus, if we are working backwards to a predetermined result, the condition will seem apposite. However it is difficult to see any independent justification for it. Joyce's rationale

---

[162] An easy way to see this is to consider credences $(x_{dom}+\varepsilon)$ among the diagonal set (7) in the immediate vicinity of the dominating point x_{dom}, for $n>1$. The symmetry of scoring rule L_i will manifest in the vanishing of the cubic term in ε^3 in the power series expansion

$L_i(x_{dom} + \varepsilon) = L_i(x_{dom}) + \varepsilon\, L_i'(x_{dom}) + \varepsilon^2/2\, L_i''(x_{dom}) + \varepsilon^3/6\, L_i'''(x_{dom}) + \ldots$

However $L_i'''(x_{dom})=0$ only in the case of $n=2$.

(p. 597) merely restates what the formula says in words and suggests that Symmetry somehow precludes an improper favoring of one credence over another.

By the writing of his (2009), Joyce had presumably recognized the fragility of positing these conditions unequivocally. They were, he conceded, "not all well justified" (p. 264) and a reappraisal was undertaken. Indeed at times the commitment to the overall project is equivocal. The decline predicted earlier seems well underway. We are told (p. 266):

> Readers will be left to decide for themselves which of the properties discussed below conform to their intuitions about what makes a system of beliefs better or worse from the purely epistemic perspective.

A proof has scant foundations if acceptance of its premises depends on the intuitions of individual readers. My intuitions about angles and lines are immaterial to the proof of Pythagoras' theorem or the impossibility of duplicating the cube. In a notable compromise of the entire program of providing quantitative, normative guides to credences, we are informed that the idea that "epistemic goodness or badness for partial beliefs can be made sufficiently precise and determinate to admit of quantification" is merely a "useful fiction." We are told (p. 267) of a newly named condition "admissibility" that "is not a substantive claim about epistemic rationality" but is a way to "capture one's sense of what is valuable about beliefs from a purely epistemic perspective." Nonetheless it is used to restrict the choice of scoring rules, although apparently on rather infirm ground.

One should not fear that Joyce (2009) has abandoned the original project entirely. For eventually, Joyce settles on what is offered as the "least restrictive" of the theorems that employ dominance ideas to demonstrate the necessity of probabilities. The theorem, details of which are found in Joyce (2009, pp. 287-88), depends, among others, upon the condition of "Coherent Admissibility." (p. 280) This condition dismisses a scoring rule as "unreasonable" if it assigns a worse score to a probabilistic credence than to a non-probabilistic one in the case of all outcomes.

Leitgeb and Pettigrew (2010, p. 246) seem to me to give the correct appraisal. Coherent Admissibility is far from benign since…

> … it accords a privileged status to probability functions. We are inclined to ask: Why is it that we are justified in demanding that every probability function is admissible? Why are we not justified in demanding the same of a belief function that lies outside that class? And, of course, we must not make this demand of any nonprobability function;…

Just this sort of privileging of probabilities seems quite benign if one is working backwards from the predetermined conclusion that credences must be probabilities, for the condition says that a scoring rule cannot preclude probabilities, as Joyce says, "a priori" (p. 280). It does not appear benign to those who have not already prejudged the outcome.

A real difficulty for probabilists is that once one becomes convinced that credences have to be probabilities, it is hard to conceive how alternatives could be cogent. This may be behind Joyce's (2009, p, 283) concerns that the all-zero valued credences that can dominate with power scoring rules when $0<n\leq 1$. His assessment is severe. He calls them "logically inconsistent," since:

> The believer minimizes expected inaccuracy by being absolutely certain that every [proposition] is false even though logic dictates that one of them must be true.

This accusation of logical inconsistency will be unwelcome to proponents of the Shafer-Dempster theory of belief functions. Complete ignorance is represented there by assigning zero valued belief functions Bel to all outcome sets excepting the universal set. We see here that Joyce's assessments are driven by a prior commitment to interpret credences as probabilities, so that zero credence coincides with certain falsity. [163] In the Shafer-Dempster theory, a zero belief function can be interpreted as demarcating an interval of belief stretching from zero to one.

In my view, the most promising avenue for restriction of scoring rules is through the class of "strictly proper" scoring rules that are much used elsewhere. Joyce (2009, §8) discusses and defends them. Let us first review them.

8. Strictly Proper Scoring Rules

This class of scoring rules arose in a different context, that of scoring a predictor's performance and of the elicitation of subjective probabilities. It addresses the problem that most alternatives to the Brier rule do not deliver probabilistic credences at their minima.

For example, we can generalize the Brier rule by replacing its exponent 2 by an arbitrarily selected n, as in the n-power rule of (5a) above. It is shown in Appendix B below that the only value of n that gives a rule that correctly elicits probabilities is $n=2$. For all $n>2$ (and $n>2$), the power rule (5a) elicits subadditive credences. Alternatively, if $1<n<2$, then the n-power rule elicits superadditive credences.

These general n-power rule elicitations have an awkward property something like the reverse of the $n=2$ Brier rule. We saw above in Section 3 that the Brier rule elicits an additive probability measure, even when the subject's true credences are not probabilistic. The n-power

[163] Of course, even for probabilists, zero probability does not coincide with certain falsity, but merely measure zero improbability. De Finetti's finitely additive treatment of the infinite lottery assigns zero probability to each outcome individually. That a dart strikes any particular point on the board is a probability zero outcome, although one must happen.

rule (for n not 2) elicits credences that are not probabilities, even when the subject's true credences are probabilities.

The upshot is that the formal properties of the credences elicited by the scoring rule method will only be probabilities if the rule used is very carefully tuned to give just that result. The standard response in the literature on elicitation and on assessment of a predictor's performance is to restrict the scoring rules under consideration to "strictly proper" scoring rules.

As background to the notion, we recall that a general scoring rule employs two functions: $g_1(x)$ to reward a credence x in what turns out to be the true outcome; and $g_0(x)$ to punish a credence x in an outcome that turns out not to be true. The loss score assigned to elicited credences $\mathbf{x} = \langle x_1, x_2, \ldots, x_r \rangle$ for true probabilistic credences or true frequencies $\mathbf{p} = \langle p_1, p_2, \ldots, p_r \rangle$ is

$$\begin{aligned} L(\mathbf{p}, \mathbf{x}) = \; & p_1\, g_1(x_1) + \ldots + p_1\, g_0(x_i) + \ldots + p_1\, g_0(x_r) \\ & + \ldots \\ & + p_i\, g_0(x_1) + \ldots + p_i\, g_1(x_i) + \ldots + p_i\, g_0(x_r) \\ & + \ldots \\ & + p_r\, g_0(x_1) + \ldots + p_r\, g_0(x_i) + \ldots + p_r\, g_1(x_r) \end{aligned} \qquad (10a)$$

The most direct definition (such as given in Gneiting and Raftery, 2007, p. 359) simply asserts that:

Strictly Proper I

A scoring rule L is strictly proper just if $L(\mathbf{p}, \mathbf{x}) \geq L(\mathbf{p}, \mathbf{p})$, for all p_i in $0 \leq p_i \leq 1$, $i = 1, \ldots, r$, with equality only when $\mathbf{x} = \mathbf{p}$.

This definition explicitly rules out by fiat any scoring rule that fails to elicit $\mathbf{x}$ as a probability measure. Note that the definition is so strong that, like the Brier rule, a strictly proper scoring rule will elicit a probability even when subject's true credences are not probabilities. To see this, imagine that the subject's true credences are a non-probabilistic $\mathbf{q} = (q_1, q_2, \ldots, q_r)$. We can normalize them to a probability

$$\mathbf{p} = \langle p_1, p_2, \ldots, p_r \rangle = \mathbf{q}/Q = \langle q_1/Q, q_2/Q, \ldots, q_r/Q \rangle$$

by dividing by $Q = (q_1 + q_2 + \ldots + q_r)$. If the subject's true probability is $\mathbf{p}$, we know that the scoring rule will elicit $\mathbf{x} = \mathbf{p}$. By the definition of strictly proper scoring rules, $\mathbf{x} = \mathbf{p}$ is the unique value of $\mathbf{x}$ that minimizes $L(\mathbf{p}, \mathbf{x})$. However, $L(\mathbf{p}, \mathbf{x})$ is linear in $\mathbf{p}$, so that $L(\mathbf{p}, \mathbf{x}) = L(\mathbf{q}, \mathbf{x})/Q$. Hence $\mathbf{x} = \mathbf{p}$ will also minimize $L(\mathbf{q}, \mathbf{x})$ uniquely. That is, if the subject's true credences are a non-probabilistic $\mathbf{q}$, a strictly proper scoring rule will reward the subject most if the subject lies and reports a probabilistic, normalized credence $\mathbf{p} = \mathbf{q}/Q$.

9. Strictly Proper Scoring Rules in the Dominance Argument

This favoring of probabilities by strictly proper scoring rules is unproblematic in the context in which the notion was introduced. For when they are used to elicit probabilities from a subject, we begin with the assumption that the subject's credences are already probabilities. Correspondingly, when we use the rule to assess the performance of a predictor against the actual frequencies of outcomes, these actual frequencies are also additive measures.

The use of strictly proper scoring rules ceases to be benign, however, when they are used as part of a vindication of probabilities. For strictly proper scoring rules are engineered to favor probabilities and will yield them even then they are not the subject's credences. They exhibit the same favoring of probabilities if they are used as accuracy measures in the dominance arguments used to vindicate probabilities. A much-noted theorem in the scoring rule literature (see, for example, Predd et al., 2009, p. 4788) asserts exactly this: any non-probabilistic credence **q** is strongly dominated by a probabilistic credence **p**, where "strongly dominated" means that **p** has a strictly lower score than **q** for all possible outcomes, when the scoring rule used is strictly proper.

A simpler but less transparent definition of a strictly proper scoring rules lets us display the dominance in an example.

> *Strictly Proper II*[164]
>
> A scoring rule L is strictly proper just if $pg_1(x) + (1\text{-}p)g_0(x)$ is uniquely minimized at $x=p$ for all $0 \leq p \leq 1$.

This definition is equivalent to the definition *Strictly Proper I.* (For a demonstration of the equivalence, see Appendix D.)

This simpler form of the definition lets us see quickly how probabilistic credences dominate in a special case, that of the "diagonal" set (7) of credences above. For the general scoring rule, the generalization of the r loss functions (4) and (4a) above is:

$$\begin{aligned} L_1 &= g_1(x_1) + g_0(x_2) + g_0(x_3) + \ldots + g_0(x_r) \\ L_2 &= g_0(x_1) + g_1(x_2) + g_0(x_3) + \ldots + g_0(x_r) \\ &\ldots \\ L_r &= g_0(x_1) + g_0(x_2) + g_0(x_3) + \ldots + g_1(x_r) \end{aligned} \tag{4a}$$

For the diagonal set (7) of credences, all these loss functions reduce to the same expression:

$$L = L_1 = L_2 = \ldots = L_r = g_1(x) + (r\text{-}1)\, g_0(x) = r\,.\,[\,(1/r)\, g_1(x) + (1\text{-}1/r)\, g_0(x)\,]$$

---

164 Predd et al (2009, p. 4787) also include the requirement that the functions $g_0(x)$ and $g_1(x)$ are continuous. Schervish, Seidenfeld and Kadane (2009, p. 205) relax the condition of continuity. Some of my analysis assumes differentiability of these functions, however.

The second definition of strict propriety tells us directly that all these loss functions are uniquely minimized when

$$x = x_1 = x_2 = \ldots = x_r = 1/r$$

That is, all credences in the set are strongly dominated by this probabilistic credence.

The selection of a strictly proper scoring rule in the accuracy driven vindication of probability amounts to a delicate fine-tuning of the analysis to give just the probabilistic result antecedently desired. The extent of the fine-tuning depends on just how sparsely strictly proper scoring rules are distributed among scoring rules that we would intuitively judge to be admissible measures of accuracy.

In short, the strictly proper rules are very sparsely distributed among this larger class of rules. This is already suggested by theorems such as in Schervish (1989) that show how all strictly proper scoring rules can be generated from selection of a small class of functions. We can more directly gauge the sparseness by means of the second definition above. In brief, we have considerable freedom in selecting either of the functions $g_0(x)$ or $g_1(x)$. But once one is fixed, then so is the other; and we can generate arbitrarily many scoring rule that are not strictly proper simply by selecting different functions for the second.

To see this, assume that $g_0(x)$ is fixed at some function suitable for penalizing a credence x on an outcome that does not obtain. We have from the second definition that $pg_1(x) + (1\text{-}p)g_0(x)$ has a unique minimum, for fixed p, when $x=p$. This minimum arises when the derivative with respect to x vanishes

$$p\frac{dg_1(x)}{dx} + (1-p)\frac{dg_0(x)}{dx} = 0$$

Substituting $x=p$ at this minimum, we have

$$x\frac{dg_1(x)}{dx} + (1-x)\frac{dg_0(x)}{dx} = 0$$

Since p can have any value in $0 \le p \le 1$, this relation is a restriction on the functions $g_0(x)$ and $g_1(x)$ for any x in the same range. It follows that

$$g_1(x) - g_1(0) = -\int_0^x \left(\frac{1-y}{y}\right)\frac{dg_o(y)}{dy}dy \qquad (11)$$

Reading from right to left in this formula, fixing $g_0(x)$ fixes $g_1(x)$ up to the additive constant $g_1(0)$. Selecting any other function for $g_1(x)$ will yield a scoring rule that is not strictly proper. For example, if we fix $g_0(x) = x^n$ for $n>1$, then a short calculation shows that $g_1(x)$ must be

$$g_1(x) = x^n - \left(\tfrac{n}{n-1}\right)x + 1$$

up to the additive constant $g_1(0)=1$. Any other choice of function for $g_1(x)$, such as the apparently "natural" n-power rule (5a), fails to be strictly proper.

10. Justifying Strict Propriety

A dominance-accuracy argument for probabilities that employs strictly proper scoring rules must provide independent grounds for the restriction to strictly proper scoring rules. That these rules are popular in the broader elicitation literature provides no such grounds. Indeed, it is quite the reverse. Since strictly proper scoring rules have been designed explicitly to favor probabilities, using them to preclude non-probabilistic credences is prima facie circular. Their favoring is so strong that, used as a means of elicitation, they will reward a subject with non-probabilistic credences who lies and declares probabilistic credences.

All that can now prevent the analysis collapsing into circularity is some independent justification of the use of strictly proper scoring rules. Joyce (2009, pp. 277-79) attempts such a justification by means of the notion of "immodesty." The quantity $L(\mathbf{p}, \mathbf{x})$ of (10a) is the probabilistically expected score using rule L of a credence $\mathbf{x}$, according to the expectations of probabilistic credence $\mathbf{p}$. A "modest" credence will judge $L(\mathbf{p}, \mathbf{x}) < L(\mathbf{p}, \mathbf{p})$. That is, it will judge some other credence $\mathbf{x}$ to have a lower expected score and thus to be more accurate than $\mathbf{p}$ itself. This is a poor situation for credence $\mathbf{p}$, since considerations of expected accuracy indicate that, by $\mathbf{p}$'s own assessment, credence $\mathbf{x}$ is the better one. The credences we should seek are, therefore, "immodest." They are such that they are, by their own lights, the most accurate.

This favoring of immodest credences is, in effect, a guide for selecting scoring rules, for a credence can only be immodest or modest in relation to a scoring rule. This guide leads us directly to strictly proper scoring rules. We are asking for rules in which $L(\mathbf{p},\mathbf{p})$ takes the minimum value in comparison with all other $L(\mathbf{p}, \mathbf{x})$. But just this property of a scoring rule is strict propriety, in form of definition I of Section 8 above.

The justification of a restriction just to strictly proper scoring rules is still not complete. For nothing so far precludes another scoring rule that might render some non-probabilistic credence immodest. The analysis stalls at this point since we have no precise characterization of this last sort of scoring rule. Note that the score $L(\mathbf{p}, \mathbf{x})$ of a strictly proper rule is the expected score for credence $\mathbf{x}$ according to probability $\mathbf{p}$. If we seek an immodest, non-probabilistic credence $\mathbf{y}$, then we would replace $\mathbf{p}$ in the score by $\mathbf{y}$. But then $L(\mathbf{y}, \mathbf{x})$ is no longer an

expectation. It is unclear how the quantity should be interpreted.[165] We have no clear way to characterize an immodest, non-probabilistic credence.

The regress of reasons must continue. In an attempt to complete the justification, Joyce considers cases of physical chances in which we naturally choose probabilistic credences. What credence can we have in the each of the six outcomes of a fair die throw, other than a probability of 1/6? Thus we should demand the hospitality condition of "Minimal Coherence" of our scoring rules: they should not preclude in advance probabilistic credences. That way credences concerning physical chance can be accommodated. If, however, we require both immodesty and the possibility of rules that favor probabilistic credences in their expectations, then we are led to strictly proper scoring rules. They are, by their definition, the only rule that can serve.

As we have seen so often before, this latest step in the regress of reasons will seem quite compelling to someone who antecedently favors probabilities. It is surely benign, they might think, to demand that we use scoring rules that are minimally hospital to probabilities in the sense that they do not automatically preclude them. To someone who has not prejudged the outcome, the demand is anything but benign.[166] For the burden of the analysis shows that this demand is enough to force probabilistic credences in all cases.

If our earnest desire is not to prejudge, then should we not ask that our scoring rules be hospitable to more than just probabilistic credences? What we seem to learning is a troubling dogmatism in the whole approach of scoring rules. Once we demand hospitality for one favored type of credence, no others are sustainable. It seemed benign merely to demand a place in the lifeboat for the first class passengers. But now we see that this benign demand fills the boat and all the other passengers must perish.

If this last vindication is unsatisfactory, might we find another? Pettigrew (2016, Ch.4) offers another vindication of strictly proper scoring rules. The analysis depends upon positing several conditions on an inaccuracy measure that include what he calls:

Divergence Additivity, Divergence Continuity and Decomposition

---

[165] For example, expectation-like quantities computed using a non-probabilistic $\mathbf{y}$ fail to meet minimal conditions of an expectation. For example, the expectation for a quantity $\mathbf{Q} = \langle Q_1, Q_2, \ldots, Q_r \rangle$ in the special case in which $Q_1 = Q_2 = \ldots = Q_r = Q$, should be Q. However the sum $\sum_i y_i Q_i = \sum_i y_i Q$ is equal to Q only when $\sum_i y_i = 1$, which is the case of probabilistic credence $\mathbf{y}$.

[166] Let us set aside the quibble that considerations of strict dominance in accuracy have been replaced by considerations of expected accuracy. That weakens the whole argument since maximizing expectations is not automatically always the best.

We find once again that these conditions are congenial for a probabilist who knows that they will yield the required result. They appear arbitrary, however, to someone not antecedently committed to probabilities.

Divergence Additivity requires that the inaccuracy of some set of credences $\langle x_1, x_2, \ldots, x_r \rangle$ is measured by taking the arithmetic sum of the inaccuracies of the individual credences, using $g_1(x_i)$ or $g_0(x_i)$, according to whether the credence x_i is in the true state or not. Summation seems, initially, to be an innocent requirement. Pettigrew (p. 49) calls it "the natural thing to do." However it is far from innocent. For it represents a particular rule for determining the import of variation among the individual inaccuracy measures. Take the case of five credences, r=5, and assume that we have two different sets of inaccuracies provided by the functions $g_1(x_i)$ or $g_0(x_i)$:

$$0.1, 0.1, 0.1, 01, 0.1 \text{ and } 0.01, 0.01, 0.01, 0.01, 0.46$$

How are we to summarize the combined inaccuracy in each case? Is the combined inaccuracy of the first the same as the second? Or does the presence of the large inaccuracy 0.46 in the second render the second case more inaccurate than the first? Or is this second case less inaccurate since four of its five components are very small, 0.01? Divergence Additivity measures the combined inaccuracy by summing the components. Since the components in each of the two cases sum to 0.5, this condition judges them equal in combined inaccuracy. That is a quite specific way to trade off the import of non-uniformities of the second case. Since it competes with many other possible ways of trading of non-uniformities, merely finding it "natural" falls well short of the independent justification needed.

Similar arbitrariness troubles the other two conditions. Briefly, Divergence Continuity requires the analogs of the functions $g_1(x)$ or $g_0(x)$ to be continuous in x. In the abstract, the requirement seems innocent. However requirements of continuity can be far from innocent. In geometry, we might think it innocent to require that some two-dimensional surface can be covered continuously by the familiar $\langle x,y \rangle$ coordinate system. However that condition restricts us to surfaces that are topologically "R^2". It precludes the surfaces of a sphere or a torus, even though both surfaces are, in a geometric sense, everywhere continuous. Finally, Decomposition arises from two further conditions, Calibration and Truth-Directedness, each of which, independently, looks quite natural. The difficulty is that these two conditions turn out to be incompatible, so that at least one is wrong. Once again naturalness proves to be a poor guide. Decomposition is a compromise condition that attempts to mediate between them. We may well wonder why it is a good idea to mediate between two conditions, one or both of which might be wrong. The mediation uses a formula that in turn appears arbitrarily chosen, unless one knows that it will enable to demonstration of the result sought.

All these efforts end up offering no escape from the problem that has dogged the accuracy-based vindication of probabilities from the start. We are trapped in an endless regress of reasons. The requirement of accuracy alone, it turns out, gives us very little. What really determines the outcome is our choice of scoring rule. Merely among *n*-power scoring rules, we can select any desired extent of super or subadditivity of our credences just by choosing a suitable *n*. If we are to vindicate a restriction to probabilistic credences, we must find further reasons that favor them. We find new reasons that seem natural; and then we realize that they are only natural if judged by our antecedent prejudice for probabilistic credences. Still further reasons are needed and the regress of reasons proceeds.

11. Naturalness Gone Astray

Selten (1998) provides a sobering illustration of the precariousness of accepting conditions on the basis of their naturalness. His interest is what he calls "the quadratic scoring rule." It is used in something like an elicitation context in which a predicted probability distribution p is scored against a true probability distribution x by means of the "expected score loss." His quadratic scoring rule is given in one form (p. 48) as

$$L(\mathbf{p} \mid \mathbf{x}) = \sum_{i=1}^{r} (x_i - p_i)^2$$

where the two distributions $\mathbf{x} = \langle x_1, \ldots, x_r \rangle$ and $\mathbf{p} = \langle p_1, \ldots, p_r \rangle$ adopt the indexed values x_i and p_i over outcomes $1, \ldots, r$. Selten (p. 43) reports: "As far as the author knows, Brier (1950) was the first one who described this rule." The principal result of the paper is a demonstration that its four axioms are satisfied uniquely by the quadratic scoring rule.

This uniqueness is a strong result, so Selten goes to some pains to justify the naturalness of what might be the most contentious of the axioms, the fourth axiom, "neutrality." It requires that the loss function L be symmetric in the two distributions:

$$L(\mathbf{p}|\mathbf{x}) = L(\mathbf{x}|\mathbf{p})$$

Selten's (p. 54) plea for the axiom is strong and plausible:

> The interpretation of axiom 4 becomes clear if one looks at the hypothetical case that one and only one of two theories p and q is right, but it is not known which one. The expected score loss of the wrong theory is a measure of how far it is from the truth. It is only fair to require that this measure is "neutral" in the sense that it treats both theories equally. If p is wrong and q is right, then p should be considered to be as far from the truth as q in the opposite case that q is wrong and p is right.

> A scoring rule should not be prejudiced in favor of one of both theories in the contest between p and q. The severity of the deviation between them should not be judged differently depending on which of them is true or false.
>
> A scoring rule which is not neutral is discriminating on the basis of the location of the theories in the space of all probability distributions over the alternatives. Theories in some parts of this space are treated more favorably than those in some other parts without any justification. Therefore, the neutrality axiom 4 is a natural requirement to be imposed on a reasonable scoring rule.

It is easy to accept this plea and, with it, neutrality as a reasonable demand for any scoring rule. The comfort will surely evaporate quite rapidly when one realizes that Selten's naturalness requirements establish the uniqueness of a scoring rule (his "quadratic" rule above) that differs from Brier's score (2a). Indeed Selten's formula is incompatible with the general scheme (10a) of strictly proper scoring rules now widely employed in the scoring rule literature.[167] It precludes all strictly proper scoring rule.

12. Conclusion

What makes the circularity of this accuracy based approach harder to see at the outset is that it draws on a well-established literature on scoring rules in meteorology, economics and subjective Bayesianism. That literature developed the scoring rules for other purposes. They were used to reward meteorologists for their probabilistic predictions, when scored against the actual frequencies of weather conditions; or they were used to encourage subjects to match their publicly declared probabilities with their true but hidden probabilities. For these purposes, it was appropriate to work with a narrow subset of scoring rules, adapted antecedently to probability measures. Using different rules, ill-adapted to probabilities would have no point.

Matters change when we try to use scoring rules to demonstrate the necessity of probabilities. Now the careful selection of these same scoring rules ceases to be the practical adaption of the rules to the intended use. It amounts to the covert assumption of the very thing that is to be proven. For these favored rules—the Brier score and its generalization as strictly proper scoring rules—strongly favor probabilistic credences. As we saw above, if a subject harbors non-probabilistic credences and these scoring rules are used to elicit them, the subject will be rewarded for lying and reporting probabilistic credences.

---

[167] To see this, note that (10a) is linear in the probability measure p_i, whereas Selten's measure is quadratic in it.

All would be well with accuracy based vindications if solid, independent grounds could be found for use of these favored rules. However, no such grounds have emerged and, I argue, none can emerge. For all such grounds must covertly assume exactly what they seek to demonstrate. Instead, inevitably and as we have seen repeatedly in the present literature, the latest grounds will succumb under scrutiny. We are forever trapped in an endless regress of reasons.

Appendices

Appendix A. Dominance Relations for *n*-Power Scoring Rule with *n*>1

The *n*-power loss functions

$$\begin{aligned} L_1 &= (1 - x_1)^n + x_2^n + x_3^n + \ldots + x_r^n \\ L_2 &= x_1^n + (1 - x_2)^n + x_3^n + \ldots + x_r^n \\ &\ldots \\ L_r &= x_1^n + x_2^n + x_3^n + \ldots + (1 - x_r)^n \end{aligned} \tag{4a}$$

admit dominating points that lie on an *r*-1 dimensional hypersurface of the *r* dimensional space of credences, $x_1, x_2, \ldots, x_r$. Each point on the surface is a minimum for all *r* loss functions among a set of points lying on a curve in the space of credences. We write this curve as $x_i(\lambda)$, $i=1, \ldots, r$, where λ is a path parameter. A dominance point is identified by means of the derivatives of the loss functions with respect to λ. The first derivatives are:

$$\frac{dL_1}{d\lambda} = -n(1-x_1)^{n-1}\frac{dx_1(\lambda)}{d\lambda} + nx_2^{n-1}\frac{dx_2(\lambda)}{d\lambda} + \ldots + nx_r^{n-1}\frac{dx_r(\lambda)}{d\lambda} \tag{13}$$

and similarly for $L_2, \ldots, L_r$. The second derivatives are

$$\begin{aligned} \frac{d^2L_1}{d\lambda^2} &= n(n-1)(1-x_1)^{n-2}\frac{dx_1(\lambda)}{d\lambda} - n(1-x_1)^{n-1}\frac{d^2x_1(\lambda)}{d\lambda^2} \\ &+ n(n-1)x_2^{n-2}\frac{dx_2(\lambda)}{d\lambda} + nx_2^{n-1}\frac{d^2x_2(\lambda)}{d\lambda^2} + \ldots \\ &\ldots + n(n-1)x_r^{n-r}\frac{dx_r(\lambda)}{d\lambda} + nx_r^{n-1}\frac{d^2x_r(\lambda)}{d\lambda^2} \end{aligned} \tag{14}$$

and similarly for $L_2, \ldots, L_r$. To identify a dominance point, we set all the first derivatives (13) to zero. The results for $dL_1/d\lambda = 0$ and $dL_i/d\lambda = 0$ are, respectively,

$$-(1-x_1)^{n-1}\frac{dx_1}{d\lambda}+\ldots+x_i^{n-1}\frac{dx_i}{d\lambda}+\ldots+x_r^{n-1}\frac{dx_r}{d\lambda}=0 \tag{15}$$

$$x_1^{n-1}\frac{dx_1}{d\lambda}+\ldots-(1-x_i)^{n-1}\frac{dx_i}{d\lambda}+\ldots+x_r^{n-1}\frac{dx_r}{d\lambda}=0$$

Subtracting the second from the first, we recover

$$\frac{dx_i/d\lambda}{dx_1/d\lambda}=\frac{\left[x_1^{n-1}+(1-x_1)^{n-1}\right]}{\left[x_i^{n-1}+(1-x_i)^{n-1}\right]} \tag{16}$$

This expression (16), with i=2, 3, … , r, can be used to replace expressions for $dx_2/d\lambda$, $dx_3/d\lambda$, … , $dx_I/d\lambda$ in (15), rewritten as:

$$(1-x_1)^{n-1}=x_2^{n-1}\frac{dx_2/d\lambda}{dx_1/d\lambda}+\ldots+x_i^{n-1}\frac{dx_i/d\lambda}{dx_1/d\lambda}+\ldots+x_r^{n-1}\frac{dx_r/d\lambda}{dx_1/d\lambda}$$

After some manipulation, the reconfigured equation (15) reduces to the expression that identifies the r-1 dimensional hypersurface of dominance points:

$$1=\frac{x_1^{n-1}}{\left[x_1^{n-1}+(1-x_1)^{n-1}\right]}+\ldots+\frac{x_i^{n-1}}{\left[x_i^{n-1}+(1-x_i)^{n-1}\right]}+\ldots+\frac{x_r^{n-1}}{\left[x_r^{n-1}+(1-x_r)^{n-1}\right]} \tag{12}$$

In the special case of n=2, the Brier score, this relation identifies the hypersurface of additive credences that conform with the probability calculus:[168]

$$1 = x_1 + \ldots + x_i + \ldots + x_r$$

To determine the disposition of the hypersurfaces of the remaining cases, we write the individual terms of (12) as

$$y_i=\frac{x_i^{n-1}}{\left[x_i^{n-1}+(1-x_i)^{n-1}\right]}$$

They can be inverted to yield

$$x_i=\frac{y_i^{1/(n-1)}}{\left[y_i^{1/(n-1)}+(1-y_i)^{1/(n-1)}\right]} \tag{17}$$

where, following (12), we have

$$1 = y_1 + \ldots + y_i + \ldots + y_r$$

A special case is r=2, for any n>1. For then $y_2 = (1- y_1)$ we have

$$x_1=\frac{y_1^{1/(n-1)}}{\left[y_1^{1/(n-1)}+(1-y_1)^{1/(n-1)}\right]}=\frac{y_1^{1/(n-1)}}{\left[y_1^{1/(n-1)}+y_2^{1/(n-1)}\right]}$$

[168] For this case, n-1=1 and $x_I^{n-1} + (1-x_i)^{n-1} = x_i + (1 - x_i) = 1$.

$$x_2 = \frac{y_2^{1/(n-1)}}{\left[y_2^{1/(n-1)} + (1-y_2)^{1/(n-1)} \right]} = \frac{y_2^{1/(n-1)}}{\left[y_2^{1/(n-1)} + y_1^{1/(n-1)} \right]}$$

so that the dominance points are also additive: $1 = x_1 + x_2$.

Otherwise, for $r>2$ and $n>2$, we have from (17) that

$$x_1 > \frac{y_1^{1/(n-1)}}{y_1^{1/(n-1)} + y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)}}$$

since

$$(1-y_1)^{1/(n-1)} = (y_2 + \ldots + y_r)^{1/(n-1)} < y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)} \qquad (18)$$

by means of inequality (23) below. Using similar relations for $x_2, x_3, \ldots, x_r$, we recover

$$x_1 + x_2 + \ldots + x_r > \frac{y_1^{1/(n-1)} + y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)}}{y_1^{1/(n-1)} + y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)}} = 1$$

It follows that $r>2$ and $n>2$ is the case of subadditive credences. Repeating the above analysis for $r>2$ and $1<n<2$, using inequality (24), we recover:

$$x_1 + x_2 + \ldots + x_r < 1$$

from which it follows that this is the case of superadditive credences.

The hypersurface (12) is picked out by the vanishing of the first derivatives, $dL_1/d\lambda = dL_2/d\lambda = \ldots = dL_r/d\lambda = 0$ for the curves $x_i(\lambda)$, $i=1, \ldots, r$. To complete the analysis, we need to show that these points are true minima for the loss functions along the curves, so that the points on the hypersurface are dominance points. This in turn requires identification of the curves.

It will be sufficient to identify one set of curves as follows.[169] In brief, we find the slope of the curve at each point on the hypersurface. We then take as the curve $x_i(\lambda)$ through that point, the straight line that has this slope as its slope everywhere. Select some point on the hypersurface, whose credences X_i satisfy equation (12). We have from (16) that

$$\frac{dx_i}{d\lambda} = \frac{K}{\left[X_i^{n-1} + (1-X_i)^{n-1} \right]}$$

where K is some undetermined constant that is the same for all $x_i(\lambda)$. The constant is undetermined since its differing values give us the freedom to rescale the parameter λ arbitrarily. We can, for example, alter the value of K if we introduce a new parameterization $\lambda'(\lambda)$ for which

[169] The properties described above do not, I suspect, uniquely define the curves $x_i(\lambda)$. Identifying one set of curves is sufficient to display the dominance properties of the points of the hypersurface.

$$\frac{dx_i}{d\lambda'} = \frac{dx_i}{d\lambda} \cdot \frac{d\lambda}{d\lambda'}$$

To ensure that the path parameterization introduces no nuisance pathologies, it is convenient to set it, by stipulation, proportional to the natural Euclidean path length through

$$d\lambda^2 = \text{constant} \cdot (dx_1^2 + dx_2^2 + \ldots + dx_r^2)$$

We select the constant in this expression so that the undetermined constant K is set to one. That is we now have

$$\frac{dx_i}{d\lambda} = \frac{1}{\left[X_i^{n-1} + (1 - X_i)^{n-1} \right]} = m_i(X_1, \ldots, X_r) > 0 \tag{19}$$

where $m_I > 0$ since $0 \leq X_I \leq 1$ for all i. The straight line with this slope m_i that passes through the hypersurface point X_i at $\lambda = 0$ is

$$x_I(\lambda) = m_I\lambda + X_i$$

For all such curves, we have

$$\frac{dx_i}{d\lambda} = m_i > 0 \text{ and } \frac{d^2x_i}{d\lambda^2} = \frac{dm_i}{d\lambda} = 0 \qquad i = 1, \ldots., r$$

Substituting these properties into the r expressions for $d^2L_I/d\lambda^2$, i=1, …, r, analogous to (14), and recalling n>0, it is easy to see that all the second derivative terms are greater than zero. Hence the point of intersection of each curve X_i with the hypersurface (12) is a true minimum along each curve for all the loss functions $L_1, \ldots, L_r$.

Appendix B. Credences Elicited by *n*-Power Scoring with *n*>1

The n-power scoring rule is generated by the functions (5a). The credences $\mathbf{x} = \langle x_1, x_2, \ldots, x_r \rangle$ it elicits for a subject's true probabilistic credences $\mathbf{p} = \langle p_1, p_2, \ldots, p_r \rangle$ are those that minimize the loss function.

$$
\begin{aligned}
L(\mathbf{p},\mathbf{x}) = \; & p_1(1-x_1)^n + \ldots + p_1 x_I^n + \ldots + p_1\, x_r^n \\
& + \ldots \\
& + p_i\, x_1{}^n + \ldots + p_i\,(1-x_i)^n + \ldots + p_i\, x_r^n \\
& + \ldots \\
& + p_r\, x_1{}^n + \ldots + p_r\, x_I^n + \ldots + p_r\,(1-x_r)^n \qquad\qquad (10b)
\end{aligned}
$$

To keep the analysis simple, consider only the generic case in which $p_I>0$, all i. The first and second derivatives of $L(\mathbf{p},\mathbf{x})$ with respect to x_1 are

$$\frac{\partial L}{\partial x_1} = -p_1 n(1-x_1)^{n-1} + (p_2+\ldots+p_r)n x_1^{\,n-1} = -p_1 n(1-x_1)^{n-1} + (1-p_1)n x_1^{\,n-1}$$

$$\frac{\partial^2 L}{\partial x_1^{\,2}} = p_1 n(n-1)(1-x_1)^{n-2} + (1-p_1)n(n-1)x_1^{\,n-2}$$

and similarly for $x_2, \ldots, x_r$. We seek the minimum loss with respect to $\mathbf{x}$ by setting all first derivatives to zero. We find for $i = 1, \ldots, r$, that $\partial L/\partial x_I = 0$ leads to

$$\left(\frac{x_i}{1-x_i}\right)^{n-1} = \left(\frac{p_i}{1-p_i}\right)$$

The values selected by this condition represent a true minimum since $\partial^2 L/\partial x_I^2 > 0$ for $0 \le x_I \le 1$, for all i. Solving for x_I, the credences elicited are

$$x_i = \frac{(p_i)^{1/(n-1)}}{(p_i)^{1/(n-1)} + (1-p_i)^{1/(n-1)}} \qquad (20)$$

The credences elicited will correspond to probabilities p_I only in the case of the Brier rule, n=2. For then we have

$$x_i = \frac{(p_i)^{1/(2-1)}}{(p_i)^{1/(2-1)} + (1-p_i)^{1/(2-1)}} = \frac{(p_i)}{(p_i)+(1-p_i)} = p_i$$

When n is not 2, but r=2, the rule will return additive credence x_1 and x_2:

$$x_1 = \frac{(p_1)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (1-p_1)^{1/(n-1)}} \;\text{ and }\; x_2 = \frac{(p_2)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (1-p_2)^{1/(n-1)}}$$

These elicited credences x_1 and x_2 will not correspond to the probabilities p_1 and p_2 unless we have the exceptional cases of p_1 =0 or p_1 =0.5 or p_1 =1.

In all other cases for n>1, we recover subadditive credences (for n>2) or superadditive credences (for 1<n<2).

To begin, consider the case of n>2. For r>2, we have from inequality (23) below that:

$$(p_2 + p_3 + \ldots + p_r)^{1/(n-1)} < (p_2)^{1/(n-1)} + (p_3)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)} \quad (21)$$

Using 1- $p_1 = p_2 + \ldots + p_r$, it becomes

$$(1 - p_1)^{1/(n-1)} < (p_2)^{1/(n-1)} + (p_3)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}$$

Substituting into (20) for the case of i=1, we have

$$x_1 = \frac{(p_1)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (1-p_1)^{1/(n-1)}} > \frac{(p_1)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (p_2)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}}$$

with similar formulae for $x_2, \ldots, x_r$. We see that these credences are subadditive if we sum them:

$$x_1 + x_2 + \ldots + x_r > \frac{(p_1)^{1/(n-1)} + (p_2)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (p_2)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}} = 1$$

where the credence in the set of all outcomes is 1. For the case of 1<n<2, using (24) below, we have, instead of (21), the inequality:

$$(p_2 + p_3 + \ldots + p_r)^{1/(n-1)} > (p_2)^{1/(n-1)} + (p_3)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)} \quad (22)$$

Following analogous reasoning, we arrive at superadditive credences

$$x_1 + x_2 + \ldots + x_r < 1$$

Appendix C. Useful Inequalities

The equalities used above are derived by considering the function

$$f(x) = (x+y)^{1/(n-1)} - x^{1/(n-1)} - y^{1/(n-1)}$$

for some fixed value of y>0. Its first derivative is

$$\frac{df(x)}{dx} = \frac{1}{n-1}\left((x+y)^{(2-n)/(n-1)} - x^{(2-n)/(n-1)}\right)$$

For n>2, the exponent satisfies -1<(2-n)/(n-1)<0. It follows that $df(x)/dx < 0$ for all x>0. Since f(0)=0, we have after integration of $df(x)/dx$ that $f(x)$<0. That is, for all x>0 and y>0, n>2,

$$(x+y)^{1/(n-1)} < x^{1/(n-1)} + y^{1/(n-1)}$$

Applying this inequality to $(z_2 + z_3 + \ldots + z_r)^{1/(n-1)}$ for all z_i>0, we recover

$$(z_2 + z_3 + \ldots + z_r)^{1/(n-1)} < (z_2 + z_3 + \ldots + z_{r-1})^{1/(n-1)} + (z_r)^{1/(n-1)}$$

and then

$$(z_2 + z_3 + \ldots + z_{r-1})^{1/(n-1)} + (z_r)^{1/(n-1)} < (z_2 + z_3 + \ldots + z_{r-2})^{1/(n-1)} + (z_{r-1})^{1/(n-1)} + (z_r)^{1/(n-1)}$$

Further iteration eventually leads to:

$$(z_2 + z_3 + \ldots + z_r)^{1/(n-1)} < (z_2)^{1/(n-1)} + (z_3)^{1/(n-1)} + \ldots + (z_r)^{1/(n-1)} \tag{23}$$

For $1<n<2$, we have that the exponent in $f(x)$ satisfies $(2-n)/(n-1)>0$. Proceeding as before we now have

$$(x+y)^{1/(n-1)} > x^{1/(n-1)} + y^{1/(n-1)}$$

which eventually leads to:

$$(z_2 + z_3 + \ldots + z_r)^{1/(n-1)} > (z_2)^{1/(n-1)} + (z_3)^{1/(n-1)} + \ldots + (z_r)^{1/(n-1)} \tag{24}$$

Appendix D. Equivalent Definitions of Strictly Proper Scoring Rules

To show the equivalence of the two definitions I and II of strictly proper scoring rules, it is sufficient to show that definition II entails definition I; and to show the converse entailment.

Strictly Proper II entails Strictly Proper I

The loss function $L(\mathbf{p}, \mathbf{x})$ of (10a) consists of a sum of r terms:

$$p_1\, g_0(x_i) + \ldots + p_i\, g_1(x_i) + \ldots + p_r\, g_0(x_i)$$

where $i = 1, \ldots, r$. Definition II entails that each of these r terms individually is minimized when $x_i = p_i$. To see this for $i=1$, the term is rewritten as

$$\begin{aligned} & p_1\, g_1(x_1) + p_2\, g_0(x_1) \ldots + p_i\, g_0(x_1) + \ldots + p_r\, g_0(x_1) \\ & \quad = p_1\, g_1(x_1) + (p_2 + \ldots + p_i + \ldots + p_r)\, g_0(x_1) \\ & \quad = p_1\, g_1(x_1) + (1 - p_1)\, g_0(x_1) \end{aligned}$$

Hence this term is minimized uniquely, according to definition II, when $x_1 = p_1$. The corresponding results for the remaining $x_2, x_3, \ldots$ follow analogously. Since $\mathbf{x} = \mathbf{p}$ minimizes each term uniquely, it follows that $\mathbf{x} = \mathbf{p}$ minimizes their sum, $L(\mathbf{p}, \mathbf{x})$, uniquely, which is definition I.

Strictly Proper I entails Strictly Proper II

Definition I applies for all p_i in $0 \le p_i \le 1$, $i = 1, \ldots, r$. Thus it applies to the case in which only $p_1>0$ and $p_2>0$, but $p_3 = p_4 = \ldots = p_r = 0$. In this special case, the loss function reduces to

$$\begin{aligned} L(\mathbf{p}, \mathbf{x}) = \; & p_1\, g_1(x_1) + p_1\, g_0(x_2) + \ldots + p_1\, g_0(x_i) + \ldots + p_1\, g_0(x_r) \\ & + p_2\, g_0(x_1) + p_2\, g_1(x_2) + \ldots + p_2\, g_0(x_i) + \ldots + p_2\, g_0(x_r) \end{aligned}$$

There are no terms in $L(\mathbf{p}, \mathbf{x})$ in $g_1(x_3), g_1(x_4), \ldots, g_1(x_4)$, but these variables only appear in $g_0(x_3), g_0(x_4), \ldots, g_0(x_r)$. Since all suitable functions for $g_0(x_i)$ are strictly increasing, the condition for minimization must include $x_i = 0 = p_i$, for $i = 3, 4, \ldots, r$. Hence the minimization of definition I reduces to the simpler problem of minimizing:

$$L(p_1, p_2, x_1, x_2) = p_1\, g_1(x_1) + p_1\, g_0(x_2) + p_2\, g_0(x_1) + p_2\, g_1(x_2)$$

That is, definition I requires minimization for fixed p_1 and p_2 of:

$$L(p_1, p_2, x_1, x_2) = p_1\, g_1(x_1) + (1\text{-}p_2)\, g_0(x_2) + (1\text{-}p_1)\, g_0(x_1) + p_2\, g_1(x_2)$$

Definition I stipulates that this minimum is achieved uniquely when $x_1 = p_1$ and $x_2 = p_2$. Since x_1 and x_2 can be varied independently in seeking the minimum, that minimum can only arise when the terms in which they appear

$$p_1\, g_1(x_1) + (1\text{-}p_1)\, g_0(x_1) \quad \text{and} \quad p_2\, g_1(x_2) + (1\text{-}p_2)\, g_0(x_2)$$

are individually, uniquely minimized by $x_1 = p_1$, for the first, and $x_2 = p_2$, for the second.

Either of these is equivalent to definition II, with the restriction that $0<p<1$. The complete definition II allows $0 \le p \le 1$. The two missing cases, p=0 and p=1, always conform with definition II, trivially. Hence definition I entails definition II.

References

Brier, Glenn W. (1950) "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, **78** (No. 1), pp. 1-3.

Brier, Glenn W. and Allen, Roger A. (1951) "Verification of Weather Forecasts," pp. 841-48 in T. F. Malone, ed., *Compendium of Meteorology*. Boston, MA: American Meteorological Society.

De Finetti, Bruno (1965) "Methods for DiscriminatingLevels of Partial Knowledge Concerning a Test Item," *The British Journal of Mathematical and Statistical Psychology*, **18**, pp. 87-123.

De Finetti, Bruno (1974) *Theory of Probability: A Critical Introductory Treatment.* Vol. 1. Chichester: John Wiley & Sons.

Gneiting, Tilmann, Raftery, Adrian T. (2007) "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, **102**, pp. 359-378.

Leitgeb, Hannes and Pettigrew, Richard (2010), "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy," *Philosophy of Science*, **77**, pp. 236-272.

McCarthy, John (1956) "Measures of the Value of Information," *Proceedings of the National Academy of Sciences*, **42**(9), pp. 654-55.

Joyce, James (1998) "A Nonpragmatic Vindication of Probabilism," *Philosophy of Science*, **65**, pp. 575–603.

Joyce, James (2009) "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief," in F. Huber and C. Schmidt-Petri, eds., *Degrees of Belief*. Synthese Library, 342. Sringer.

Pettigrew, Richard (2016) *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Predd, Joel B et al. (2009) "Probabilistic Coherence and Proper Scoring Rules," *IEEE Transactions of Information Theory*, **55**, pp. 4786–4792.

Rosenkrantz, Roger D. (1981) *Foundations and Applications of Inductive Probability*. Atascadero, CA: Ridgeview Publishing.

Savage, Leonard J. (1971) "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*. **66**, pp. 783-801.

Schervish, Mark (1989) "A General Method for Comparing Probability Assessors," *The Annals of Statistics*, **17**, pp. 1856-1879.

Schervish, Mark; Seidenfeld, Teddy; and Kadane, Joseph, (2009) "Proper Scoring Rules, Dominated Forecasts and Coherence," *Decision Analysis*. **6**, pp. 202-221.

Selten, Reinhard (1998) "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, **1**, pp. 43-62.

April 6, 16, 28, 2015
November 19, 2017; June 21, 2018

Chapter 12

No Place to Stand:

The Incompleteness of All Calculi of Inductive Inference[170]

1. Introduction

The previous two chapters have sought to show that the probability calculus cannot serve as a universally applicable logic of inductive inference. We may well wonder whether there might be some other calculus of inductive inference that can be applied universally. It would, perhaps, arise through a weakening of the probability calculus. The principal source of difficulty addressed in those chapters was the additivity of the probability calculus. Such a weakening seems possible as far as additivity is concerned. Something like it is achieved with the Shafer-Dempster theory of belief functions. However there is a second, lingering problem. Bayesian analyses require prior probabilities. As we shall see below, these prior probabilities are never benign. They always make a difference to the final result.

For a long time, I hoped to find an extension of or alternative to the probability calculus that would afford us a truly neutral initial state. We could then proceed to incorporate the evidence, free from the worry that the unsupported choice of a prior state might somehow compromise the analysis. These efforts failed, again and again. Eventually I came to see that they failed for a good reason of principle: there is no calculus of inductive inference that can support this fully neutral initial state and still admit the nontrivial incorporation of new evidence.

A technically detailed statement and demonstration of this result is in Norton (forthcoming) and readers are referred to it for these details. The burden of this present chapter is to give an introductory account of the result and its import, suppressing as much as possible of the distracting technical details. For, as we shall see, the result itself rather simple in conception. Indeed it is so simple that I believe the only reason we have not had the result as a staple in our literature is that no one thought to look for it.

The earlier Sections 2-5 below describe what it would be for a calculus of inductive inference to be complete, using the illustration of the Bayesian analysis of simplicity; and they

[170] I am grateful for helpful discussion especially to Wayne Myrvold and to Yann Benétreau-Dupin and the Fellows of the Center for Philosophy of Science, Spring Term, 2015, who urged me to write this introductory account.

explain why completeness is desirable, if only it could be secured. In brief, completeness is achieved when computations in the calculus are carried out in a domain sufficiently large so that the computations do not need to call upon inductive content that is external to the domain. Completeness provides us an evidentially neutral "place to stand"[171] prior to any considerations of evidence. We then modify this initial state, moving away from neutrality, under the import of evidence. This neutral starting point would allow us to characterize inductive inference merely as inference that conforms to the calculus at issue, for no external inductive content would be needed. Any deviations from neutrality would solely result from the import of evidence. This characterization would provide a clear and simple solution to the enduring foundational problems of inductive inference. All such problems would be reduced to questions answerable by computation in the calculus.

This attractive solution to the foundational problems fails. Non-trivial calculi of inductive inference are incomplete. None provide an evidentially neutral place to stand. These incomplete calculi include many more than just the probability calculus. This incompleteness explains why particular calculi of inductive inference are beset by lingering difficulties. The Bayesian system is perpetually struggling to overcome the problem of the priors. Augmented calculi are repeatedly proposed to solve problems in older calculi, while none manages without its own, new problems. All these problems arise because we are really trying to formulate a complete calculus of inductive inference. That they must linger unsolved does not derive from a failure of our imagination to hit upon just the right solution. It is a necessity derived from incompleteness.

Section 6-14 below provide a simplified guide to the full proof of this failure, given elsewhere. Here is a terse summary of the main result that will be introduced and explained in greater detail in this chapter. The incompleteness arises from the combination of two desirable properties of calculi of inductive inference.

The first property is an expression of completeness: we can find a sufficiently large set of propositions in which the inductive strengths of support are fixed by relations in the set, without the need to import any inductive content from outside it. Since the only other inferential resources within the set are the deductive relations among the propositions, this amounts to requiring that the inductive strengths of support are fixed by the deductive relations among the propositions in the set. This requirement is unremarkable. The Kolmogorov axioms of probability theory are a routine part of such a specification. These axioms adapt the probabilities to the deductive structure. They need only a small supplement to fix the probabilities uniquely.

---

[171] This phrase alludes to Archimedes' celebrated boast in the context of the principle of the lever: "Give me a place to stand and I shall move the world."

The second property involves disjunctive refinements of propositions. Through them we replace a proposition

"Person X is in Boston."

by a disjunction of its disjunctive parts:

"Person X is in Boston-location-1 or Person X is in Boston-location-2 or
… or Person X is in Boston-location-r."

Such disjunctive refinement increases the expressive power of the set of propositions and leads to adjustments of the inductive strengths of support. The requirement of asymptotic stability asserts that continuing disjunctive refinement eventually provides such diminished further power that the inductive strengths of support among some fixed set of propositions stabilize to limiting values. Further refinement eventually becomes inert, inductive hair-splitting.

The failure of the completeness resides in the impossibility of sustaining both properties.[172] In briefest terms, the deductive closure of any set of propositions is highly symmetric. Each of the non-contradictory, logically strongest propositions—the "atoms"—enter into the same deductive relations. As a result, a deductively definable logic of induction must treat them alike. Each new disjunctive refinement will alter the atoms and, as a result, the inductive strengths throughout the set. It turns out that a deductively definable logic of induction will continue to respond without stabilization to suitably crafted, continuing disjunctive refinements, unless it is a trivial logic that assigns the same limiting inductive strengths everywhere.

One might be tempted by an obvious rejoinder: if continuing refinement causes continuing problems, stop refining! Declare one specific refinement as preferred; or declare that its propositions comprise a preferred language. That resolves the problem. But the decision of when to stop or which is the preferred language must be made on external, inductive grounds. It privileges certain propositions and thus amounts to the introduction of external inductive content, in violation of the requirement of completeness.

The concluding Sections 15-18 of this chapter take stock and review possible responses.

2. The Appeal of a Calculus of Inductive Inference

Those who have read through the earlier chapters in this book should be in no doubt of one thing: rule-based accounts of inductive inference are not in good shape. Simple enumerative induction fails more than it succeeds. It is *almost never* the case that, when some As are B, it also

[172] The proof strategy is an extension of the familiar problems introduced by the principle of indifference probabilistic logic. See Norton (2008) for discussion.

happens that all As are B. The replicability of experiment is the gold standard of science, we are told, never to be discounted, except when we do discount it. If we seek the formal template to which arguments from analogy must conform, we find prescriptions of ever growing complexity that never reach an endpoint. We should infer to the best explanation. Yet it is an instruction that is hard to follow since we are offered no precise characterization of just what is a good explanation or why explaining, whatever it is, has such evidential powers. Finally, to mention an example to which we will return below, evidence favors simpler hypotheses, we are told. But we have no serviceable characterization at the most general level of what makes an hypothesis simpler or why such hypotheses should be favored.

These are just the beginnings of the difficulties. Over the centuries, inductive inference has attracted a fulsome collection of general problems that threaten the very cogency of this form of inference. We have Hume's problem, Hempel's raven, Goodman's grue and Quine's underdetermination. The difficulties are so enduring that mere mention of induction calls philosophical pain to mind.

The tenacity of these problems stands in striking contrast with deductive inference. While there are always complications at the fringes, the core is stable to the point of tedium. Modus ponens is a valid argument. Affirming the consequent is a fallacy. These facts of logic leave no room for doubt or debate. We separate the valid from the invalid deductive inferences merely by checking whether the argument form used is one of the approved argument forms in a logic textbook. The exercise is reminiscent of making travel plans by checking a train timetable.

In this regard, deductive logic is more like arithmetic than inductive inference. It is an uncontested, particular fact of arithmetic that 7,919 is the thousandth prime number; and it is merely a matter of tedious computation using standard algorithms to check it. More general facts have a similar security. That there are infinitely many prime numbers is proved by a theorem known since the time of Euclid. Anyone who doubts the infinity of the primes can consult the proof and, by working through its steps, receive all the assurance a reasonable person could require.

Might the problems of inductive inference be resolvable in a similar way? Might the puzzles of induction be converted into queries that can be put to and answered by mechanical computation in some suitable calculus? The presently most popular approach to inductive inference, the Bayesian approach, holds out the promise of such a solution. The approach is based on the supposition that inductive support or warranted belief is captured by the mathematical calculus of probabilities. Much of Bayesian analysis is the tedious working of proofs in the calculus. The strength of inductive support provided by some item of evidence for some hypothesis is computed numerically as a conditional probability. General facts about inductive inference are established as theorems of the probability calculus, much as Euclid

proved the infinity of the primes. In each case, we have the comforting assurance that, one way or another, a computation will provide precise answers to our questions.

3. A Bayesian Analysis of Simplicity

Here is an illustration of such a computation. A familiar principle is that evidence favors a simpler hypothesis. For example, as we saw in earlier chapters, when we fit a curve to data, we may find a good enough fit from the hypothesis of a straight line and a slightly better fit from a parabola. We are routinely willing to forgo a slightly better fit by a parabola for the lesser fit of straight line, because we prefer to use the simpler hypothesis.

This preference for the simpler can be vindicated in Bayesian analysis. The key to it is that there are fewer of the simpler hypotheses. A straight line—"$y = ax + b$"—is fixed by just two adjustable parameters, a and b. A parabola—"$y = ax^2 + bx + c$"—is fixed by three parameters, a, b and c. Hence there are many more of the more complicated hypotheses. The straight line hypotheses form a two dimensional space. The parabolic hypotheses form a three dimensional space.

A still simpler example uses this fact and will suffice to get to the key point. Imagine that we have to choose between a simple hypothesis and a more complicated one. Let us say that the simple hypotheses is drawn from a ten-membered set $\{H_{sim1}, H_{sim2}, \ldots, H_{sim10}\}$ of hypotheses of comparable simplicity. The complicated hypothesis is drawn from a much larger, one-hundred-membered set $\{H_{com1}, H_{com2}, \ldots, H_{com100}\}$ of hypotheses of comparable complication. We shall assign equal prior probability to each set:

$$P(\{H_{sim1}, H_{sim2}, \ldots, H_{sim10}\}) = P(\{H_{com1}, H_{com2}, \ldots, H_{com100}\}) \quad (1)$$

where conditionalization on a background Ω is supposed but not represented. We then spread the probability uniformly within each set. Since the second set has ten times as many members as the first, the prior probability of any of individual simple hypothesis $H_{sim\ i}$ is ten times as great as the prior probability of any of the complicated hypotheses $H_{com\ k}$.

$$\frac{P(H_{sim\ i})}{P(H_{com\ k})} = 10 \quad (2)$$

Let us say that the two hypotheses $H_{sim\ i}$ and $H_{com\ k}$ fit roughly equally well with the evidence. That is, the supposition of each makes the evidence E roughly equally probable:

$$P(E|H_{sim\ i}) \approx P(E|H_{com\ k})$$

so that the ratio of likelihoods $P(E|H_{sim\ i}) / P(E|H_{com\ k}) \approx 1$. The relative strength of support from the evidence and background together for the hypotheses is expressed by the ratio of

posterior probabilities $P(H_{\text{sim i}}|E)$ / $P(H_{\text{com k}}|E)$. It can be calculated with the ratio form of Bayes' theorem:

$$\frac{P(H_{sim\ i}\,|\,E)}{P(H_{com\ k}\,|\,E)} = \frac{P(E\,|\,H_{sim\ i})}{P(E\,|\,H_{com\ k})} \cdot \frac{P(H_{sim\ i})}{P(H_{com\ k})}$$

Since the likelihood ratio is approximately one, the ratio of the priors (2) is the deciding factor that gives a large boost to the probability of the simpler hypotheses:

$$\frac{P(H_{sim\ i}\,|\,E)}{P(H_{com\ k}\,|\,E)} \approx \frac{P(H_{sim\ i})}{P(H_{com\ k})} = 10 \qquad (3)$$

In brief, since there are fewer simpler hypotheses, a natural spreading of prior probabilities (1) can assign higher prior probability to the simpler hypotheses. When the evidence is equivocal in choosing among the hypothesis, this higher prior probability gives the simpler hypothesis the decisive advantage.

While this captures the essentials of the Bayesian analysis, more realistic cases are messier. There are almost always infinitely many hypotheses grouped into one complexity class and then, in addition, infinitely many such classes. Simply counting hypotheses no longer works. More sophisticated analyses are needed, while the essentials remain the same. Jeffreys (1961, p. 47) measures the complexity of classes of curves by the sum of the order, the degree and the absolute values of the coefficients of a suitably reduced differential equation that governs the curves. Solomonoff (1964) measures complexity as algorithmic complexity; that is, the measure is the size of the smallest universal Turing machine program needed to generate the hypothesis. They both then exponentially penalize the prior probability of each complexity class so that the probabilities can sum to unity.

4. External Inductive Content

In many examples like this, Bayesian analysis has been able to reduce an inductive puzzle to a computation in the probability calculus. In each case, however, it turns out that the analysis is not self-contained. Each requires supplement by external inductive content. That is, the computation depends on direct or indirect specification of inductive strengths of support by considerations external to the computation.

Take the case of the analysis of simplicity above. We assigned equal probability to the two complexity classes in (1) and then spread the assigned probability uniformly within each class. The outcome was that each of the simpler hypotheses was assigned a greater prior probability; and this was key to the whole analysis. Yet nothing within the probabilistic computation forced this assignment. We could merely have assigned the same prior probability to each hypothesis individually

$$P(H_{\text{sim1}}) = P(H_{\text{sim2}}) = \ldots = P(H_{\text{sim10}}\})$$
$$= P(H_{\text{com1}}) = P(H_{\text{com2}}) = \ldots = P(H_{\text{com100}}) \quad (1')$$

This alternative assignment would have defeated the analysis. For then, instead of (2), we would have had:

$$\frac{P(H_{sim\ i})}{P(H_{com\ k})} = 1 \quad (2')$$

and the simpler hypothesis would have received no probabilistic boost:

$$\frac{P(H_{sim\ i} \mid E)}{P(H_{com\ k} \mid E)} \approx \frac{P(H_{sim\ i})}{P(H_{com\ k})} = 1 \quad (3')$$

The point is not that the assignment of (1) is unjustifiable. One could certainly conceive circumstances in which we would be warranted in assigning a higher prior probability to a simpler hypothesis. And we could conceive others in which this might not be so.

The point is that the assignment of (1) is provided externally to the probabilistic computation that takes us from (1) to the main result (3). This means that the recovery of the result (3) by the computation is not inductively self-contained. Essential inductive content is provided from an external source. To preclude confusion, by "inductive content" I mean merely the assignments of probability in (1) or (1').

5. The Ideal of Completeness

A natural response to the presence of the external inductive content in the Bayesian analysis of simplicity is that we have set our boundaries too narrowly. That the simpler hypotheses ought to be assigned a higher prior probability is something that can in turn be learned inductively. In Jeffreys' analysis of simplicity, we are to assume that nature favors curves drawn from the simpler of his complexity classes. In Solomonoff's analysis, we are to assume that nature favors hypotheses that are algorithmically simpler. Neither of these are a priori truths. They are contingent facts about the world. Ascertaining their truth is a matter of further inductive investigation. If we extend the boundaries of our computation, we would hope to capture those considerations as well.

What if those considerations in turn depend upon further external inductive content? We would then extend our boundaries still further. Let us suppose that it is possible to extend the boundary of the computational domain so far that no external inductive content is needed. What would result is an account of all the relations of inductive support within the domain that is fully contained in a single, enormous computation in the probability calculus.

While such an enormous computation would surely outstrip any human powers of

comprehension, its possibility in principle is of profound foundational importance. It would mean that the probability calculus is all we need for a full understanding of inductive inference within a suitably large domain.

All particular facts of inductive support within that domain would be expressible by particular probabilistic relations among its propositions. That the straight-line hypothesis is better supported by the evidence would be expressed by its greater probability; and so on for every other particular fact of inductive support.

The same would be true for general facts about inductive inference. Every general fact could, in principle, be captured by some general theorem within this huge computation. If, for example, simpler hypotheses are favored evidentially in this domain, that general fact would be captured by a theorem. It would assert that the prior probabilities of hypotheses in simpler classes must, in general, be higher, as in (1). All this, at both the level of the particular and the general, could be known without drawing upon any inductive content from outside the domain. The analysis would be self-contained.

6. Its Failure

What is shown in Norton (forthcoming) and will be reviewed below is that this ideal of completeness is unattainable. A very large class of possible calculi that likely includes any calculus one might realistically consider, proves unable to support this ideal of completeness. This failure is profound foundationally. It tells us something important about the nature of inductive inference itself: it cannot be fully characterized merely by a calculus.

To get a sense of this import, it is helpful to compare it with the familiar incompleteness of arithmetic. It was once quite reasonable to expect that all the truths of arithmetic could be captured by a few axioms. For example, Peano's axioms lay down a few simple properties of natural numbers: 1 is a number; every number has a unique successor; and so on. We would hope that we could identify all of arithmetic with all the truths that can be deduced from these axioms.

Famously, Gödel demonstrated that no finite axiom system can capture all arithmetic truths in this way. The truths of arithmetic are something more than what can be deduced from any fixed, finite system of axioms. We may, of course, be able to derive very many important and interesting arithmetic truths from our favorite axiom system. However, no matter which finite axiom system we favor, there will always be arithmetic truths that are external to its theorems.

I hesitate to draw a comparison with Gödel's result, for his result is profound and his methods extraordinarily ingenious. The corresponding methods for inductive calculi are simple and mechanical and the result rather banal. But the significance of the result for inductive logic is comparable.

We may have a favored calculus for inductive inference and be able to infer many important and useful results within it. We might then seek to characterize inductive inference merely as inference that conforms with some specific calculus, such as the probability calculus. The incompleteness tells us that characterization fails. There is always more to inductive support than can be captured by the calculus. Searching for theorems within a favored calculus can only ever return a partial understanding. Inductive inference cannot be reduced to inference that conforms with some favored calculus.

7. Deductive Preliminaries

7.1 Deductive Structure

How is the incompleteness demonstrated? The first step is to fix the environment in which the inductive logic is applied. We take a fixed set of propositions

$$\{A_1, A_2, \ldots, A_m\}.$$

and our concern will be to determine the inductive relations prevailing among these propositions. This set is intended to be not just large, but very large. It might be all the hypotheses entertained in science, all the evidence statements that may support them and every other proposition that in some way mediates between them. That set—all the propositions we have entertained in science—will be large. But it will still be finite. For there have only been finitely many scientists and, given some finite limit of the length of sentences, only finitely many propositions expressible.

These propositions come with a deductive structure. That structure is just the set of all deductive entailment relations among the m propositions. It may turn out, for example, that A_{1235} deductively entails A_{441}; or that A_{57} and A_{103} are logically incompatible, so that their conjunction entails the contradiction $\varnothing$. The deductive structure is the totality of these deductive relations.

It will be essential for what follows to see that this structure is highly symmetric. That symmetry is harder to see if we consider merely the propositions $A_1, A_2, \ldots, A_m$ by themselves. Rather we take the larger set of propositions generated by Boolean operations; that is, by taking all negations ("not" ~), disjunctions ("or" v) and conjunctions ("and" &) of the propositions. The set of sentences that results is infinite. However the set of logically distinct propositions is not. The set contains many logically equivalent sentences. The sentence A_1, for example, is logically equivalent to all of $\sim\sim A_1$, $\sim\sim\sim\sim A_1$, $A_1 \text{v} A_1$, $A_1 \& (A_2 \text{v} \sim A_2)$, etc.

7.2 A Boolean Algebra of Propositions

The deductive structure, with all these duplications eliminated, is best characterized by identifying its "atoms." These are the logically strongest (non-contradictory) propositions. A finite set of propositions can support only finitely many atoms. Take the simple case of two propositions in the set $\{A, B\}$, where we assume they are logically compatible and do not exhaust the space. Then there are four distinct atoms:

$$a_1 = A\&B \quad a_2 = A\&{\sim}B \quad a_3 = {\sim}A\&B \quad a_4 = {\sim}A\&{\sim}B$$

Each of the propositions a_1, a_2, a_3 and a_4 is an atom since nothing (other than the contradiction $\varnothing$) entails it.

These four atoms generate a four-atom Boolean algebra of finitely many propositions, which has five distinct logical levels

the universal proposition: $\Omega_4 = a_1 \vee a_2 \vee a_3 \vee a_4$

three-atom disjunctions: $a_1 \vee a_2 \vee a_3$, $a_1 \vee a_2 \vee a_4$, $a_1 \vee a_3 \vee a_4$, $a_2 \vee a_3 \vee a_4$

two-atom disjunctions: $a_1 \vee a_2$, $a_1 \vee a_3$, $a_1 \vee a_4$, $a_2 \vee a_3$, $a_2 \vee a_4$, $a_3 \vee a_4$

atoms: a_1, a_2, a_3, a_4

the contradiction: $\varnothing$

The original propositions A and B reside within this Boolean algebra as $A = a_1 \vee a_2$ and $B = a_1 \vee a_3$. Figure 1 is a picture of the algebra, showing the distinct levels. The arrows represent deductive entailment.

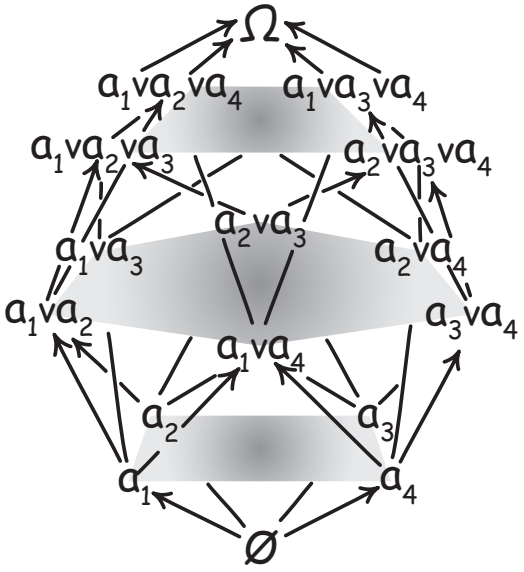


Figure 1. A Four-Atom Boolean Algebra

7.3 Symmetries of Deductive Structure

A Boolean algebra is a highly symmetric structure. Informally speaking, each level is homogeneous. That is, the entire algebra "looks the same" from any proposition we pick in the level. For example, take the two-atom disjunction level of the four-atom algebra. Each disjunction in it is entailed by two atoms; and each disjunction in the two-atom layer in turn entails just two three-atom disjunctions. The only change, as we move around within one of the levels is the labeling of the atoms that appear in the deductive entailments.

When there are very many atoms in the algebra, the basic structure remains the same. There are now, however, many more levels: the one-atom level, the two-atom level, the three-atom level, and so on for very many more levels. As before, each level in the algebra is homogenous. That is, the algebra looks the same, as far as deductive relations are concerned, from each proposition in the same level.

More formally, this symmetry is expressed as a labeling invariance. That is, the total deductive structure is unchanged if we permute the labels attached to the atoms. Take the four atoms

$$a_1, a_2, a_3, a_4$$

and permute their labels any way you please. You might just switch the first two, so that the atoms are now labeled

$$a_2, a_1, a_3, a_4$$

Or you might cyclically permute them to

$$a_2, a_3, a_4, a_1$$

In both cases, propagate the labeling change through the remainder of the algebra. For these permutations and for any others, the total deductive structure will remain unchanged. If a_1 entails $a_1 \vee a_2$ entails $a_1 \vee a_2 \vee a_3$ prior to the permutations of atomic labels, the same will be true for the relabeled propositions.

The symmetry is easier to see geometrically in a simpler figure that shows just a three-atom algebra. Figure 2 shows the same three-atom algebra, differing only in the arbitrary labeling for the atoms. Labels a_1 and a_2 are switched on one side; and atom labels a_1, a_2 and a_3 are cyclically permuted on the other:

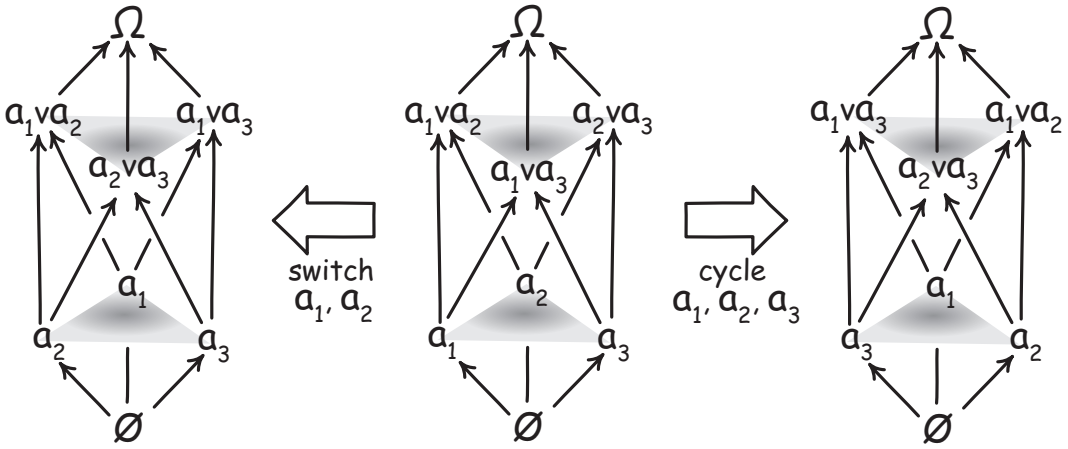


Figure 2. Relabelings of a Three-Atom Algebra

8. Deductively Definable Logics of Induction

8.1 Rules Define Strengths of Inductive Support

A calculus of inductive inference will here be built around the fundamental quantity "[A|B]", which is the strength of the inductive support afforded proposition A by proposition B. The strength might be a conditional probability, which means that it conforms with the probability calculus. The strength need not be a probability. It may be a strength that conforms with one of many other calculi.

Other choices are possible for the basic quantity. We could instead use "[A|B,C]", which could be interpreted as the strength of inductive support afforded proposition A by B with respect to background C. It will become clear that the arguments leading to incompleteness can be mounted in variant form for each of these choices. We will proceed with just [A|B] since it is all that is needed to see how the arguments run.

A calculus of inductive inference is a system of rules that enables the assignment by purely mechanical computation of all the strengths $[A_i|A_k]$ for propositions in the set $\{A_1, A_2, \ldots, A_m\}$. They key question is which resources these rules may use. If the domain in which the set resides is sufficiently large for completeness to obtain, then the rules may not use any inductive content from outside the domain. That is, it may not set any of the $[A_i|A_k]$ by external considerations independent of the rules of the calculus.

This restriction then leaves as the sole resource the deductive relations among the propositions in the set $\{A_1, A_2, \ldots, A_m\}$ and their deductive relations with the other propositions in the larger algebra Ω in which it resides. A calculus that employs just this

deductive structure in specifying its strength is "deductively definable."

8.2 Two Sample Logics

At first it may seem that deductive definability is excessively restrictive. It is not. Rather it is the standard way of specifying a calculus of this type. As a general matter, the definitions of the strengths $[A_i|A_k]$ may be supplied by explicit or implicit definitions.

The latter implicit definitions are more commonly used. The celebrated Kolmogorov axioms (1950) for the probability axioms provide implicit definitions solely in terms of the deductive structures among the propositions in the outcome space. These axioms, used to define an additive measure m on the algebra, assert:

$$\text{For any } A,\ m(A)\geq 0. \tag{4a}$$

$$m(\Omega) = 1 \tag{4b}$$

$$\text{If } A\&B=\varnothing, \text{ then } m(A \text{v} B) = m(A) + m(B) \tag{4c}$$

This is an implicit definition of the additive measure m. It consists of three sentences in which the measure appears; and those sentences otherwise only mention the deductive structure of the algebra. For example, (4b) assigns unity to the universal proposition Ω, distinguished by the fact that it is deductively entailed by all the propositions in the algebra. The summation rule relates the measure of a disjunction to the measures of the disjuncts, in the special case in which the disjuncts are deductively incompatible.

The Kolmogorov axioms constrain the measure m, but do not definite it uniquely. In any given algebra, there will be infinitely many measures compatible with the axioms. We can assure uniqueness of m in some algebra by adding further conditions, such as:

$$\text{For all atoms, } a_1, a_2, \ldots, a_n, \tag{5}$$

$$m(a_1) = m(a_2) = \ldots = m(a_n)$$

Once again, this sentence mentions only deductive structure. The atoms $a_1, a_2, \ldots, a_n$ are the propositions in the algebra that are deductively entailed by no other propositions (other than the contradiction, $\varnothing$).

This uniquely defined additive measure can now be used to introduce the familiar inductive strength of support, a conditional probability. For all propositions A and B, where $m(B)$ is not 0

$$[A \mid B]_P = P(A \mid B) = \frac{m(A \,\&\, B)}{m(B)} \tag{6}$$

In order to underscore that these results apply to many calculi, we can also define a different calculus—a "specific conditioning" logic—by replacing (6) by the following.[173] For all

[173] For more details of the properties of a special conditioning logic, see Norton (2010, §11.2).

propositions A and B, where neither $m(A)$ nor $m(B)$ is 0

$$[A \mid B]_{SC} = P(A \mid B) = \frac{m(A \,\&\, B)^2}{m(A)m(B)} \tag{7}$$

We will see shortly in an example what motivates this logic.

8.3 General Form of the Definitions

The conditions (4), (5) and (6) implicitly define a probabilistic calculus of inductive inference. The conditions (4), (5) and (7) implicitly define a distinct "specific conditioning" calculus of inductive inference. What will matter in what follows is the general form of the definitions:

> *General form of the implicit definition:*
> A set of sentences that mention the strengths $[A_i|A_k]$ and deductive relations among the members of the set $\{A_1, A_2, \ldots, A_m\}$ and the other propositions in the algebra.

These two examples are just two of many possible deductively definable logics of induction. More are described in Norton (2010).

A simple and natural one derives from the basic notion of hypothetico-deductive confirmation. According to it, if hypothesis H deductively entails evidence E, then evidence E inductively supports H. This much provides for a single value "*supports*" for $[H|E]$ via the explicit definition:

> If H deductively entails E, then $[H|E]$ = *supports*.

There is much scope to enhance the definition. We might replace the single value with increasing numerical values the closer that H is to E in terms of the levels of the Boolean algebra. If, for example, $H = a_1 \vee a_2$ from the level of two atom disjunctions and $E = a_1 \vee a_2 \vee a_3 \vee a_4$ from the level of four atom disjunctions, then the strength of support might be defined as 2/4. Then the closer they are in levels, the stronger the support. This gives the augmented definition[174]

> If H from the level of m atom disjunctions
> deductively entails E from the level of n atom disjunctions,
> then $[H|E] = m/n$.

This second example illustrates the general form of an explicit definition of inductive strengths:

> *General form of the explicit definition:*
> The strengths $[A_i|A_k]$ are determined by a formula that mentions only the deductive relations among the members of the set $\{A_1, A_2, \ldots, A_m\}$ and

[174] This definition induces a product rule. If A entails B entails C, then [A|C] = [A|B]x[B|C].

the other propositions in the algebra.

In the example, the formula is "m/n", where the quantities n and m are related to atom counts and are thus recoverable from the deductive structure of the Boolean algebra.

This hypothetico-deductive model could be enhanced still further by rewarding hypotheses with stronger support if they are more explanatory or simpler. To do this requires that we have some way of identifying which hypotheses are more explanatory or which are simpler. If that can be done by adding further propositions to the algebra, then the definition of the inductive strengths can still meet the requirement that they draw only on resources within the domain. If that cannot be done and these judgments require resources outside the domain, then we have already established that these particular augmentations of the hypothetico-deductive scheme are not complete.

9. The Quest for an Art Thief

As an illustration of the application of these logics, we will imagine an inductive problem presented to the police in their efforts to track down the location of a notorious art thief. They know, we shall say, that the art thief is in one of four cities: Boston "BOS", New York "NY", Philadelphia "PHL" or Pittsburgh "PIT." That is we have

$$\Omega = \text{BOS v NY v PHL v PIT}$$

These four propositions are the atoms of the algebra. Their evidence is that the thief is in an east coast, Atlantic port city "EC":

$$\text{EC} = \text{BOS v NY v PHL}$$

We can then ask how much support EC provides to the various possibilities. We have from the Kolmogorov axioms (4) and condition (5) that

$$m(\text{BOS}) = m(\text{NY}) = m(\text{PHL}) = m(\text{PIT}) = 1/4$$

It follows from the definition (6) that the evidence EC gives the same support to the hypothesis BOS as it does to the disjunction BOS v PIT

$$P(\text{BOS} \mid \text{EC}) = P(\text{BOS v PIT} \mid \text{EC}) = 1/3$$

This is a familiar property of conditional probability. Since the proposition PIT contradicts the evidence EC, forming a disjunction with BOS does not alter the conditional probability.

While it is familiar, this is an oddity of probabilistic support. Unless we have honed our sense of evidential support on probabilistic notions, we would judge the support provided by EC for BOS to be weakened when we form a disjunction with a city PIT that contradicts the evidence. The evidence specifically supports BOS, not PIT. Within the probabilistic analysis, we can recover the fact that the PIT disjunct plays no role in the support accrued to BOS v PIT by noting that the probability is unchanged when we eliminate the PIT disjunct. The awkwardness is

that we have to do this additional computation to learn that the evidence points better to BOS rather than BOS v PIT.

The specific conditioning logic (7) is designed to remedy this defect. It does the work of discriminating between BOS and BOS v PIT by assigning a lower strength of support to BOS v PIT. That is, we have

$$[BOS \mid EC]_{SC} = \frac{m(BOS \,\&\, EC)^2}{m(BOS)m(EC)} = \frac{1^2}{1 \cdot 3} = \frac{1}{3}$$

whereas

$$[BOS \vee PIT \mid EC]_{SC} = \frac{m((BOS \vee PIT) \,\&\, EC)^2}{m(BOS \vee PIT)m(EC)} = \frac{1^2}{2 \cdot 3} = \frac{1}{6}$$

so that 1/6 = [(BOS v PIT) | EC]$_{SC}$ < [BOS | EC]$_{SC}$ = 1/3. Perhaps in this case, the advantage of the specific conditioning logic is unclear. But that is only because we can "see through" the example and recognize the odd, disjunctive character of the hypothesis BOS v PIT. In more complicated cases, this might not be possible and we would benefit from the specific conditioning logic doing the work of recognizing the oddity for us.

10. Symmetry Constraints on Deductively Definable Inductive Logics

Two properties of the systems developed here combine to place powerful constraints on the inductive logics.

First, the inductive logic is deductively definable. It follows directly from the above general implicit and explicit definitions that, if two sets of propositions agree in their deductive relations, then they must agree in their inductive relations. That is, assume that a set of proposition are A, B, C, … and can be mapped to the second set A', B', C', … in a way that preserves deductive structure. It follows that the inductive strengths formed from A, B, C, … must agree with the corresponding strengths formed from A', B', C', …

Second, the deductive structure is highly symmetric. This means that the deductive structure preserving map can be implemented within a single algebra of propositions merely by relabeling the propositions. It then follows that many of the inductive strengths formed within the single algebra must be equal.

10.1 An Illustration

We can see how these equalities arise in the example of the art thief. Consider the support afforded by EC for each of BOS and NY. That is, compare [BOS|EC] and [NY|EC]. We shall see that they must be equal.

To see this, we relabel BOS and NY as:

BOS' = NY and NY' = BOS

The two remaining atom labels are unchanged other than for the addition of a prime:

PHL' = PHL and PIT' = PIT.

One sees immediately that the deductive structure of the propositions with the primed labels is the same as the deductive structure of the propositions with the unprimed labels. That is, for every deductive entailment in the first there is a corresponding deductive entailment in the second; and vice versa. For example, BOS deductively entails EC = BOS v NY v PHL. Correspondingly BOS' deductively entails EC' = BOS' v NY' v PHL'.

Since the inductive logic is deductively definable, it now follows that all corresponding inductive strengths must agree. That is we have:

[BOS|EC] = [BOS'|EC']

[NY|EC] = [NY'|EC']

[PHL|EC] = [PHL'|EC']

[PIT|EC] = [PIT'|EC']

etc,

The primed propositions are merely relabelings of the unprimed propositions. In particular, BOS' = NY and EC' = EC. Making the replacements in the first equality [BOS|EC] = [BOS'|EC'] gives the result promised

[BOS|EC] = [NY|EC].

We can see informally how this equality comes about. It arises because the BOS-EC relationship is, roughly speaking,

"single atomic proposition deductively entails three-atom disjunction."

The NY-EC relationship is the same. Since the deductive structures involved are the same, the correspondingly inductive strengths must be the same.

10.2 The Symmetry Theorem

The symmetry constraint can be generalized. Take a slightly more general case of a deductively definable logic in which the inductive strengths [*A*|*B*] are fixed by the deductive relations among *A* and *B* and the remaining propositions of the algebra. When might we have an equality of two strengths [*A*|*B*] and [*C*|*D*]? It arises when there is some relabeling possible for the atoms in the algebra, so that *A* and *B* are relabeled as *A*' and *B*' and

$$A'\&B' = C\&D$$

$$A'\&\sim B' = C\&\sim D$$

$$\sim A'\&B' = \sim C\&D$$

$$\sim A'\&\sim B' = \sim C\&\sim D$$

This relabeling will be possible just in case the conjunctions to be set equal are formed from the

same number of atoms. That is, the same number of atoms disjoined to form $A\&B$ and to form $C\&D$; as so on for the remaining equalities, so that

$$\#A\&B = \#C\&D$$
$$\#A\&\sim B = \#C\&\sim D$$
$$\#\sim A\&B = \#\sim C\&D$$
$$\#\sim A\&\sim B = \#\sim C\&\sim D$$

where the notation "#proposition" indicates the number of atoms disjoined to form the proposition.

Then, by reasoning analogous to that of the last section, we can show that the deductive relations into which A and B enter are the same as those into which C and D enter. It now follows that the inductive strength $[A|B]$ is fixed by the atom counts of these four conjunctions. That is:

> *Symmetry Theorem*
> For each deductively definable logic in which the inductive strengths $[A|B]$ are fixed by the deductive relations among A and B and the remaining propositions of the algebra, there exists a function f such that $[A|B] = f(\#A\&B, \#A\&\sim B, \#\sim A\&B, \#\sim A\&\sim B)$

We can illustrate this theorem in the case of the two logics considered above. For the probabilistic logic we have

$$[A \mid B]_P = P(A \mid B) = \frac{\#A\,\&\,B}{\#A\,\&\,B \;+\; \#\sim A\,\&\,B} = \frac{\#A\,\&\,B}{\#B}$$

For the specific conditioning logic, we have

$$[A \mid B]_{SC} = \frac{(\#A\,\&\,B)^2}{(\#A\,\&\,B \;+\; \#A\&\sim B)\cdot(\#A\,\&\,B \;+\; \#\sim A\,\&\,B)} = \frac{(\#A\,\&\,B)^2}{\#A\cdot\#B}$$

In general, the specification of a new inductive logic merely requires the specification of a new function f in the theorem.

This formulation of the symmetry theorem is not the most general formulation. In general, the strengths $[A_i|A_k]$ are fixed by deductive relations among the large set $\{A_1, A_2, \ldots, A_m\}$ and their deductive relations with the other propositions in the larger algebra Ω in which it resides. The obvious generalization of the theorem is given in Norton (forthcoming, §4.2)

10.3 How Might Deductive Definability Fail?

The requirement of deductive definability is fragile and easily broken. Since that might not be immediately apparent, here is an example of a failure. Consider the deductively definable logic of induction specified by (4) and (5) above. Replace (5) by

For all atoms, $a_1, a_2, \ldots, a_n$, (5')

$m(a_1) = m(a_2)/2 = \ldots = m(a_n)/n$

That is equivalent to setting the normalized measures of the atoms to

$$m(a_1) = 2/(n+n^2), \quad m(a_2) = 2\cdot 2/(n+n^2), \quad m(a_3) = 2\cdot 3/(n+n^2), \quad \ldots, \quad m(a_n) = 2\cdot n/(n+n^2)$$

The corresponding conditional probabilities are

$$P(a_1|\Omega) = 2/(n+n^2), \quad P(a_2|\Omega) = 2\cdot 2/(n+n^2),$$

$$P(a_3|\Omega) = 2\cdot 3/(n+n^2), \ \ldots, \ P(a_n|\Omega) = 2\cdot n/(n+n^2) \qquad (5'')$$

The key fact about these assignments is that they are non-uniform. That uniformity is unsustainable in a deductively definable logic of induction. Each of the atoms $a_1, a_2, \ldots, a_n$ enters into exactly the same deductive relations with the other propositions in the algebra. Hence deductive definability requires the equality of all these conditional probabilities

$$P(a_1|\Omega) = P(a_2|\Omega) = P(a_3|\Omega) = \ldots = P(a_n|\Omega).$$

For the condition (5') to be upheld, we must have some way of distinguishing among the atoms. Atom a_1 will be assigned the smallest measure m; atom a_2 will be assigned the next largest measure m; and so on.

Distinguishing among them cannot be done in terms of the deductive structure. It must be done by means external to the algebra. These means amount to external inductive content and lead to specification of the non-uniform probabilities (5'').

Finally, since the logic is no longer deductively definable, it is no longer possible to define the conditional probabilities of (5'') purely as a function of atoms counts, so the symmetry theorem does not apply to this logic.

11. The Need for Disjunctive Refinements

The example of the art thief shows how a simple deductively definable logic of induction can be inadequate for its intended purpose. We would like to know whether the evidence EC better supports that the art thief is in New York (NY), say, rather than in Boston, (BOS). However the logic requires [BOS|EC] = [NY|EC]. So differential support is not possible.

This problem will persist as long as the propositions form a small Boolean algebra based on just four atoms BOS, NY, PHL and PIT. The remedy is to increase the expressive power of the algebra by increasing the number of atoms. For example, we may judge that there are a large number of possible lairs in Boston in which our thief may hide. If we write BOS_i as the proposition that the thief is hiding in the *i*th of *r* possible lairs, then we create a disjunctive refinement of original algebra by replacing the atom BOS by the disjunction of new atoms

$$BOS = BOS_1 \text{ v } \ldots \text{ v } BOS_r$$

Correspondingly we can expand the remaining atoms as

$$NY = NY_1 \text{ v } \ldots \text{ v } NY_s$$

$$PHL = PHL_1 \text{ v } \ldots \text{ v } PHL_t$$

$$PIT = PIT_1 \text{ v } \ldots \text{ v } PIT_u$$

The small four-atom algebra has now been replaced by a larger algebra with $r+s+t+u$ atoms.

This larger algebra gives us a great deal more expressive power. We can assign widely varying support to propositions like BOS or NY, according to the values selected for r, s, t and u. In the probabilistic logic, we now have

$$P(\text{BOS}|\text{EC}) = r/(r+s+t) \qquad P(\text{NY}|\text{EC}) = s/(r+s+t)$$

If there are many more likely places to hide in New York than in Boston, we would have $r<s$ and P(BOS|EC) < P(NY|EC). For the specific conditioning logic, we now have

$$[\text{BOS}|\text{EC}]_{SC} = r/(r+s+t)$$

$$[(\text{BOS v PIT})|\text{EC}]_{SC} = r^2/[(r+t)(r+s+t)] = r/(r+t)\ [\text{BOS}|\text{EC}]_{SC}$$

Then $[(\text{BOS v PIT})|\text{EC}]_{SC}$ would be reduced in relation to $[\text{BOS}|\text{EC}]_{SC}$ according to how large t is in relation to r.

12. Asymptotic Stability

This last example illustrates a general property of deductively definable logics of induction. By disjunctively refining the atoms, we introduce new possibilities that alter the inductive strengths. Part of that content comes in the inductive relations among the new atoms and the original propositions. The part that will concern us here, however, involves just the relations among the old propositions.

Here is an example. We fix just three for examination: BOS, NY and EC and ask after the support BOS accrues from evidence EC and the support NY accrues from EC. As we refine and add more atoms, the relative strengths of support [BOS|EC] and [NY|EC] will change. Initially, these changes reflect the incorporation of new information into the algebra of propositions. There may be, for example, many more lairs in New York in which the art thief can hide.

In this process, we are not altering the evidence proposition directly. We are asking the same question repeatedly: what is the support accrued to NY from the evidence EC? What changes is the background deductive and inductive structure in which the propositions NY and EC appears. Those changes should be reflected, to greater or lesser degree, in the strength [NY|EC].

Eventually, we expect that the new information incorporated will have diminishing import

inductively. If NY_1 happens to be the proposition that the art thief is in a luxurious Fifth Avenue penthouse apartment in New York, then we might refine it further as

$$NY_1 = NY_{1\text{-NE}} \vee NY_{1\text{-NW}} \vee NY_{1\text{-SE}} \vee NY_{1\text{-SW}}$$

where $NY_{1\text{-NE}}$ is the proposition that the art thief is, at this moment, in the North-East corner of penthouse; and so on for the remaining three quadrants NW, SE and SW. Presumably this refinement would lead at best to a small change in the inductive strength [NY|EC].

Or perhaps not. Perhaps there is some evidential import in the location of the art thief in the penthouse that the inductive logic can discern. Then we might refine further to incorporate still more inductively relevant information. Through the refinements, we may add new sorts of propositions, perhaps concerning the history of the art thief's behavior, the climate in New York and elsewhere, the public transport system in various cities, and so on.

The requirement of asymptotic stability is that, eventually, continuing refinement will produce diminishing returns, in the sense that the original strengths like [NY|EC] alter less and less. Once we are at this point, strengths involving these propositions stabilize. They many stop changing completely. Or they may approach their limiting values asymptotically. For example, if [NY|EC] has the limiting value $[NY|EC]_{lim}$, then, once we are at this point of diminishing returns, the actual value of [NY|EC] will be close to $[NY|EC]_{lim}$ and the sole change introduced by further refinement is to bring [NY|EC] closer to the limiting value, $[NY|EC]_{lim}$.[175]

The idea behind asymptotic stability is just that there is a right choice for the strength of support [NY|EC] once all relevant background information is incorporated into the algebra; and that the inductive logic implemented is able to find it, at least asymptotically.

The alternative is to allow that the strength [NY|EC] never stabilizes. That would mean that, no matter how much additional information we incorporate into the algebra of propositions, the value of [NY|EC] would keep changing without ever settling down. An inductive logic that behaves this way is of no value to us, for it is unable to implement the idea that there is a definite strength of support that EC affords NY in the context of even the fullest specification of background facts.

This discussion so far has dealt with a special case of the art thief. The general case is no different. As indicated above, we concern ourselves with some fixed set of proposition $\{A_1, A_2,$

[175] More precisely, when we require that [NY|EC] approaches the limiting value $[NY|EC]_{lim}$ asymptotically we just mean this. Pick any measure of closeness to $[NY|EC]_{lim}$ you like: within 1%, within 0.1%, within 0.001%, etc. Then it is always possible to refine the algebra so that the actually value of [NY|EC] lies within those bounds and so that it remains there under all possible, subsequent refinements.

… , A_m}, where that set is very large and may include all the propositions considered in science. The requirement of asymptotic stability is that sufficient disjunctive refinement of the atoms leads each of the pairwise strengths [A_i| A_k] to settle down asymptotically to its limiting value, from which still further refinement cannot remove it. The limiting value is the best representation of the inductive support A_k affords A_i.

13. The Two Requirements Conflict

Now the trouble starts. We require two things of our logic of induction, each well motivated. First, we require it to be deductively definable, as a consequence of our requirement that the logic be complete. Second, we require asymptotic stability, as a consequence of our requirement that the logic can eventually lead to stable inductive strengths under continued disjunctive refinements.

The two requirements conflict and visit disaster on the logic. That is, if the logic is deductively definable, then it must be so responsive to different disjunctive refinements that it never settles down to limiting inductive strengths. Asymptotic stability proves unsustainable.

The instability is easily recoverable in the example of the art thief. Imagine that the art thief has a confederate within the police headquarters who is intent on confounding the police's efforts. The confederate can confound any inductive logic merely by artful selection of disjunctive refinements.

The ease of this confounding follows directly from the symmetry theorem: inductive strengths are fixed by the atom counts in the propositions. The confederate can then confound the logic merely by refinements that artfully manipulate the atom counts and drive the inductive support in any direction the malicious confederate desires.

For example, take the probabilistic case above. We have

$$P(\text{BOS}|\text{EC}) = r/(r+s+t) \qquad P(\text{NY}|\text{EC}) = s/(r+s+t)$$

We might start with values $r = s = t = 10$, as result of the first refinement. Then we would have

$$P(\text{BOS}|\text{EC}) = 1/3 \qquad P(\text{NY}|\text{EC}) = 1/3$$

The confederate might choose to lead the police towards Boston by merely refining BOS much more than NY and PHL. So we might refine further to $r = 1000$ and $s = t = 10$. Then evidential support swings strongly towards BOS since we have

$$P(\text{BOS}|\text{EC}) = 1000/1020 = 0.9804 \qquad P(\text{NY}|\text{EC}) = 10/1020 = 0.0098$$

But had the confederate chosen instead to refine NY, we could get exactly the reversed result, from $r = t = 10$ and $s = 1000$:

$$P(\text{BOS}|\text{EC}) = 10/1020 = 0.0098 \qquad P(\text{NY}|\text{EC}) = 1000/1020 = 0.9804$$

No matter how far advanced the disjunctive refinements may be, this possibility for confounding by further, malicious refinement will always be there. There can be no stabilization of the two probabilities. For, if ever the probabilities seem to stabilize, further malicious refinement can drive them away from what appeared to be their limiting values. The logic has no protection from this malice. Nothing within it can distinguish a refinement that reflects proper inductive import from one that merely deceives.

One might imagine the following escape. The malicious refinements are blocked merely by halting the disjunctive refinements at a stage at which further refinements would only advance the deception. This escape would succeed, but its success would depend upon knowing when is the appropriate stage of refinement at which to halt. That fact is not recoverable within the propositions of the algebra. It must be supplied by external considerations. These external considerations would then be supplying important inductive content in violation of the requirement of completeness of the inductive logic. That is, we escape instability by admitting incompleteness.

The example above is drawn from a probabilistic logic of induction. The same malicious deception can be visited upon *any* non-trivial logic of induction. The symmetry theorem tells us that the strengths in any deductively definable logic of induction are fixed by the atom counts. As long as the logic assigns different inductive strengths when the atom counts change, a malicious confederate will always be able to steer the weight of inductive support in any desired direction.

14. Triviality of a Complete Logic of Induction

The escape that preserves completeness is an unhappy one: if the logic of induction fails to adjust its strengths of inductive support when the atom counts change, then it is immune to deception by malicious disjunctive refinements. However a logic that is unresponsive to the atom counts, or merely unresponsive in its limiting behavior, is a trivial logic that assigns the same limiting inductive strength in all cases, no matter what the atoms counts in the propositions might be.

In short, deductive definability and asymptotic stability forces the inductive logic to be the trivial logic that assigns the same limiting value to all inductive strengths. The discussion here does not provide a proof of this result. It merely recounts an example to illustrate how the result comes about. The full demonstration of Norton (forthcoming), its "no-go" result, requires a great deal more logical accountancy. But those details introduce no further matters of principle. The essential manipulations have already been illustrated in the example above.

There is a technical complication in the full demonstration. To get the simplest version of the no-go result, a third condition of continuity is needed. It merely requires that inductive strengths do not make discontinuous jumps in their dependence on atom counts, when the atom counts are large. Without it, one still has triviality forced on the inductive logics, but the triviality comes in the form of a unique limiting value for each inductive strength, according to the class of deductive structure to which they belong. The notion of class is defined in Norton (forthcoming).

15. Escapes

The no-go result is developed in a precise setting: the deductive structure is given by propositional logic with finitely many propositions; and the inductive structure is given by inductive strengths that are represented by the binary quantity $[A|B]$. The temptation is to look for ways of escaping the result by altering the setting. The prospects of such an escape are poor.

As far the deductive structure is concerned, the logic employs just the Boolean operators, They reappear in most, more developed deductive logics. All these logics will then admit the disjunctive refinements that power the present analysis. More generally, the decisive property of the deductive structure is that it is highly symmetric. This symmetry can be replicated in richer logics. For example, if we have a simple predicate logic with monadic predicates only, $P_1(.), \ldots, P_n(.)$, then the logic will be symmetric under permutation of the predicates.

Similarly, a richer inductive structure will also generate corresponding no-go results. For example, we may replace $[A|B]$ by a tertiary quantity, "$[A|B,C]$," as suggested earlier. It could be interpreted as the strength of inductive support afforded proposition A by B with respect to background C. The discussion above would remain largely unchanged except in the details. If the inductive logic is deductively definable, the strength of support would still turn out to be a function solely of the atom counts in propositions A, B and C. As a result it would be subject to confounding by malicious disjunctive refinement, as before, and the logic would be forced to triviality.

More briefly stated, the no-go result developed here is likely to be replicable in almost any setting precisely because there is rather little in it. Deductive structures are, generally, highly symmetric; and asymptotic stability is hard to deny, for otherwise the inductive logic would fail to assign a stable limiting value for the strengths of inductive support. With these properties pervasive, a version of the incompleteness result is always nearby.

16. Subjective Bayesianism

Because of the present popularity of subjective Bayesianism, it is worth indicating how it interacts with the no-go result. To begin, the fact that prior probabilities can be assigned arbitrarily, according to our personal whim, does break the symmetry essential to the no-go result. However it breaks it at great cost, for the conditional probabilities cease to be measures of inductive support. They have become, initially, pure statements of opinion and, after conditionalization on evidence, an amalgam of opinion and evidential warrant. [176]

One might hope that the amalgam of opinion and warrant can be separated into its elements by a confirmation measure. It would be defined in terms of the subjective Bayesians' probabilities but would extract just the evidential warrant from the amalgam, stripping out any subjective contributions. What the no-go result asserts, however, is that any such confirmation measure must be trivial, if it is to be complete. For such a measure would conform to the conditions that lead to the no-go result.

17. The Recalcitrance of Problems of Induction Explained

This analysis establishes that any non-trivial calculus of inductive inference is incomplete. In retrospect that fact is not so surprising. The literature on calculi of inductive inference has been beset with persistent problems. We can now see that their recalcitrance is explicable as an inevitable outcome of incompleteness.

The traditional failure is the notorious problem of the priors in Bayesian analysis. The hope has been that we can push our inductive investigations back far enough to a neutral starting point, prior to the inclusion of any relevant evidence. There we seek a prior probability distribution that is vacuous in the sense that it inductively favors no particular proposition over any others. Yet no such vacuous prior has been found. All prior probability distributions exert an influence on the subsequent analysis and can only be used responsibly if they reflect the presence of further evidence outside the calculation.

That is just what incompleteness predicts. For a vacuous prior would enable a calculus to be complete. Moreover the incompleteness result predicts that this problem of the priors will reappear in some form in any non-trivial calculus, not just a probabilistic calculus.

---

[176] The celebrated "washing out of the priors" theorems fall short of what is needed. There is a reverse, indelibility result. Loosely speaking, for any fixed likelihoods and any fixed posterior probability we may choose, there will always be some perversely chosen prior probability compatible with it.

Another recurring problem is that the unadulterated probability calculus is not elastic enough to accommodate all inductive inference problems. There have been many extensions proposed. We may suppose, for example, that a simple probability measure is insufficient and it is replaced by a set of measures; or by a structure that uses interval values; and so on. Or we may alter the calculus in fundamental ways, such as the violation of additivity in the Shafer-Dempster calculus. Whatever successes these expansions meet, they are always limited. Further problems arise and call for still more extensions.

If we reconceive these proposals for altered calculi as efforts to find the one, true and complete logic of inductive inference, then their limited success ceases to be an unexpected annoyance. It is merely the reflection of a necessity: there can be no non-trivial, complete logic of inductive inference.

18. Conclusion

In the light of these results, what should we think of calculi of inductive inference? The import of the results is limited. They do not tell us that we must give up the idea of calculi of inductive inference. Rather they tell us that we should give up the quest for a single, all-purpose calculus that will give us a complete treatment of inductive inference. In its place, we should conceive of inductive inference locally. In any domain of investigation, no matter how big or how small, we may seek a calculus to govern our inductive inferences. If we find one that works in the domain, that calculus will never provide a complete account of the inductive relations in that domain. We will always need further inductive content to be supplied externally to the domain. No matter what our domain, there will always be an external background to which we must resort for inductive content.

This local reconception of inductive inference fits well with the material theory of induction developed in this book. In each domain, there will be relations of inductive support peculiar to it. They are not warranted by conformity with some universal calculus. They are warranted by the particular background facts prevailing in that domain. If those relations are regular enough to be described abstractly, we may identify a calculus for those inductive relations. However whether there is such a calculus and what its rules are will depend on the background facts prevailing in that domain. We should expect the calculus to differ from domain to domain. There is no universal calculus of inductive inference. That is the final moral of incompleteness.

References

Jeffreys, Harold (1961) *Theory of Probability*. 3rd ed. Oxford: Clarendon Press.

Kolmogorov, A. N., (1950), *Foundations of Probability*. Trans N. Morrison. New York: Chelsea Publishing Company.

Norton, John D. (2003) "A Material Theory of Induction" *Philosophy of Science*, **70**, pp. 647-70.

Norton, John D (2008) "Ignorance and Indifference." *Philosophy of Science*, **75**, pp. 45-68.

Norton, John D (2010) "Deductively Definable Logics of Induction." *Journal of Philosophical Logic*. **39**, pp. 617-654.

Norton, John D. (forthcoming) "A Demonstration of the Incompleteness of Calculi of Inductive Inference." *British Journal for the Philosophy of Science*.

Solomonoff, Ray (1964). "A Formal Theory of Inductive Inference," *Information and Control* **7** pp. 1–22; pp. 224-254.

November 17, December 3, 2016;
January 3, 6, March 12, 22, 2017. July 18, 2017; November 8, December 8, 2017.
April 27, June 10, 22, 2018
October 20, November 8, 2018

Field that initiates equation numbering:

Chapter 13

Infinite Lottery Machines

1. Introduction

No single calculus of inductive inference can serve universally. There is even no guarantee that the inductive inferences warranted locally, in some domain, will be regular enough to admit the abstractions that form a calculus. However, in many important cases, when the background facts there warrant it, inductive inferences can be governed by a calculus. By far the most familiar case is the probability calculus.

That many alternative calculi other than the probability calculus are possible is easy to see. Norton (2010) identifies a large class of what are there called "deductively definable" logics of induction. Generating a calculus in the class is easy. It requires little more than picking a function from infinitely many choices.

The harder part is to see whether some specific calculus is warranted in some particular domain. This and the following chapters will provide a few illustrations of unfamiliar cases. In them, the warranted calculus is not the probability calculus. The systems to be investigated are: in this chapter, infinite lottery machines; and, in subsequent chapters, continuum-sized outcome sets, which include nonmeasurable outcomes; indeterministic physical systems; and the quantum spin of electrons.

The focus of this chapter, a fair infinite lottery machine, selects among a countable infinity of outcomes, 1, 2, 3, … without favor. It allows us to pose a series of inductive problems. In this arrangement, how much support inductively is given to the outcome of some particular number, say 378? Or to some finite set of numbers, say all those between 37 to 256? Or to some infinite set of numbers, such as the even numbers or the prime numbers? The answers to these questions will be supplied by the inductive logic applicable to this domain.

The warranting facts that pick out the logic will be the physical properties of the infinite lottery machine. The inductive logic will be the same for all properly functioning infinite lottery

machines. Thus the pertinent warranting facts will be just those that they have in common. That is the fact that they choose a number without favoring any.

The example of the infinite lottery machine has already proven troublesome. We shall see in Section 2 that an unreflective application of the probability calculus to it fails. The literature has explored several ways of modifying the calculus to accommodate the infinite lottery. They include dropping countable additivity and introducing infinitesimal probabilities. In subsequent sections, I will argue that neither of these modifications succeeds. The defining characteristic of the infinite lottery is that it chooses its outcomes without favoring any one. That characteristic is captured formally in the condition of "label independence" of Section 3. It says that the chance of an outcome with some definite number or a set of them is unaffected if we permute the numbers that label the outcomes. This condition, it is argued in Sections 4 and 5, is incompatible with the (finite) additivity of a probability measure. This additivity is the familiar property that, if we have two mutually exclusive outcomes, then we can add their probabilities to find the probability of their disjunction. Thus the chance properties of an infinite lottery machine cannot be represented by a probability measure. Attempts to continue to do so, it is argued in Section 6, amount to altering the background facts presumed. These attempts do not solve the problem but merely exchange it for a different problem that can be solved with a probability measure. Section 7 explores a non-standard calculus that is warranted by specific configurations of an infinite lottery machine. Section 8 outlines how we can give intuitive meaning to the values in the non-standard calculus and use it to make predictions. Section 9 extends the logic to repeated, independent drawings of the lottery. Section 10 uses the extension to show that the chances of frequencies of outcomes in these repeated drawings do not conform with probabilistic expectations so that frequencies cannot be used to reintroduce probabilities. Section 11 defends the failure of what is there identified as the "containment principle." Section 12 reports briefly on work elsewhere on the unexpected complications found when we try to determine the extent to which an infinite lottery machine is physically possible. Section 13 concludes.

Finally, Appendix A reviews the so-called "measure problem" of eternal inflation in modern cosmology. It turns out to be essentially the same as the difficulty of fitting an additive probability measure to an infinite lottery machine.

2. The Initial Difficulty

The infinite lottery machine entered the literature because it poses an immediate problem if we wish to use the probability calculus as the applicable inductive logic. That problem arises from a tension between two conditions. First, the machine chooses each number without favor. So each outcome n must have equal probability $P(n)$:

$$\varepsilon = P(1) = P(2) = \ldots = P(n) = \ldots \quad (1)$$

Second, the outcomes are mutually exclusive and at least one must happen. Hence all these probabilities must sum to unity in the infinite sum:

$$P(1) + P(2) + \ldots + P(n) + \ldots = 1 \quad (2)$$

No value of ε can satisfy both (1) and (2). For if we choose some $\varepsilon > 0$, no matter how close this ε to zero, then (2) is the summing of infinitely many non-zero ε's. Summing only finitely many will eventually exceed the unity required in (2). If, instead, we set $\varepsilon = 0$, then (2) is the summing of infinitely many zeroes, which is zero.

Two types of solutions have been proposed in the literature. The most popular, advocated by Bruno de Finetti (1972; §5.17), targets the fact that (2) requires the summing of an infinity of probabilities. This infinite sum operation is qualitatively different from merely summing finitely many probabilities. For the infinite summation is carried out in two steps. First, one sums finitely many terms, up to some large number N, say:

$$S(N) = P(1) + P(2) + \ldots + P(N)$$

One then takes the limit of $S(N)$ as N grows infinitely large. De Finetti proposed that we discard this rule of "countable additivity"[177] and employ only the first step, "finite additivity," in which we are allowed to add only finitely many probabilities. The outcome is that we no longer require summation condition (2) for the infinite lottery machine; and we can now employ $\varepsilon = 0$ in (1), without running into contradictions. De Finetti's proposal has been subject to extensive critical scrutiny. See, for example, Bartha (2004), Blackwell and Diaconis (1996), Kadane, Schervish, and Seidenfeld (1986), Kadane and O'Hagan (1995) and Williamson (1999).

Setting $\varepsilon = 0$ amounts to setting the probability of each individual number outcome (or any finite set of them) to zero. That seems too severe to some. Might we not manage by assigning a very, very tiny probability—an "infinitesimal" amount—to each outcome? Non-standard analysis provides a mathematically clean way of doing just this. The possibility has been explored by, for example, Benci, Horsten, and Wenmackers (2013) and Wenmackers and Horsten, (2013); and it has been subjected to critical scrutiny by, for example, Pruss (2014), Williamson (2007) and Weintraub (2008).

---

[177] The full condition of countable additivity applies to any infinite set of mutually incompatible outcomes $\{A_1, A_2, \ldots, A_n, \ldots\}$ and asserts that $P(A_1 \text{ or } A_2 \text{ or } \ldots) = P(A_1) + P(A_2) + \ldots$, where the ellipses "…" indicate that the formulae continue for all n.

Neither of these repairs to probabilistic analysis will be pursued further here since, as I will now argue, no such repair is adequate. The infinite lottery requires an even greater departure from normal ideas of probability.

3. Label Independence

To proceed, we must clarify just what is meant by "choosing *without favor*" or, as it is sometime said, having a "fair" lottery. Taking this to mean that each outcome has equal probability is untenable since it presumes that the probabilistic treatment is adequate. We need an analysis that does not make this presumption. In the following, I shall speak of the "chance" of an outcome, where the term will no longer designate a probability. Just what it designates will be determined through the development of the inductive calculus that governs it, in the sections that follow.

What it is to choose without favoring any outcome can be specified through the requirement of "label independence." The driving intuition is that, when outcomes are chosen *with* favor, then the chances will, in general, differ with different outcomes. Holding a ticket for the outcome labeled "37" may be preferable to, say, "18," if the outcome labeled "37" is favored over the one labeled "18." If, however, the choice is made *without* favor, then we should be indifferent to whether we have the outcome labeled "37," "18" or any other label. Moreover, that indifference should remain no matter how the lottery machine operator switches the labels around over the various outcomes. We should not care to which outcome our label "37" is attached, for none is favored.

The general requirement is that the chances are unaffected by any permutation of the labels. A permutation moves labels from outcomes to outcomes such that every outcome starts and ends with exactly one a label; no labels are discarded; and no new labels are introduced. More formally, the requirement is:

Label independence

> All true statements pertinent to the chances of different outcomes remain true when the labels are arbitrarily permuted.

We can see how it works by taking the case of a finite randomizer, the roulette wheel. Such a wheel has, in the American case, 38 equally sized pockets on its perimeter. It is spun and a ball projected in the opposite direction. The pockets are numbered from 1 to 36, 0 and 00; and the outcome is the pocket in which the ball eventually comes to rest. As long as the wheel is well balanced with equal sized pockets and the croupier spins and projects with vigor, the ball with pass over the wheel many times and arrive with equal chance in each pocket. Under those conditions, the choice of labeling of the pockets is immaterial. We could, without compromising

the fairness of the wheel, peel off the labels that mark each pocket and rearrange them in any way we please.

To apply label independence, we start with a statement true of a properly made roulette wheel:

Pockets 11 and 23 are the same size.

Under a permutation that switches label 11 with label 3 and label 23 with label 10, the proposition now asserts a truth expressed in the old labeling as:

Pockets 3 and 10 are the same size.

Proceeding with further permutations, we see that the label independence of the statement amounts to the assertion that any two pockets have the same size. Similarly the following is true of any well functioning roulette wheel:

The ball ends up in pockets 1 to 12,
roughly as often as it does in pockets 13 to 24.

Under label independence, it remains true if we permute the labels of pockets 13 to 24 with those of pockets 25 to 36. It now expresses a truth expressed in the old labeling as

The ball ends up in pockets 1 to 12,
roughly as often as it does in pockets 25 to 36.

Thus the label independence of the second statement reflects the fact that the relative frequency of outcomes in a set of pockets depends merely on the number of pockets in the set.

The qualification "pertinent to the chances" is essential, for there are many statements true of a roulette wheel whose truth is not preserved under arbitrary permutation of the pocket labels. For example, in an American wheel:

Pockets 3 and 4 are diametrically opposite on the wheel.

This statement does not remain true under most permutations of the pocket labels. However, since the statement is not pertinent to the randomizing function of the wheel, the failure does not violate label independence.

4. Abandoning Finite Additivity

There are no surprises when label independence is used to characterize how a finite randomizer, such as a roulette wheel, picks outcomes without favor. Matters change when label independence is applied to an infinite lottery machine. The reason is that labels on infinite sets of outcomes can be permuted in ways that are impossible for finite sets. It is easy to permute them so that the labels for some infinite set of outcomes end up assigned to one of its proper subset. It follows from label independence that the set and its proper subset have the same chance. If chances are probabilities, that means that they have the same probability. Assembling several

permutations like this soon contradicts the requirement that the probability of an outcome is the sum of the probabilities of its disjoint parts. That is a striking result that bears being repeated. If outcome A is the disjunction of mutually exclusive outcomes B or C or D, that is,

$$A = (B \text{ or } C \text{ or } D),$$

and B, C and D pairwise contradict, then we can have cases in which

$$\text{Chance}(A) = \text{Chance}(B) = \text{Chance}(C) = \text{Chance}(D) \quad (3)$$

which is incompatible[178] with finite additivity,[179] which requires

$$P(A) = P(B) + P(C) + P(D) \quad (4)$$

That is, the label independence of an infinite lottery machine requires us to abandon finite additivity for a measure of the chance of sets of outcomes. Since finite additivity is essential to the definition of probability, it follows that chances cannot be probabilities for an infinite lottery machine.

5. An Example of the Failure of Finite Additivity

An illustration of the failure of finite additivity in (3) and (4) is provided by an example reported in Bartha (2004, §5) and Norton (2011, pp. 412-15). Assume that the chance function "Ch(.)" measures the chance of the different sets of outcomes of an infinite lottery machine, recalling that the notion of chance employed here, so far, is only loosely defined and need not be a probability measure. For some numbering of the outcomes, the labels on the sets of even numbered outcomes[180]

$$\textit{even} = \{2, 4, 6, 8, \ldots\}$$

and on the sets of odd numbered outcomes

$$\textit{odd} = \{1, 3, 5, 7, \ldots\}$$

can be switched one-one by a permutation:

$$1 \leftrightarrow 2,\ 3 \leftrightarrow 4,\ 5 \leftrightarrow 6,\ 7 \leftrightarrow 8,\ \ldots$$

Hence, by label independence, the two sets must have equal chance:

$$\text{Ch}(\textit{even}) = \text{Ch}(\textit{odd}) \quad (5)$$

---

[178] Unless all the probabilities are zero.

[179] The full condition of finite additivity applies to any finite set of mutually incompatible outcomes $\{A_1, A_2, \ldots, A_n\}$ and asserts that $P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$.

[180] Here and henceforth I move without warning between a set representation of an outcome, *even* = {2, 4, 6, …} and an equivalent propositional representation, *even* = 2 or 4 or 6 or …

Now consider the four sets of every fourth number.

$$\textit{one} = \{1, 5, 9, 13, \ldots\}$$
$$\textit{two} = \{2, 6, 10, 14, \ldots\}$$
$$\textit{three} = \{3, 7, 11, 15, \ldots\}$$
$$\textit{four} = \{4, 8, 12, 16\}$$

By similar reasoning each of *one*, *two*, *three*, and *four* have equal chance:

$$\text{Ch}(\textit{one}) = \text{Ch}(\textit{two}) = \text{Ch}(\textit{three}) = \text{Ch}(\textit{four}) \qquad (6)$$

So far, nothing untoward has happened. All this is compatible with the Ch(.) function being a probability measure. This will now change.

Consider two sets of outcomes: *one* and the set whose members are in (*two* or *three* or *four*). Since all the sets are countably infinite, we can have the following two-part permutation of the labels. The first switches one to one the labels on *odd* with those on *one*:

$$1\leftrightarrow 1, 3\leftrightarrow 5, 5\leftrightarrow 9, 7\leftrightarrow 13, \ldots$$

The second part switches one to one the labels on *even* with those of (*two* or *three* or *four*):

$$2\leftrightarrow 2, 4\leftrightarrow 3, 6\leftrightarrow 4, 8\leftrightarrow 6, 10\leftrightarrow 7, 12\leftrightarrow 8, 14\leftrightarrow 10, 16\leftrightarrow 11, \ldots$$

For convenience, since the set *one* now carries the labels that originated in *odd,* let us also call it *odd**; and similarly (*two* or *three* or *four*) is also called *even**. That is, we have two names for each outcome set:

$$\textit{one} = \textit{odd}^* \quad (\textit{two} \text{ or } \textit{three} \text{ or } \textit{four}) = \textit{even}^*$$

Since the new labels of outcomes in *odd** and *even** can also be switched one-one with each other, analogously to (5), they must also have equal chance. That is:

$$\text{Ch}(\textit{even}^*) = \text{Ch}(\textit{odd}^*) \qquad (7)$$

Combining we have

$$\begin{aligned}
\text{Ch}(\textit{two}) = \text{Ch}(\textit{three}) &= \text{Ch}(\textit{four}) && \text{[from (6)]}\\
&= \text{Ch}(\textit{one}) && \text{[from (6)]}\\
&= \text{Ch}(\textit{odd}^*) && \text{[since } \textit{one} \text{ and } \textit{odd}^* \text{ name the same set]}\\
&= \text{Ch}(\textit{even}^*) && \text{[from (7)]}\\
&= \text{Ch}(\textit{two} \text{ or } \textit{three} \text{ or } \textit{four}) && \text{[since } (\textit{two} \text{ or } \textit{three} \text{ or } \textit{four}) \text{ and } \textit{even}^*\\
& && \text{name the same set]}
\end{aligned}$$

These last equalities violate[181] finite additivity (4), since a finitely additive probability measure P(.) must satisfy:

$$\text{P}(\textit{two}) + \text{P}(\textit{three}) + \text{P}(\textit{four}) = \text{P}(\textit{two} \text{ or } \textit{three} \text{ or } \textit{four})$$

[181] Unless all the probabilities are zero.

6. Finite Additivity Must Go

The simple example shows that label independence for an infinite lottery is incompatible with the finite additivity of a probability measure. To proceed, at least one of them must be given up. Both Bartha (2005, §5) and Wenmackers and Horsten (2013, p. 41) find giving up finite additivity too great a sacrifice. In my view, we have no choice but to sacrifice finite additivity. For label independence is a defining characteristic of an infinite lottery machine. Without it, we can no longer say that the infinite lottery machine chooses its outcomes without favoring any. There is no comparable necessity for probability measures, other than our comfort and familiarity with them.

To persist in describing the chance properties of an infinite lottery machine by a probability measure is, in effect, to change the problem posed. For no single probability measure can satisfy all the equalities derived above from label independence. We must choose which subset will be satisfied. That choice amounts to adding extra conditions on the operation of the infinite lottery machine. While the augmented problem may be quite well-posed and even interesting, it is a different problem. The extra conditions must breach label independence, so that we no longer describe a device that chooses outcomes without favor. We have not solved the original problem, but merely changed it to a different problem we like better.

To see how this favoring can come about, consider the two equalities (5) and (7). If the chance function is a probability function P(.), then they become

$$P(\textit{even}) = P(\textit{odd}) = 1/2 \qquad (5a)$$

$$P(\textit{even}^*) = P(\textit{odd}^*) = 1/2 \qquad (7a)$$

We cannot uphold both if we note that the probabilistic version of (6) requires

$$P(\textit{one}) = P(\textit{two}) = P(\textit{three}) = P(\textit{four}) = 1/4 \qquad (6a)$$

For then $P(\textit{odd}^*) = P(\textit{one}) = 1/4$; while $P(\textit{even}^*) = P(\textit{two}) + P(\textit{three}) + P(\textit{four}) = 3/4$, in contradiction with (7a).

To preserve the applicability of a probability measure, we have to block one of (5a) or (7a). A simple strategy is to select a preferred numbering of the outcomes, such as the original labeling, and then define the probability of each set of outcomes in the natural way. That is, we consider the sequence of finite, initial sets

$$\{1\}, \{1, 2\}, \{1, 2, 3\}, \ \ldots, \{1, 2, 3, \ldots, n\} \ , \ldots \qquad (8)$$

The probability of some nominated outcome set is defined as the limit of the frequency of outcome set members in this sequence. For the outcome *even*, we have

$$\begin{aligned} P(\textit{even}) &= \mathrm{Lim}_{n\to\infty}\, n/2n = 1/2 && n \text{ is even} \\ &= \mathrm{Lim}_{n\to\infty}\, (n+1)/2n = 1/2 && n \text{ is odd} \end{aligned} \qquad (9)$$

Definitions of the form (9) using the sequence (8) gives the expected probabilities (5a) and (6a) for P(*even*), P(*odd*), P(*one*), P(*two*), P(*three*) and P(*four*). However they fail to return (7a), since, as before, we have P(*odd**) = P(*one*) = 1/4 and P(*even**) = P(*two* or *three* or *four*) = 3/4.

There is a second, parallel "starred" analysis that preserves the equality of (7a) while giving up (5a). It proceeds exactly as above, but replaces the sequence (8) with one natural to the starred labeling of outcomes. That is, the starred labels assigned to outcomes after the permutation conform with

$$odd^* = \{1^*, 3^*, 5^*, 7^*, \ldots\} = \{1, 5, 9, 13, \ldots\}$$
$$even^* = \{2^*, 4^*, 6^*, 8^*, \ldots\} = \{2, 3, 4, 6, 7, 8, 10, 11, 12, \ldots.\}$$

In place of (8), it has the sequence:

$$\{1^*\} = \{1\}, \{1^*, 2^*\} = \{1, 2\}, \{1^*, 2^*, 3^*\} = \{1, 2, 5\}, \{1^*, 2^*, 3^*, 4^*\} = \{1, 2, 5, 3\}, \ldots \quad (8a)$$

Using the sequence (8a), definitions of probability based on relative frequencies akin to (9), will give starred results that are the reverse of the unstarred results. That is, we shall secure (7a) P(*even**) = P(*odd**) = 1/2, but not (5a).

In comparing the unstarred and starred analysis, we see how each improperly favors certain outcomes in the judgment of the other. The unstarred analysis gives P(*odd**) = 1/4 and P(*even**) = 3/4, improperly favoring *even** over *odd**, according to a starred analysis. However the starred analysis gives gives P(*odd*) = 1/4 and P(*even*) = 3/4, improperly favoring *even* over *odd*, according to an unstarred analysis.

Thus describing an infinite lottery machine with a probability measure replaces the original requirement of selection without favor, by selection under by the added restriction that the selection must respect also a preferred numbering scheme and the limiting ratios native to it.

That some such change in the problem is required if probabilities are to be retained was noted by Edwin Jaynes. He was a leading proponent of objective Bayesianism and a master of the memorable riposte, which he formulated for this case as follows (2003, p.xxii).

> Infinite-set paradoxing has become a morbid infection that is today spreading in a way that threatens the very life of probability theory, and it requires immediate surgical removal. In our system, after this surgery, such paradoxes are avoided automatically; they cannot arise from correct application of our basic rules, because those rules admit only finite sets and infinite sets that arise as well-defined and well-behaved limits of finite sets. The paradoxing was caused by (1) jumping directly into an infinite set without specifying any limiting process to define its properties; and then (2) asking questions whose answers depend on how the limit was approached.
>
> For example, the question: 'What is the probability that an integer is even?' can have any answer we please in (0, 1), depending on what limiting process is used to

define the 'set of all integers' (just as a conditionally convergent series can be made to converge to any number we please, depending on the order in which we arrange the terms).

In our view, an infinite set cannot be said to possess any 'existence' and mathematical properties at all – at least, in probability theory – until we have specified the limiting process that is to generate it from a finite set.

The bluster of Jaynes' riposte cannot cover the fact that he can offer no good reason for eschewing infinite sets that do not come with a preferred ordering or numbering scheme. If we must eschew all such sets, then we are precluding from inductive analysis cases that arise in real science. The problems just rehearsed in Sections 5 and 6 above have played out almost exactly as a foundational problem in recent inflationary cosmology, the "measure problem," where the lack of a preferred order on an infinite set of pocket universes has precluded introduction of a probability measure over them. The problem is reviewed in the Appendix. This should quell fears that that the problem of fitting a probability measure to an infinite lottery machine is merely the contrarian whimsy of eccentric theorists and idle philosophers. The problem has a connection and application in real science.

7. The Inductive Logic Warranted for an Infinite Lottery Machine

The defining characteristic of an infinite lottery machine is that its choice of outcomes respects label independence. That characteristic rules out an inductive logic whose strengths of support are probability measures. According to the material theory of induction, the background facts warrant the inductive logic appropriate to the domain. Label independence, the characteristic common to all infinite lottery machines, is the key, warranting fact. It acts powerfully and leads us to the following inductive logic.

7.1 Equal Chance Sets

The logic divides outcomes sets into types such that all sets of the same type must have the same chance. To implement this division, we require that two outcomes sets are of the same type if the members of the two sets can be mapped one-one to one another by a permutation of labels. That means that the outcome sets must have the same size (i.e. cardinality). In addition, the complements of the sets must also be the same size, else the requisite permutation of labels will not be possible. What results are sets of outcomes of the following types:[182]

[182] Co-infinite means that the complement of the set is infinite. Co-finite means that the complement of the set is finite.

$finite_n$: a set with n members, where n is a natural number.

Examples of $finite_3$ are {1, 2, 3}, {27, 1026, 5000} and {24, 589, 2001}.

$infinite_{\text{co-infinite}}$: an infinite set whose complement is also infinite.

An example is the infinite set of even numbers {2, 4, 6, …} since its complement is the infinite set of odd numbers {1, 3, 5, …}

$infinite_{\text{co-finite-}n}$: an infinite set whose complement is finite of size n.

An example of $infinite_{\text{co-finite-10}}$ is the set of all numbers greater than 10: {11, 12, 13,…) since its complement is the finite set {1, 2, 3, …, 10}.

7.2 Chance Values

The requirement of label independence entails that sets of outcomes of the same type must be assigned the same chance. Thus the chance function Ch(.) in this logic can only have the following set of values:

$$\text{Ch}(finite_n) = V_n, \text{ where } n = 1, 2, 3, \ldots \quad (10a)$$

$$\text{Ch}(infinite_{\text{co-infinite}}) = V_\infty = \text{“as likely as not.”} \quad (10b)$$

$$\text{Ch}(infinite_{\text{co-finite-n}}) = V_{-n}, \text{ where } n = 1, 2, 3, \ldots \quad (10c)$$

And for completeness we add in the two special cases

$$\text{Ch}(\textit{empty-set}) = V_0 = \text{“certain not to happen”} \quad (10d)$$

$$\text{Ch}(\textit{all-outcomes}) = V_{-0} = \text{“certain to happen”} \quad (10e)$$

According to (10a), all equal-sized finite sets of outcomes have the same chance: any n membered finite set has the same chance V_n. This is required by label independence since some permutation can always switch the labels between any two finite sets, as long as they are the same size. Similarly, (10b) tells us that all infinite sets that are co-infinite have the same chance. We have already seen an example above in (5) and (7):

$$\text{Ch}(\textit{even}) = \text{Ch}(\textit{odd}) = \text{Ch}(\textit{even*}) = \text{Ch}(\textit{odd*}) = V_\infty$$

Since each of the four infinite sets are co-infinite, there is a permutation that switches their labels. By label independence, they have the same chance. Since every co-infinite infinite set of outcomes is assigned the same value V_∞ as its complement set, we informally name this value “as likely as not.” Finally, (10c) can be interpreted similarly to (10a).

7.3 Comparing Chance Values

The conditions (10) are powerful restrictions. They preclude the chance function Ch(.) being an additive probability measure. However they leave the logic underspecified. We do not yet know whether the values V_n, V_∞, V_{-n} are the same or different; and, if they are different,

how they compare with one another. To arrive at the conditions (10), we used label invariance only. Further restrictions can enrich the logic.

A qualitative ranking of the strengths of support derives from the idea that the chance of a set of outcomes cannot be diminished if we add further outcomes to the set. This condition induces the relation "≤, " which is read as "is no stronger than." It obtains between values A and B when the outcomes that realize a value A can be a subset of the outcomes that realize a value B. As a result, the relation inherits the properties of set theoretic inclusion. It is antisymmetric, reflexive and transitive. It is easy to see that:

$$V_0 \leq V_1 \leq V_2 \leq V_3 \leq \ldots \leq V_\infty \leq \ldots \leq V_{-3} \leq V_{-2} \leq V_{-1} \leq V_{-0} \qquad (11)$$

One might think this condition unavoidable. It is not. It is merely familiar and amounts to one construal of the meaning of strength of support. A somewhat similar condition fails in the "specific conditioning logic" of Norton (2010, §11.2).

Further discriminations, if they happen at all, must be warranted by further background facts, whose truth must be recovered from the physical properties of the pertinent chance process. One case that is easy to motivate physically arises if we have an additive measure that is not normalizable. That is, the total measure of its space is infinite. It arises if we have a space in which lengths, areas or volumes are defined, the total space has infinite length, area or volume and the chances of some event occurring in a region of the space are measured by its length, area or volume. This case is developed more fully in the next chapter on "Uncountable Problems" in Section 4. An illustration recounted there derives from steady state cosmology. According to it, the chance of a hydrogen atom being created in some region of our cosmic infinite Euclidean space is proportional to the region's volume.

To apply the infinite lottery logic this case, we divide the space into infinitely many parts of equal length, area or volume. An outcome finite_n arises when the event is realized in some subset of the space of n of these parts. Its chance is measured by n. Correspondingly, the chance associated with any infinite volume of space will be measured by ∞. That is, we have:

$$\mathrm{Ch}(\mathit{finite}_n) = V_n = n \qquad \text{where } n = 1, 2, 3, \ldots \qquad (12)$$

$$\mathrm{Ch}(\mathit{infinite}_{\text{co-infinite}}) = V_\infty = \mathrm{Ch}(\mathit{infinite}_{\text{co-finite-n}}) = V_{-n} = \infty$$

The inequalities relating the various values of V_n in (11) become strict inequalities.

$$V_0 < V_1 < V_2 < V_3 < \ldots < V_\infty \qquad (11a)$$

If the outcome of the infinite lottery machine lies in some finite set of outcomes, then the chance relations (12) match those of a finite probabilistic randomizer with the same finite set of outcomes. That is, the chances of different outcomes in the finite set will behave like probabilities defined as:

$$P(A|B) = \mathrm{Ch}(A)/\mathrm{Ch}(B) \qquad (13)$$

where A is a subset of B and B is a finite set of outcomes.

The conditions (11a) and (13) are not assured. They can fail, depending on the particular physical instantiation of the infinite lottery machine. Such a failure would arise if the randomizer is based on the non-probabilistic, indeterministic systems described in Chapter 15 below. The conditions succeed for the "Spin of a pointer on a dial" device of Norton (2018).

Correspondingly, while label independence does not force it, we may require as an additional assumption in some more specific logic that:[183]

$$V_{\infty} < \ldots < V_{-3} < V_{-2} < V_{-1} < V_{-0} \qquad (11b)$$

In the following section, we shall see why this additional assumption fits naturally into the formal properties of the chance function.

These inequalities along with relations (10), (11), (12) and (13), all assumed henceforth, characterize an inductive logic native to an infinite lottery machine well enough for us to see that such logics differ significantly from a probabilistic logic.

A curious outcome of the analysis is that this logic is the reverse of the one de Finetti (1972; §5.17) proposed for an infinite lottery. In his logic, additivity was preserved for outcomes comprised of infinite sets; but it was trivialized for outcomes of finite sets, since these latter were all assigned zero probability. In the present logic, non-trivial additivity is maintained for finite sets through (12) and (13), but additivity fails through (10b) for most infinite sets.

8. Interpreting the Inductive Logic

The chance function Ch(.) of Section 7 specifies an inductive logic. Its formal properties are clear. However we may well ask what its quantities mean. What should we think when we learn that some outcome has such and such a chance value? This question is asking less than is usually asked, in the analogous circumstance, when we seek an interpretation of probability. It is not asking for an explicit definition, such as is sought by a relative frequency interpretation of probability or from the subjectivist Bayesian definition of probability in terms of betting quotients. One can have an understanding of a magnitude, adequate for practical applications, without an explicit definition of it. Since the values of the chance function (10) are so unfamiliar, that is all that is sought here.

---

[183] Considerations of cardinality make natural the strict inequality $V_{\infty} < V_{-n}$ for all n. However, unlike the case of V_n, I have been unable to conceive possible background facts that would warrant strict inequalities among the individual values of V_{-n} as shown in (11b). Might an inventive reader be able to conceive such facts?

8.1 The Probabilistic Model

The problem of developing some informal understanding of an initially abstruse quantity arises also for ordinary probabilities. We can use its solution as a model for the new chance function. Take the simple case of a coin toss, whose outcomes can be heads *H* or tails *T*. How are we to understand the probability assertion that $P(H) = 0.5$? How are we to distinguish that probability assertion from nearby assertions like $P(H) = 0.4$ or $P(H) = 0.6$? To be told that a probability of 0.4 is weaker than a probability of 0.5 is true but merely qualitative and falls well short of the precision we expect.

We gain a better understanding of such assertions, sufficient to discriminate among them, by contriving associated circumstances of either very high or very low probability. For example:

> If $P(H) = 0.5$, then, with probability near one, the frequency of *H* among many, independent coin tosses will be close to 0.5.
>
> If $P(H) = 0.4$, then, with probability near one, the frequency of *H* among many, independent coin tosses will be close to 0.4.

Sentence like these, by themselves, are not sufficient to give informal meaning to the quantity $P(.)$. All we have is one probability statement, that $P(H) = 0.5$, associated with another statement concerning an outcome with probability near one. Without something further, we will be trapped forever in a self-referential web of statements in which probabilistic assertions are made about other probabilistic assertions, without otherwise clarifying what any probabilistic assertion means. The axioms and definitions used to deduce all these assertions can be modeled in many systems with an extensive quantity whose magnitude is additive. To break out of the self-referring trap, we use a rule that coordinates large and small values of probability with informal judgments of expectation about chancy outcomes:

> *Rule of coordination for probability.*
>
> Very low probability outcomes generally do not happen; and very high probability outcomes generally do.

Thus we come to some understanding of the difference between $P(H) = 0.5$ and $P(H) = 0.4$: we expect each to deliver roughly 50% or 40% *H* respectively in repeated, independent coin tosses.

This interpretive rule, in various forms, has a long history and has come to be known as "Cournot's Principle."[184] In his canonical treatment of the foundations of probability theory,

---

[184] For a brief survey, see Shafer (2008, §2). One must be careful to treat the rule as nothing more than an informal guide. Otherwise the danger is that one misidentifies very low probability events as strictly impossible and very high probability events as necessary. For de Finetti's view of the rule, see de Finetti (1974, pp. 180-181). My use of the term "rule of coordination" is intended to recall Reichenbach's notion of a coordinative principle.

Kolmogorov (1950, p. 4) has a version of this rule that employs the locution "practically certain":

> (a) One can be practically certain that if the complex of conditions $\mathfrak{S}$ [Fraktur capital S] is repeated a large number of times, n, then if m be the number of occurrences of event A, the ratio m/n will differ slightly from $P(A)$.
>
> (b) If $P(A)$ is very small, one can be practically certain that when conditions $\mathfrak{S}$ are realized only once, the event A would not occur at all.

This process of conveying meaning should not be confused with subjective Bayesians' process of elicitation of probabilities. They determine, for example, that a subject has assigned probability 0.5 to H when the subject accepts even odds on either H or T. The present concern is how the subject, prior to the elicitation, came to judge that 0.5 is the appropriate probability to assign. That in turn requires some prior understanding by the subject of what probability 0.5 means.

8.2 The Analogous Analysis for the Chance Function

This same strategy can be used both to interpret the values of the chance function (10) and, at the same time, to display the predictive powers of the logic. The analogs of very low probability and very high probability outcomes are those with chance V_n and chance V_{-n}. A chance V_n outcome is realized when the number drawn resides in a finite set among the infinitely many possibilities. This is not an outcome we should expect to happen since it is thoroughly swamped by the infinitely many numbers outside the set. A chance V_{-n} happens when the number drawn resides outside some finite set. Since there are infinitely many possibilities outside the finite set that realize it, this is an outcome we should expect. That is, we have the interpretive rule:

> *Rule of coordination for chance.*
>
> Very low chance outcomes with chance V_n generally do not happen; and very high chance outcomes with chance V_{-n} generally do.

This rule divides outcomes sharply into three sets:

> outcomes in one of the $finite_n$, which we do not expect;
>
> outcomes in $infinite_{\text{co-infinite}}$, which may or may not happen "as likely as not"; and
>
> outcomes in one of the $infinite_{\text{co-finite-}n}$, which we do expect.

The application of this rule is simpler than in the probabilistic case for two reasons. First, in the present case, the division of outcomes into unexpected, intermediate and expected is sharp. This sharpness makes it natural to replace the inequalities of (11) by strict inequalities. In the probabilistic case, the division was muddier. Just how low should a probability be before its

outcome is not to be expected? If one is pressed, one eventually introduces some arbitrary cutoff, knowing that any cutoff can be challenged if sufficient contrivance is allowed.

Second, the intermediate co-infinite infinite outcomes all are assigned the same chance values of V_∞. The intermediate outcomes in the probabilistic case, however, are assigned a range of probabilities and further work is needed to distinguish them. For example, we separated the cases of probability 0.5 and 0.4 by considering a large number of independent trials. The comparable analysis is not needed for the chance function. However, as an exercise in applying the chance function, in Section 8.4 below, it is used to determine the chance of various frequencies of outcomes of even and odd numbers in many, independent drawings of an infinite fair lottery.

8.3 Applying the Rule of Coordination

To get a sense of how this rule is used, we can apply it to a simple case. Consider the chance that the number drawn is less than or equal to some large number *N*. This outcome set has *N* members and thus has chance V_N. It is an outcome not to be expected. The outcome that the number is greater than *N*, however, is in the complement set and thus has chance $V_{\text{-}N}$. It is an outcome we do expect. This must appear strange at first. For it tells us that no matter how large we make *N* – one million, one quadrillion, one million$^{\text{million}}$ — we are sure the number drawn is greater, even though we are certain that some definite, finite number is drawn. There is only strangeness here, but no problem. It is how the chances are in an infinite lottery. All our calculus does is to relate that fact to us.

9. Repeated, Independent, Infinite Lottery Drawings[185]

9.1 Applying Label Independence

To explore the application of this rule further and to see how the chance function behaves, consider the case of repeated, independent drawings from a sequence of identical infinite lottery machines. We will consider the case of *N* independent drawings from *N* machines: machine$_1$, machine$_2$, … , machine$_N$. The combined outcome of *N* drawings will form an *N*-tuple such as

$$<156, 27, 2398, \ldots, 180>_N$$

[185] The analysis of Sections 8 and 9 was decisively advanced by ideas that emerged in an energetic email exchange with Matthew W. Parker. I thank him for this and also for helpful remarks on the present text.

where the subscript N reminds us that there are N elements in the tuple. The set of all such outcomes is Ω_N. It is countably infinite since it is formed as a finite tuple of elements of a countably infinite set.

Label independence can be implemented once again. We consider permutations of the labels on the outcomes of each lottery machine individually. Under such permutations, any N-tuple can be mapped to any other N-tuple. Thus label independence requires that the outcome represented by each N-tuple has an each chance.

Label independence allows us to form equal chance sets of outcome sets, analogous to the equal chance sets of Section 7.1. Consider for example the set of all N-tuples such that every element in each of the member N-tuples is an even number. We will write this as[186]

$$\textit{all-even} = [\textit{even, even, …, even}]_N = \{<n_1, n_2, n_3, \ldots, n_N>_N : \text{all } n_i \text{ even}\}$$

Analogously we have

$$\textit{all-odd} = [\textit{odd, odd, …, odd}]_N = \{<n_1, n_2, n_3, \ldots, n_N>_N : \text{all } n_i \text{ odd}\}$$

When it happens that two sets of outcomes can be mapped onto each other by a label permutation, then label independence requires that the two sets have the same chance. Since they can be so mapped, *all-even* and *all-odd* have the same chance. They belong to the same equal chance set of outcome sets.

This shows that the inductive logic induced by label independence on repeated, independent drawings is similar in structure to that induced on single drawings. We shall see below that the full structure induced for the repeated case is more complicated. However there are simple sectors in the logic that are formally the same as the logic that applies to single drawings.

9.2 A Simple Sector

A simple sector consists of a set of equal chance sets, where those equal chance sets can be totally ordered by set inclusion. That is, the equal chance sets form a chain such the outcomes of each equal chance set is a subset of those higher in the chain. Since the set of all outcomes Ω_N is countably infinite, the equal chance sets will be of the familiar types $\textit{finite}_n$, $\textit{infinite}_{\text{co-infinite}}$ and $\textit{infinite}_{\text{co-finite-}n}$ of Section 7.1. Because they are also totally ordered, we can assign the chance values $V_0, V_1, \ldots, V_\infty, \ldots V_{-1}, V_{-0}$ of (10). If all the cardinalities are not realized by the

[186] The square bracket notation […] is used to preclude the misreading that *all-even* is a N-tuple of sets, whose first, second, third, … members are each the sets of even drawings on machine_1, machine_2, machine_3,… Note—this is a misreading!

equal chance sets, then the sector will only have a subset of these values. Thus the equal chance sets of a simple sector follow the same logic as that governing equal chance sets of single drawings.

A caution: there are many simple sectors in the outcome space of repeated drawings. The chance values only have a meaning within the sector in which they are defined, relative to the chances of the other outcomes in the sector. Without further justification, we cannot assume that the chance of $V_{something}$ in the outcome space of a single drawing has the same meaning chance of $V_{something}$ in a simple sector of the outcome space of repeated drawings.

An example of a simple sector is the set of all outcomes in which all drawings return the same number. The outcome in which number 1 is drawn every time is

$$1_N = \langle 1, 1, 1, \ldots, 1 \rangle_N$$

with an obvious extension of the notation to all 2, all 3, … outcomes. Set complementation with the simple sector gives a notion of negation. For example[187]

$$\textit{not}\ 1_N = 2_N \text{ or } 3_N \text{ or } 4_N \text{ or } \ldots$$

$$\textit{not}\ 2_N = 1_N \text{ or } 3_N \text{ or } 4_N \text{ or } \ldots$$

The outcome 1_N has a single member and is of type $\textit{finite}_1$. The complement $\textit{not}\ 1_N$ is of type $\textit{infinite}_{\text{co-finite-1}}$. Thus:

$$\text{Ch}(1_N) = V_1 \qquad \text{Ch}(\textit{not}\ 1_N) = V_{-1}$$

Applying the rule of coordination, we infer that an outcome in which all numbers drawn in N independent repetitions are 1 is not to be expected in relation to other outcomes in the sector. Correspondingly an outcome in which none of the numbers drawn is 1 is to be expected.

To identify further members in the sector, we ask whether we should expect all the N drawings to yield the same number, where the same number is found in some finite set, say {1, 2, 3}. That is, the outcome is (1_N or 2_N or 3_N). Proceeding as above, we find this outcome is not to be expected, since

$$\text{Ch}(1_N \text{ or } 2_N \text{ or } 3_N) = V_3.$$

We get a different result if we ask after the outcome in which all the numbers drawn are the same, but that number can be any in an infinite set of type $\textit{infinite}_{\text{co-infinite}}$, such as the set of all even numbers; or the set of all odd numbers. These two outcomes are (2_N or 4_N or 6_N or …) and (1

[187] As before, I move without warning between the set representation of the outcome $\textit{not}\ 1_N = \{2_N, 3_N, 4_N, \ldots\}$ and its equivalent propositional representation $\textit{not}\ 1_N = 2_N \text{ or } 3_N \text{ or } 4_N \text{ or } \ldots$.

$_N$ or $\mathbf{2}_N$ or $\mathbf{3}_N$ or …). Since these two outcomes can be mapped onto each other by a permutation of labels and because they are of type $infinite_{\text{co-infinite}}$, we assign the same value

$$\text{Ch}(\mathbf{2}_N \text{ or } \mathbf{4}_N \text{ or } \mathbf{6}_N \text{ or } \ldots) = \text{V}_\infty$$

$$\text{Ch}(\mathbf{1}_N \text{ or } \mathbf{3}_N \text{ or } \mathbf{5}_N \text{ or } \ldots) = \text{V}_\infty$$

These outcomes are "as likely as not" in this sector.

9.3 A Finite Simple Sector

All the finite outcome sets in this last simple sector are subsets of another simple sector. Consider the outcome in which all the numbers drawn in the N repetitions are less than or equal to some big, finite number *Big*, where the numbers drawn need not be the same. This outcome corresponds to a set of Big^N tuples in the outcome set $\mathbf{\Omega}_N$. Thus we have

$$\text{Ch}(\text{all numbers less than or equal to } Big) = \text{V}_{Big^N}.$$

That is, since Big^N is finite, the outcome is one that will generally not happen according to the rule of coordination.

This is a new sector since a permutation of labels cannot map the set of tuples here assigned the value V_{Big^N} onto the set assigned the value V_{Big^N} in the simple sector of Section 9.2. For example, consider the $finite_2$ equal chance sets in each sector. The sector this section will have outcomes like

$$<2, 1, 1, \ldots, 1>_N \text{ or } <3, 1, 1, \ldots, 1>_N.$$

No permutation of labels can map these onto the tuples such as

$$<4, 4, 4, \ldots, 4>_N \text{ or } <5, 5, 5, \ldots, 5>_N$$

in the corresponding $finite_2$ equal chance sets of the simple sector of Section 9.2

We cannot directly compare chance values across different sectors. However our rule of coordination enables us to make some coarser judgments. What of the outcome that at least one of the numbers in N independent drawings is greater than *Big*? This outcome set is the complement of the last set considered with Big^N members. Thus this outcome set is co-finite infinite so that the outcome is to be expected according to the rule of coordination. That is, no matter how big we make *Big* we must always expect that at least one of the numbers drawn in N drawings will be greater than it.

Similarly we cannot directly compare the chance values across the different sectors of Sections 9.2 and 9.3. However our rule of coordination, applied to tuples of drawings, tells us that outcomes realized by finitely many tuples of drawings generally do not happen. If we now assume that outcomes realized by infinitely many tuples of drawings are more likely than the finite case, we arrive at a result that is surely surprising to someone whose intuitions about

chance have been tutored by the probability calculus. It is more likely that all N numbers drawn are the same than it is that all N numbers drawn are less than or equal to some number *Big*, no matter how big we make *Big*. This holds no matter how large we make N.

9.4 A "Likely as Not" Sector

Here are examples illustrating outcomes to which the "as likely as not" chance of V_∞ is assigned. Consider the numbers drawn in N independent repetitions of the infinite lottery:

all-even: all numbers drawn are even numbers
all-odd: all numbers drawn are odd numbers
all-powers: all numbers drawn are powers of 10,
that is, 10, 10^2, 10^3, 10^4, …
not-all-powers: all numbers drawn are NOT powers of 10,
that is, not and of 10, 10^2, 10^3, 10^4, …

Each of these outcomes corresponds to sets of tuples in $\mathbf{\Omega}_N$ of type $infinite_{\text{co-infinite}}$. They can each be mapped into any other by a permutation of the labels on the individual lottery machines. It follows that they have equal chance:

$$\text{Ch}(\textit{all-even}) = \text{Ch}(\textit{all-odd}) = \text{Ch}(\textit{all-powers}) = \text{Ch}(\textit{not-all-powers}) = V_\infty$$

This will seem surprising if we think that there are vastly fewer outcomes in *all-powers* than in *not-all-powers*, since there are vastly fewer powers of ten than numbers that are not powers of ten. Any surprise should be eradicated by recalling that both these sets are countably infinite. The impression that one is bigger than the other is purely an artifact of labeling. Label independence warns us that such artifacts of labeling should be ignored. The two sets in these examples are equinumerous and equinumerous in their complements and can be mapped onto each other by a label permutation.

9.5 Further Sectors

The chance logic of repeated independent infinite lottery drawings includes further sectors with more complicated properties. An indication of their nature follows from consideration of two independent drawings. Consider the outcome that the first number drawn is 1 and that the second number drawn is even, that is [1, *even*], and then another outcome [1 or 2, *even*]. Both can be mapped one-one by label permutations onto infinite-co-infinite sets of pairs. However no permutation of labels can map [1, *even*] to [1 or 2, *even*]. Thus they cannot be required by label independence to have the same chance value. We would need to assign them different chance values. In an obvious notation they might be $V_{1,\infty}$ and $V_{2,\infty}$. In this notation, the outcome [*even*, *even*] would be assigned the value $V_{\infty,\infty}$. The applicable chance logic would

then reside in relations analogous to those of (11), such as $V_{1,\infty} \leq V_{2,\infty} \leq \ \ldots \ \leq V_{\infty,\infty}$; and $V_{1,\infty} = V_{\infty,1}$; $V_{2,\infty} = V_{\infty,2}$; etc.

10. Relative Frequencies of "as likely as not" Outcomes

10.1 Can frequencies reintroduce probabilities?

The inductive logic induced by label independence precludes an ordinary probabilistic logic. We might wonder, however, whether probabilities can be reintroduced indirectly by an empirical approach. We carry out many, independent drawings and let the limiting behavior of the frequencies reintroduce probabilities. This approach would succeed with a finite lottery. In independent repetitions, we expect with high probability, that roughly half the numbers drawn will be even and half of them odd. That is a consequence of the probabilistic fact that an even number is drawn with probability 1/2.

We should not expect similar results in an infinite lottery, for the value V_∞ assigned to both even and odd outcomes is quite removed in its formal properties from a probability 1/2. We shall see in this section by direct calculation that the chance function of the infinite lottery does not return the favoring of relative frequencies of odd and even outcomes such as would be needed to reintroduce a probability of one half for each.

10.2 Odd and even outcomes

Consider $N > 1$ independent drawings of the lottery as in Section 9. The outcome sets that interest us are sets of N-tuples of the form

$$[odd, odd, \ldots, even, odd, even, even]_N$$

$$= \{<n_1, n_2, n_3, \ldots, n_N>_N :$$

$$n_i \text{ is an odd number in the positions marked "}odd\text{"}$$

$$\text{and an even number in the positions marked "}even\text{"}\}$$

Since each of *odd* and *even* are realized by infinitely many numbers, the set of N-tuples realizing any particular outcome set of the form $[odd, odd, \ldots, even, odd, even, even]_N$ is infinite. Correspondingly there are infinitely many ways that the complement set could be realized. Thus the outcome is co-infinite infinite and it has chance V_∞ of the simple sector of Section 9.3.

Permuting the labels on the individual lottery machine outcomes, we find that each of these outcome sets can be mapped onto any other. For example the outcome set

$$[odd, odd, \ldots, even, odd, even, even]_N$$

can be mapped onto the outcome set

$$\textit{all-odd} = [\textit{odd}, \textit{odd}, \ldots, \textit{odd}, \textit{odd}, \textit{odd}, \textit{odd}]_N$$

We take the lottery machines in the positions marked "*even*" in the first outcome set and apply a permutation of labels that switches odd and even numbers. It follows that all the outcome sets of odd and even outcomes in this subsection have equal chances.

10.3 Frequencies of even outcomes

Our concern is not just the outcome sets of Section 10.2. We want to know the chances of n even numbers in N independent draws. Those chances are assigned to larger outcome sets. The case of n=0 is the *all-odd* tuple above. The case of n=1 is realized as the union of N outcome sets

$$1\ \textit{even} = [\textit{even}, \textit{odd}, \ldots, \textit{odd}, \textit{odd}, \textit{odd}, \textit{odd}]_N$$

$$\cup\ [\textit{odd}, \textit{even}, \ldots, \textit{odd}, \textit{odd}, \textit{odd}, \textit{odd}]_N\ \cup$$

$$\ldots$$

$$\cup\ [\textit{odd}, \textit{odd}, \ldots, \textit{odd}, \textit{odd}, \textit{odd}, \textit{even}]_N$$

In general, the number of these outcome sets to be joined to form the set of n *even* outcomes is given by the combinatorial factor $C(N,n) = N!/(n!\,(N\text{-}n)!)$. This combinatorial factor is always finite for finite N and n. It follows that there are still infinitely many N-tuples of individual outcome numbers that realize the outcome of exactly n even numbers in any order amongst the N drawings; and also infinitely N-tuples in the complement set.

As a result it is natural to assign the chance value V_∞ to each outcome of n even numbers among N draws, for any n. We might then continue with the natural supposition that each outcome of n even numbers among N draws has the same chance, for any n. I drew just this conclusion in an earlier draft of this chapter and reported it in a paper (Norton, manuscript, §9).

Unfortunately the inference to this conclusion is a fallacy and I retract it. That the outcomes have the same chance requires that they be in the same sector of the infinite logic. The values V_∞ reported might be drawn from different sectors. Then they would have an immediate meaning only within each sector. To conclude that they represent equal chances requires further argumentation. Ideally, we would need to show that permuting the labels takes us from one outcome of n even numbers to any other, which would show that they are within the same sector after all. This has not been shown and cannot be shown.

For it is easy to show that the outcome set of n=0 even numbers drawn cannot be mapped by a label permutation onto the outcome set of n even numbers drawn, where $0 < n < N$. To see this, for the purpose of a *reductio*, assume otherwise: that there is such a mapping for some particular value of $0 < n < N$. Then a permutation of labels must include mappings of N-tuples of the form

$<o_{1,1}, o_{1,2}, o_{1,3}, \ldots o_{1,N}> \rightarrow <e_{1,1}, ?, ?, \ldots, ?>$

$<o_{2,1}, o_{2,2}, o_{2,3}, \ldots o_{2,N}> \rightarrow <?, e_{2,2}, ?, \ldots, ?>$

…

$<o_{N,1}, o_{N,2}, o_{N,3}, \ldots o_{N,N}> \rightarrow <?, ?, ?, \ldots, e_{N,N}>$

Here $o_{1,1}, o_{1,2, \ldots}, o_{N,N}$ are odd numbers that enter into N-tuples that map to N-tuples with even numbers $e_{1,1}, e_{2,2}, \ldots, e_{N,N}$ in the positions shown. The "?, ?, ?, …" represent further numbers that may be odd or even, but have at least one odd number in each N-tuple.

Since the label permutations are carried out independently on each machine, it now follows that the label permutation on the set of machines must also include the map

$<o_{1,1}, o_{2,2}, o_{3,3}, \ldots, o_{N,N}> \rightarrow <e_{1,1}, e_{2,2}, e_{3,3}, \ldots, e_{N,N}>$

However this mapping is not included in the mapping supposed, for an N-tuple drawn from n=0 *even* outcome set is mapped to an N-tuple drawn from the n=N *even* outcome set. This contradiction completes the *reductio*.

While not all outcome sets with n even numbers can be mapped onto each other. There are a few mappings that do succeed. We can map the outcome set with n even numbers among N draws onto the outcome set with N-n even outcomes merely by a permutation that switches everywhere odd and even numbers in each lottery machine. Thus we have

$$\text{Ch}(n \textit{ even}) = \text{Ch}(N - n \textit{ even}) \text{ for all } 0 \leq n \leq N$$

In Appendix B, it is shown that this last possibility exhausts all the possibilities for equivalences under label permutation in the case of n *even* outcomes. That is, it is shown that a label permutation cannot map the outcome set n *even* to the outcome set m *even*, unless $n = N - m$.

In the following two sections, we shall see that we can infer enough equivalences under label permutation to show that the essential point reported is correct: the chances of n even outcomes do not make likely a stabilization of frequencies that accord with probabilistic expectations.

10.4 The Chances of N odd versus N even in N drawings

The simplest case arises with the two extremes *all-even* and *all-odd*. They are in the same sector since a permutation of the individual lottery labels can map one onto the other. To probe their chance behavior, consider another property:

$$\textit{div } m = \text{set of numbers divisible by } m$$

and its complement *not div m*. The outcomes *even* and *odd* are the special case of m=2. We have from earlier that a permutation of labels can map each of *even, odd, div m, not div m* onto each other. So they individually have the same chance. It now follows immediately that the same is true of the N tuples

all-even = [*even*, *even*, …, *even*, *even*, *even*, *even*]$_N$

all-odd = [*odd*, *odd*, …, *odd*, *odd*, *odd*, *odd*]$_N$

all-div m = [*div m*, *div m*, …, *div m*, *div m*, *div m*, *div m*]$_N$

all-not div m = [*not div m*, *not div m*, …, *not div m*, *not div m*, *not div m*, *not div m*]$_N$

They have equal chance, so we may write:

Ch(*N div m* in *N*) = Ch(*N even* in *N*) = Ch(*N odd* in *N*) = Ch(*N not-div m* in *N*)

These equalities differ markedly from probabilistic expectations. Since we have P(*div m*) = 1/*m* and P(*not-div m*) = (*m*-1)/*m*, we expect

P(*N div m* in *N*) = [1/*m*]^N << P(*N* not-*div m* in *N*) = [(*m*-1)/*m*]^N

That is, the outcome (*N not-div m* in *N*) is (*m*-1)^N times as probable as outcome (*N div m* in *N*). It is the basis of the probabilistic expectation that *not-div m* outcomes are likely to occur much more frequently than *div m* outcomes (for *m* >2). The equalities of the chance function do not reflect this probabilistic favoring or the associated expectations concerning frequencies.

10.5 Chances of Intermediate n even drawings in N drawings

The last section shows that the chance of frequencies of *div m* in *N* drawings is independent of *m* for the extreme *n* = *N* case of *all-div m*. This independence of the chances from *m* holds for all values of *n*. That is, the chance of 0, 1, 2, … occurrences of a *div m* number in *N* drawings is independent of the value of *m*. Below I sketch a diagrammatic proof for the simple case of *N*=2. The proof will then be generalized to all *N*.

In two independent drawings, we will represent the four possible outcomes sets as

OO = [*odd*, *odd*] *OE* = [*odd*, *even*] *EO* = [*even*, *odd*] *EE* = [*even*, *even*]

The frequency *n* = 0 corresponds to *OO*; *n* = 1 to (*OE* or *EO*); and *n* = 2 to *EE*. Figure 1 one lays out the pairs of individual number outcomes in a grid. (It only shows a finite corner of the infinite grid.) The first number drawn is on the horizontal axis and the second number drawn is on the vertical axis. The set of pairs that comprise *OO* is shown by the distribution of the labels "*OO*"; and so on for the remaining outcomes.

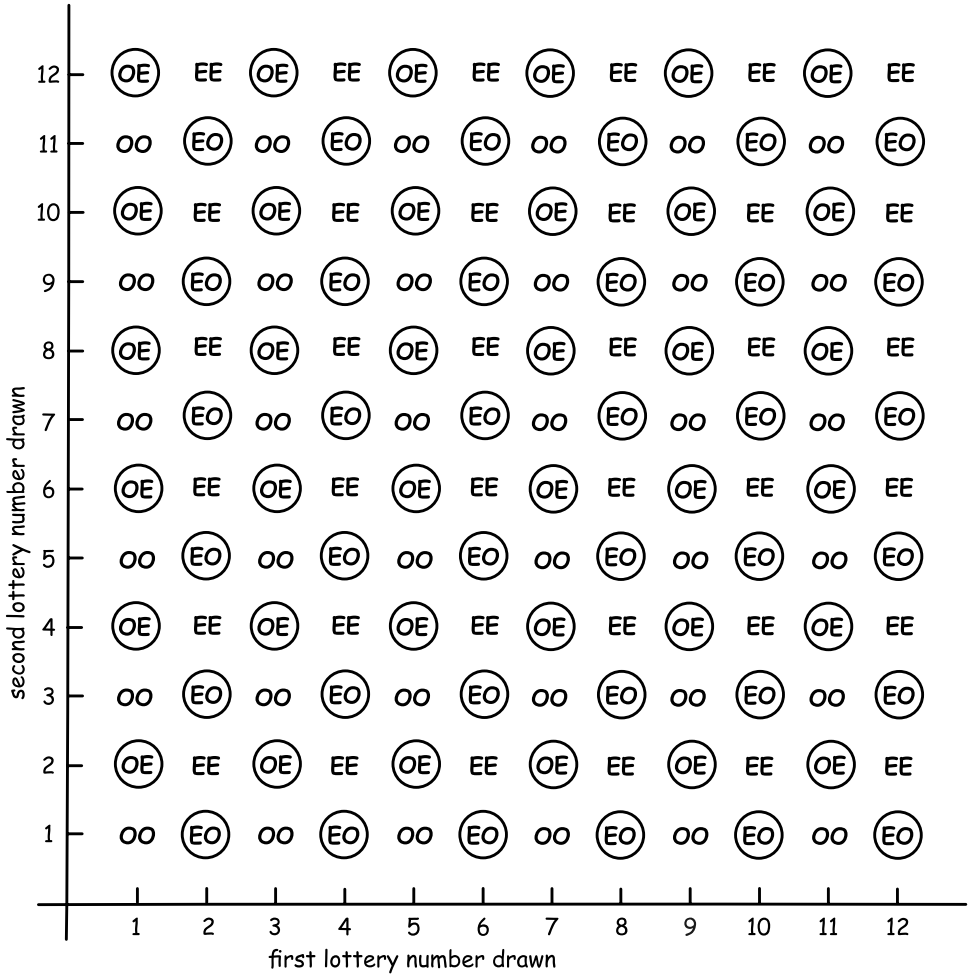


Figure 1. Distribution of outcomes *OO*, *OE*, *EO* and *EE* in a two lottery outcome space

We will permute the labels so that the outcome sets for $n = 0$, $n = 1$ and $n = 2$ *even* outcomes coincide with the outcomes sets for $n = 0$, $n = 1$ and $n = 2$ *div* 6 outcomes.

A permutation of the labels of the first lottery can be represented in the figure by leaving the labels in their positions on the axes and permuting the columns associated with the first lottery's numbers. The requisite permutation shifts the first five odd numbered columns, 1, 3, 5, 7, 9 to the left; and then places the first even numbered column 2 after it; and so on for the all the column numbers: five odd numbered columns, then an even numbered column, repeatedly. The result is shown in Figure 2.

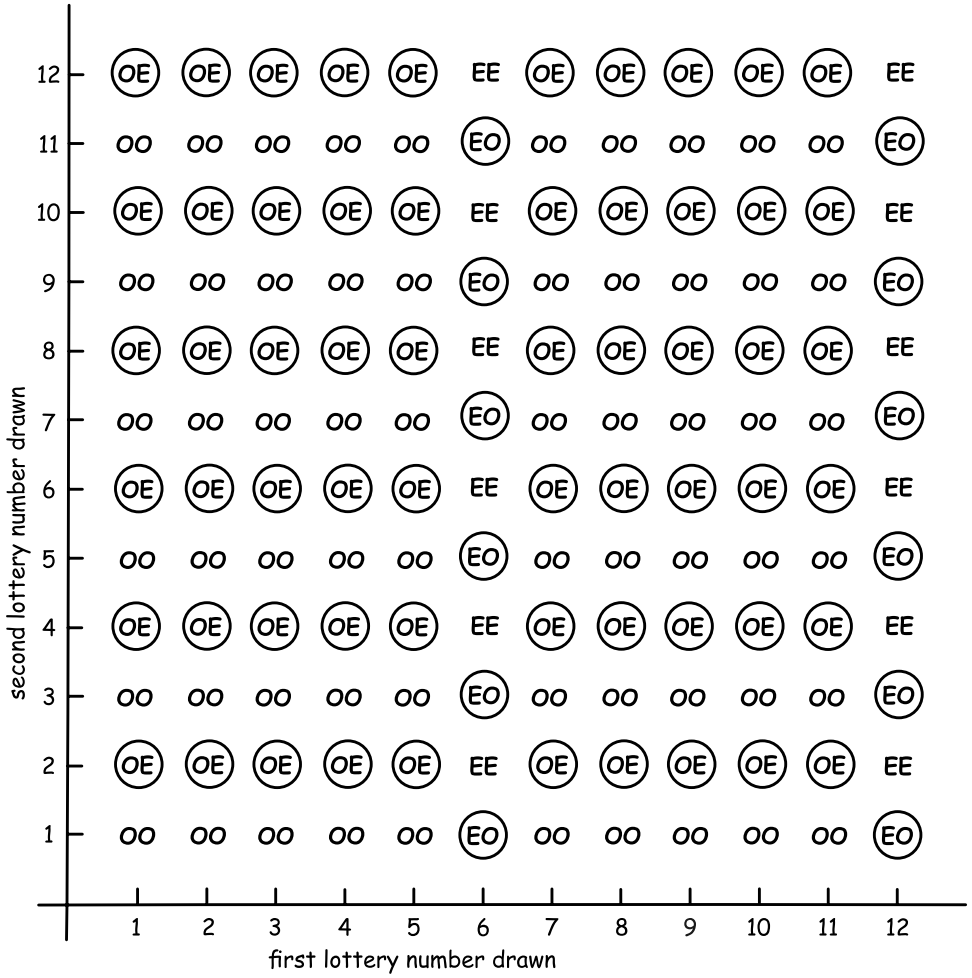


Figure 2. Result of permuting the columns

To complete the manipulation, we perform the same permutation on the labels of the second lottery. That is, we perform the corresponding permutation of the rows to which the second lottery's numbers are associated. The result is shown in Figure 3

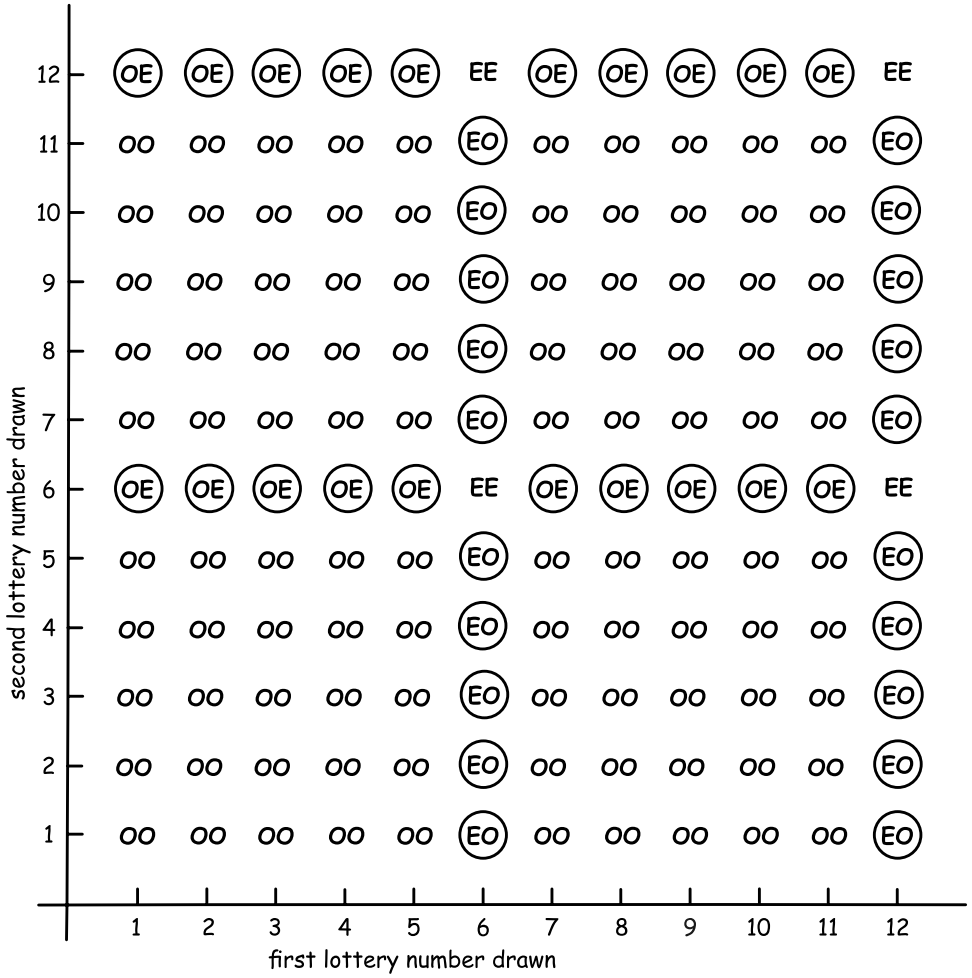


Figure 3. Result of permuting the columns and rows

We read from Figure 3 that the outcomes have been relocated as follows

$n = 0$ *even* outcomes (*OO*) coincides with $n = 0$ *div* 6 outcomes

$n = 1$ *even* outcomes (*OE* or *EO*) coincides with $n = 1$ *div* 6 outcomes

$n = 2$ *even* outcomes (*EE*) coincides with $n = 2$ *div* 6 outcomes

Thus the chances of *n even* outcomes equals the chances of *n div* 6 outcomes, for all *n.*

The figure shows the manipulation for the case of $m = 6$. It is clear that it will succeed for any value of $m>2$. It follows that the chances of the frequencies are independent of whether we are asking after even numbers, or numbers divisible by 6 or 10 or 100 or 1000. That is, the chances of these frequencies do not conform with the probabilistic expectations that even numbers appear in repeated trials roughly half the time and that those divisible by 6 or 10 or 100 or 1000 appear roughly 1/6 or 1/10 or 1/100 or 1/1000th the time, respectively.

10.6 The general case[188]

The general result is that the chances of *n div m* outcomes in *N* drawings is independent of the value of m for all $0 \leq n \leq N$.

To see it, first note that there is a permutation of the label numbers of one lottery machine such that the set *div m* is mapped exactly onto the set *div k* for any $m, k > 1$. That is, under the permutation, all number labels divisible by *m* are switched with all number labels divisible by *k*. The construction of the *N*=2 case displays the permutation for the case of $m = 2$ and $k = 6$.

Consider any *N*-tuple of outcomes that has exactly *n* outcomes divisible by *m*, that is, is drawn from the set *div m*. Under the permutation, this *N*-tuple is mapped to one that has exactly *n* outcomes divisible by *k*, that is, drawn from the set *div k*. Now consider the set of all *N*-tuples with exactly *n* outcomes divisible by *m*. The same permutation will map it to the set of all *N*-tuples with exactly *n* outcomes divisible by *k*. Thus label independence entails that the two sets have the same chance and we can write:

$$\text{Ch}(n \textit{ div } m \text{ in } N) = \text{Ch}(n \textit{ div } k \text{ in } N) = \text{Ch}(n \textit{ even} \text{ in } N)$$

for all $0 \leq n \leq N$ *and* any $m, k > 1$. Since the outcomes of *n even* and *N-n even* may be mapped into each other, we can extend these equalities of chances:

$$\text{Ch}(n \textit{ div } m \text{ in } N) = \text{Ch}(n \textit{ even} \text{ in } N) = \text{Ch}(N\text{-}n \textit{ even} \text{ in } N) = \text{Ch}(N\text{-}n \textit{ even } m \text{ in } N)$$

for all $0 \leq n \leq N$.

10.7 Frequencies do not give us probabilities

What these results show is that the tempting strategy for reintroducing probabilities fails. The temptation is to say "Do the experiment. Run many independent drawings from lottery machines. Read the limiting frequencies in many drawings. They will reveal to you the probabilities hidden in the lottery machines!"

The strategy fails since the chances of different frequencies do not mass in a way that would reveal probabilities. Probabilistic intuitions would lead us to expect that drawing all *N* numbers divisible by 100 in *N* draws would be much less likely that drawing all *N* numbers *not* divisible by 100 in *N* draws. Yet they have the same chance so we have no reason to expect the second over the first.

These same probabilistic intuitions would lead us to expect that the most likely numbers of even drawings in *N* drawings would cluster around *N*/2. Numbers of even drawings far from *N*/2 would be unlikely. From this clustering, we could recover a probability of one half for an even number. The trouble is that this same clustering around *N*/2 is likely for outcomes divisible by 10 or 100 or 1000. We would then have to infer that numbers divisible by 10, 100 or 1000 or

[188] I thank Matthew W. Parker for this proof.

any other number greater than 2 also have a probability of one half. No ordinary probability distribution can realize these probabilities.[189]

The calculations reviewed in this Section and in Appendix B show that the chances of securing *n* or *m even* numbers in *N* repeated independent draws from infinite lottery machines are incomparable for most *n* and *m*. Thus this section leaves open whether imposition of further background facts will lead to further relations that will lead to chances favoring certain frequencies of outcomes. However what has been shown is that if there is any favoring, it is not of a type that can be used to reveal underlying probabilities as long as the fair character of the infinite lottery is preserved.

11. Failure of the Containment Principle

This infinite lottery logic will likely be discomforting for someone whose intuitions are guided by probability theory. One source of discomfort may be that the removal of elements from an outcome set commonly does not reduce the chances of the outcome. It would seem natural that the set of even numbered outcomes {2, 4, 6, 8, …} must be assigned greater chance than the set of every fourth numbered outcome {4, 8, 12, 16, …}. This second set is properly contained in the first. However the present logic assigns the same chance to both. We might express the intuition more clearly as:

> *The containment principle.* If a set of outcomes *A* is properly contained in a set of outcomes *B*, then the chance of *A* is strictly less than the chance of *B*:
> Ch(A) < Ch(B).

If the background facts support it, there is no problem with a logic that conforms with this principle. However the principle cannot lay claim to a preferred status. As is always the case, whether a logic has some feature is decided by prevailing background facts. The background fact of label independence entails the failure of the containment principle.

Two further considerations reduce the appeal of the principle:

First, the containment principle has not been uniformly respected in familiar probabilistic applications. There is a probability zero of a dart hitting any particular point on a dartboard of continuum many points. The same zero probability is assigned to the dart hitting any of a

---

[189] Assume otherwise. Then the probability of drawing a number divisible by 2^r is one half, for any $r>1$. Since the probability of drawing a number divisible by 2 is also one half, it follows the probability of drawing numbers divisible only by $2^1, 2^2, \ldots, 2^{r-1}$, is zero. But since *r* can be set as large as we like, we infer that the chance of a number divisible by any power of two is zero, which contradicts the probability of one half for even numbers.

countable infinity of points on the dartboard, even if that set contains the single point originally considered. In another example, we follow de Finetti's prescription for the infinite lottery and employ a probability measure that is only finitely additive. Then the probability of drawing a one is the same the probability of drawing any number less one hundred million. Both are zero probability outcomes.

Second, the containment principle by itself is insufficient to induce chances that can compare all sets of outcomes. Since the set of even numbered outcomes is disjoint from the set of odd multiples of three $\{3, 9, 15, 21, 27, \ldots\}$, we are left unable to compare their chances. In such cases, we may be inclined to retain the chance assignments of the present logic: if disjoint outcome sets (and their complements) are equinumerous, then they are assigned the same chance. What results, however, is a non-transitive comparison relation for chances. We have from considerations of equinumerosity that:

$$\text{Ch}(\{2, 4, 6, 8, \ldots\}) = \text{Ch}(\{3, 9, 15, 21, 27, \ldots\})$$

$$\text{Ch}(\{4, 8, 12, 16, \ldots\}) = \text{Ch}(\{3, 9, 15, 21, 27, \ldots\})$$

If transitivity of the comparison relation for chances is supposed, it follows that:

$$\text{Ch}(\{4, 8, 12, 16, \ldots\}) = \text{Ch}(\{2, 4, 6, 8, \ldots\}).$$

This equality contradicts the containment principle, which tells us that:

$$\text{Ch}(\{4, 8, 12, 16, \ldots\}) < \text{Ch}(\{2, 4, 6, 8, \ldots\}).$$

If transitivity is dropped, we will be unable to assign a single value to each chance, but only assign pairwise comparisons of strength. Presumably some accommodation of the two approaches can be found eventually, but it may not be pretty or simple.

In sum, we should use the containment principle when the background facts call for it. When they do not call for it, we should feel no special loss at its failure.

12. Is An Infinite Lottery Machine Physically Possible?

The discussion so far has presumed the physical possibility of an infinite lottery machine. In what sense are they physically possible? Elsewhere (Norton, 2018; Norton and Pruss, 2018, Norton, manuscript a) I have pursued the question is greater detail. The answer proves to be more complicated and much more interesting than one might first imagine.

The natural starting point is to seek some design that employs ordinary probabilistic randomizers, such as coin tosses, die throws and pointers spun on dials. We run into difficulties immediately. We will need infinite powers of discrimination to distinguish among the infinitely many possible pointer outcomes crammed onto the scale etched onto the surface of the dial. If we use coins or dice, we will need to use infinitely many of them to create an outcome space big enough to hold the countable infinity of outcomes of the infinite lottery machine.

If we are undaunted by the task of flipping infinitely many coins or reading pointer positions with infinite precision, the prospects for an infinite lottery machine seem good. Infinitely many coin tosses produce an outcome space of continuum size, that is, an order of infinity higher than that needed for the countably infinite outcomes of the infinite lottery machine. Somewhere in it we would expect to find a countable infinity of outcomes that implement an infinite lottery machine.

However in Norton (2018), as corrected by Norton and Pruss (2018), we found a maddening problem. With some ingenuity, we can use ordinary probabilistic randomizers to form infinite lottery machines. However in every design we could imagine, there was always a probability of zero that the machine would operate successfully. The persistence and recalcitrance of the failure gave the clue that the problem was not merely one of an impoverished imagination for the design of the infinite lottery machines. There was some unidentified matter of principle defeating all attempts.

In Norton (manuscript a) that matter of principle is recovered from what I would otherwise have imagined to be the arcana of measure theory and axiomatic set theory. The probabilistic randomizers will provide us with an outcome space expansive enough to host the infinite lottery outcomes that encode results "1," "2," "3," and so on. If a probability is defined for each of these outcomes, then that probability must be the same for each and can only be zero. For otherwise, if it is greater than zero, we need only sum finitely many of the equal, non-zero probabilities $P(1)$, $P(2)$, $P(3)$, … to arrive at a sum greater than one. That sum contradicts the normalization of the probability measure to unity. If, however, we set each of the probabilities $P(1)$, $P(2)$, $P(3)$, … to zero, then the probability that any one of the infinite lottery outcomes, 1, 2, 3, …, arises is zero. For it is given by the sum

$$P(1) + P(2) + P(3) + \ldots = 0 + 0 + 0 + \ldots = 0$$

That means that the infinite lottery machine operates successfully only with probability zero.

The escape is to use infinite lottery outcomes to encode results "1," "2," "3," … that are probabilistically nonmeasurable. Norton (manuscript a) describes two designs that do this. The same difficulty besets both. Their designs presume the existence of the nonmeasurable outcome sets, but do not specify which those sets are. That means that, after the randomizers settle into some end state, we cannot know the outcome set to which they belong. The number selected as the infinite lottery outcome is inaccessible to the user, rendering the device useless.

It turns out that, as far as we know, this failure must always happen. For all known examples of nonmeasurable sets are nonconstructive and we have some reason to expect that none can be constructed. That means that we are allowed to assume their existence, commonly

by virtue of the axiom of choice of axiomatic set theory, or something equivalent to it.[190] However there is no explicit description for which they are. We are caught in a dilemma. If an infinite lottery machine based on ordinary probabilistic randomizers is to return a result we can read, it will do so successfully only with probability zero. If we demand a probability of success greater than zero, then we can have it, but the result of the infinite lottery machine will be inaccessible to us.

These results apply only to infinite lottery machines constructed from ordinary probabilistic randomizers. They do not preclude other designs. Norton (2018, manuscript a) describes designs based on quantum mechanical systems. In the simplest, one takes a quantum particle in a definite momentum state. It consists of a wave uniformly distributed over space in the direction of the momentum. Divide that space into a countable infinity of intervals of the same size, numbered 1, 2, 3, …. If we now perform a measurement on the position of the particle, it will manifest with equal chances in each interval. An infinite lottery machine has been implemented.

While the exercise of designing these infinite lottery machines is entertaining, I take a more permissive view of them. For hundreds of years, the paradigm of a probabilistic system in probability theory was the coin toss, die throw and card shuffle. Yet prior to quantum theory, our best science told us that none of these was a true randomizer. Probability theory thrived merely by supposing that these real randomizers were imperfect surrogates for true but unrealizable probabilistic randomizers: idealized ideal coin tosses, die throws and card shuffles. We can, I propose, take the same attitude to infinite lottery machines. They are an idealized case that can be added to our repertoire of idealized randomizers. We can and should ask what inductive logic is adapted them.

Finally, we should separate the issue of the cogency of the design of an infinite lottery machine from the cogency of the infinite lottery logic described in this chapter. We may not be able to specify explicitly which are the infinite lottery outcomes of a probabilistically based machine. But, on the authority of the axiom of choice, they exist. So we can ask what chance each has of being realized; and we should expect a suitable logic of induction to tell us.

13. Conclusion

The infinite lottery remains one of the most popular arguments used to establish that the countable additivity of a probability measure must be reduced to mere finite additivity. What this chapter shows is that the implications of the infinite lottery are still stronger. It requires also that

[190] For more on nonmeasurable sets and the axiom of choice, see Chapter 14.

we abandon finite additivity. The existing literature has been reluctant to accept this further conclusion for it requires abandoning probabilities as the gauge of the possibility of the various outcomes. However, as I argued in Section 6, to persist in the use of a finitely additive probability measure for this purpose is to change the problem posed by adding further conditions, such as a preferred numbering of the outcomes. The original infinite lottery problem is solved by a non-additive logic such as developed in Sections 7 and 8.

The new chance logic of these Sections will seem strange to those already steeped in probabilistic thinking. The strangeness is merely a result of its unfamiliarity. It is easy to lose sight of how abstruse is even the notion of probability. It was once unfamiliar to all of us. Imagine trying to convey to someone new to it that there is a probability of 0.5 that their unborn child will be a girl. We may eventually convey the idea by saying:

"What is the probability of a girl?
It is the same as getting heads on a fair coin toss."

This formulation uses a physical randomizer as a benchmarking device.

Now consider the cosmologists described in Appendix A. They consider the infinitely many *like* and *unlike* patches spawned by eternal inflation. They find the chance properties of the patches to conform with label independence; and they find themselves confused by the resulting chance behavior. We should be able to use the same benchmarking strategy to clarify these chance properties for them:

"What is the chance of a *like* patch?
It is the same as the chance of an even number in an infinite, fair lottery."

Appendix A: The "Measure Problem" in Eternal Inflation[191]

A1 Inflation and Eternal Inflation

Inflation in cosmology is a brief period of very rapid expansion in the very early universe. It has the same effect as taking a wrinkled rubber sheet and stretching it to an enormous size. The wrinkles are all but eliminated. This smoothing process motivated in large part the introduction of inflation into cosmological theory in the 1980s. The smoothing would explain why the cosmic matter distribution is so uniform on the largest scale and why the geometry of space is so close to flat. It also explains why, contrary to expectations of exotic particle theories, we see no magnetic monopoles. The inflationary stretching of space exiles them to parts of the cosmos we cannot see.

Under continuing criticism, the status of inflation in modern cosmology remains mixed. It was unclear that there ever was a pressing need to explain these features of the cosmos through further theory. The matter driving inflation was initially supposed to come from novel particle physics: a "GUT," that is, a grand unified theory. Those efforts failed. The driving matter is now just a novel matter field, the inflaton, posited *ad hoc* with just the right properties. Moreover, the search for a viable form of inflation has led to multiple versions, so that it is not so much a single theory as a program of research.

Nonetheless, the notion has proven quite appealing and it has become a staple, if debated, topic in cosmology. The strongest argument for it comes from its treatment of quantum fluctuations. During inflation, tiny, evanescent quantum fluctuations are amplified to cosmic scales, where they are "frozen in" as classical perturbations in matter density that match the nonuniformities we observe now.

The original idea was that there would be an early period of inflation, driven by the exotic matter of the inflaton field. This rapid expansion would cease and be followed by a more slowly expanding state, driven by familiar forms of matter and radiation. Eternal inflation is a variation in which this cessation of inflation never happens universally. Rather it happens in patches, with each patch reverting to a modestly expanding universe with ordinary matter. Each is a pocket universe or little island universe. Outside these patches, inflation continues. Since inflating space grows so much faster than the space of the patches, the universe overall persists

[191] For a fuller discussion of the measure problem and its inductive analysis, see Norton (manuscript).

eternally in an inflating state, continuously spawning non-inflating pocket universes. One of these pocket universes is our observable universe.

A2 The Measure problem: Should we be here?

The immediate question asked of eternal inflation is whether we should expect a spawned pocket universe to be like our observable universe. It would count against eternal inflation if a universe like ours is exceptional among the non-inflating universes spawned. The measure problem is the problem of finding a way to quantify how much we should expect patches like ours.

The difficulty can be seen in a simplified version of the problem in which we introduce a binary classification: pocket universes *like* ours versus pocket universes *unlike* ours. We gauge the extent to which a universe like ours will come about in eternal inflation by asking after the distribution of *like* and *unlike* over the pocket universes. It is natural to ask for the probabilities of each. That query leads to trouble.

Alan Guth introduced inflation to cosmology in the early 1980s. Here is his development of the problem (2007, p. 11):

> However, as soon as one attempts to define probabilities in an eternally inflating spacetime, one discovers ambiguities. The problem is that the sample space is infinite, in that an eternally inflating universe produces an infinite number of pocket universes. The fraction of universes with any particular property is therefore equal to infinity divided by infinity—a meaningless ratio. To obtain a well-defined answer, one needs to invoke some method of regularization.

Since there is a countable infinity of these pocket universes, we can see the similarity to the infinite lottery problem. It is like asking after the distribution of *even* and *odd* tickets in the lottery. Guth continues the above remarks by making the connection:

> To understand the nature of the problem, it is useful to think about the integers as a model system with an infinite number of entities. We can ask, for example, what fraction of the integers are odd. Most people would presumably say that the answer is 1/2, since the integers alternate between odd and even. That is, if the string of integers is truncated after the Nth, then the fraction of odd integers in the string is exactly 1/2 if N is even, and is $(N+1)/2N$ if N is odd. In any case, the fraction approaches 1/2 as N approaches infinity.
>
> However, the ambiguity of the answer can be seen if one imagines other orderings for the integers. One could, if one wished, order the integers as
>
> $$1, 3, 2, 5, 7, 4, 9, 11, 6\ ,..., \qquad (11)$$

always writing two odd integers followed by one even integer. This series includes each integer exactly once, just like the usual sequence (1, 2, 3, 4,...). The integers are just arranged in an unusual order. However, if we truncate the sequence shown in Eq. (11) after the Nth entry, and then take the limit $N \to \infty$, we would conclude that 2/3 of the integers are odd. Thus, we find that the definition of probability on an infinite set requires some method of truncation, and that the answer can depend nontrivially on the method that is used.

Guth correctly recognizes that recovering a well-defined probability requires us to add something. He calls it "regularization" and it corresponds to imposing an order on the set of outcomes quite analogous to that used in Section 6 above. The difficulty, of course, is that there are multiple choices for the ordering and each typically leads to a different probability measure.

In including regularization in the set up of the problem, Guth presumes more than is needed to arrive at it. The same problem is generated in Section 5 above merely by matching one-to-one infinite sets of the same cardinality. Paul Steinhardt is also one of the founding figures of inflationary cosmology and now one of its sternest critics. He sets up the problem using cardinality considerations alone (2001, p. 42):

> In an eternally inflating universe, an infinite number of islands will have properties like the ones we observe, but an infinite number will not. The true outcome of inflation was best summarized by Guth: "In an eternally inflating universe, anything that can happen will happen; in fact, it will happen an infinite number of times."
>
> So is our universe the exception or the rule? In an infinite collection of islands, it is hard to tell. As an analogy, suppose you have a sack containing a known finite number of quarters and pennies. If you reach in and pick a coin randomly, you can make a firm prediction about which coin you are most likely to choose. If the sack contains an infinite number of quarter and pennies, though, you cannot. To try to assess the probabilities, you sort the coins into piles. You start by putting one quarter into the pile, then one penny, then a second quarter, then a second penny, and so on. This procedure gives you the impression that there is an equal number of each denomination. But then you try a different system, first piling 10 quarters, then one penny, then 10 quarters, then another penny, and so on. Now you have the impression that there are 10 quarters for every penny.
>
> Which method of counting out the coins is right? The answer is neither. For an infinite collection of coins, there are an infinite number of ways of sorting that produce an infinite range of probabilities. So there is no legitimate way to judge

which coin is more likely. By the same reasoning, there is no way to judge which kind of island is more likely in an eternally inflating universe.

A3 No Probabilities—No Predictions

Guth seems optimistic that there will be a solution to the measure problem. Steinhardt is pessimistic and uses his pessimism as grounds for criticizing inflationary theory. However they agree that securing probabilities is essential to eternal inflation as a predictive theory. Guth (2007, p. 11) writes: "To extract predictions from the theory, we must therefore learn to distinguish the probable from the improbable." Steinhardt (2011, p. 42) is more forthright in his concern:

> Now you should be disturbed. What does it mean to say that inflation makes certain predictions—that, for example, the universe is uniform or has scale-invariant fluctuations—if anything that can happen will happen an infinite number of times? And if the theory does not make testable predictions, how can cosmologists claim that the theory agrees with observations, as they routinely do?

He then reviews with disdain the idea of imposing a measure on the islands (pp. 42-43):

> An alternative strategy supposes that islands like our observable universe are the most likely outcome of inflation. Proponents of this approach impose a so-called measure, a specific rule for weighting which kinds of islands are most likely—analogous to declaring that we must take three quarters for every five pennies when drawing coins from our sack. The notion of a measure, an ad hoc addition, is an open admission that in inflationary theory on its own does not explain or predict anything.

Guth and Steinhardt share an all or nothing view: if probabilities cannot be secured, then the theory has failed as an instrument of prediction. This view is based on a widely accepted but false presumption: that the only precise way to deal with uncertainties is through probabilities. A major goal of this entire work is to show that this presumption is too severe and too narrow. We can still deal formally with uncertainty when probabilities are inapplicable. The background facts may merely warrant an inductive logic that is not probabilistic. In this case, the inductive logic warranted is summarized in the chance function (10).

We should separate the question of whether there is an inductive logic native to the situation from the question of whether we can secure the sorts of prediction we might like. In the case of eternal inflation, there is a well-defined inductive logic applicable. However it turns out not to support the sorts of predictions the cosmologists seek. The difficulty is that the inductive logic assigns the same chance V_∞ to any universe in which there are infinitely many *like* pocket universes and infinitely many *unlike* pocket universes. Since this combination encompasses

virtually all the possibilities that can be realized,[192] the logic is unable to discriminate among them usefully, that is, in a way that might privilege *like* universes.

Some prediction is still possible. The chance function (10) has predictive powers, as shown in Sections 9 and 10 above. They may be weaker than the predictive powers of a full probability measure. But that is all that the specification of the infinite lottery permits.

More generally, we cannot demand that the universe gives us theories of the type that we happen to like. We may prefer theories of indeterministic processes always to be endowed with probabilities, for they enable strong predictions. However the world is under no obligation to provide such theories. Probabilities are not provided by the indeterministic systems described in a later chapter; and the theories are correspondingly weak in predictions. That fact does not make them failures as theories. They just happen to be the best the world will give us.

Appendix B: Inequivalences Under Label Permutation of Outcomes of Many Independent Drawings

The numbers drawn independently from N infinite lottery machines form an N-tuple $<n_1, n_2, n_3, \ldots, n_N>_N$. These N-tuples can be grouped into "ordered parity sets" such as $[odd, odd, \ldots, even, odd, even, even]_N$ defined in the main text in Section 10.2. The outcome sets of primary interest are those with n *even* numbers in any order. They are the "unordered parity sets," written "(n, N)":

(n, N) = Union of all ordered parity sets $[parity, \ldots., parity]_N$ with exactly n *even*.

where *parity* is either *even* or *odd*. The following is to be shown:

Theorem

No label permutation can map the unordered parity set (n, N) onto (m, N), for all $0 \leq n \leq N$, excepting the trivial case of $n = m$, implemented by an identity map on labels; and the case of $n = N - m$, implemented by a label permutation that switches all odd with all even numbers.

Proof

[192] There is an uncountable infinity of possible distributions of *like* and *unlike* over the countable infinity of pocket universes. The case in the main text occupies all of them excepting a countable infinity of exceptions that arise in universes finitely many *like* pocket universes, or in universes with finitely many *unlike* universes.

The case of $n = 0$ and $0 < m < N$ has been shown in Section 10.3. Switching "even" for "odd" in that demonstration shows the case of $n = N$ and $0 < m < N$. Here we need only consider $0 < n, m < N$ in the theorem.

Assume for purposes of a *reductio* that there exists a label permutation f that maps the N-tuple $<n_1, n_2, n_3, \ldots, n_N>_N$ to $<f(n_1), f(n_2), f(n_3), \ldots, f(n_N)>_N$ such that unordered parity set (n, N) is mapped onto (m, N), where n does not equal N-m.

It may be the case that a label permutation maps every member of some *ordered* parity set of (n, N) onto elements of the same ordered parity set of (m, N). The mapping is "onto" so that the image of the ordered parity set of (n, N) coincides with the ordered parity set of (m, N). We shall say that the label permutation respects ordered parity sets just if this last property is true for every ordered parity set of (n, N).

There are $N!/(n!(N\text{-}n)!)$ ordered parity sets that are subsets of (n, N); and $N!/(m!(N\text{-}m)!)$ ordered parity sets that are subsets of (m, N). Unless we have the cases excepted in the theorem, $n = m$ or $n = N - m$, these two combinatorial factors are unequal. It follows that there can be no one-one label permutation that respects ordered parity sets for the cases considered in the theorem.

For example, there are four ordered parity sets for (1,4): EOOO, OEOO, OOEO, OOOE, written here in compact notation with "E" = *even* and "O" = *odd*. There are six ordered parity sets for (2,4): EEOO, EOEO, EOOE, OEEO, OOEE. A label permutation that respects ordered parity sets would have to map the members of each of the EEOO, EOEO, … of (2,4) onto distinct ordered parity sets EOOO, OEOO, … of (1,4). Since there are six of the former and four of the latter, this is impossible.

Set n as the number of evens for which $N!/(n!(N\text{-}n)!) > N!/(m!(N\text{-}m)!)$. (There will always be an inequality since the case of equality, $n = N - m$, is excluded.) Since the label permutation cannot respect ordered parity sets, it follows that the permutation must "cross over" the boundaries somewhere of the ordered parity sets. That is, there must be two N-tuples that map as

$$\mathbf{R} = <r_1, r_2, r_3, \ldots, r_N>_N \text{ maps to } f(\mathbf{R}) = <f(r_1), f(r_2), f(r_3), \ldots, f(r_N)>_N$$

$$\mathbf{S} = <s_1, s_2, s_3, \ldots, s_N>_N \text{ maps to } f(\mathbf{S}) = <f(s_1), f(s_2), f(s_3), \ldots, f(s_N)>_N$$

where $f(\mathbf{R})$ and $f(\mathbf{S})$ belong to the same ordered parity set of (m, N), but $\mathbf{R}$ and $\mathbf{S}$ belong to different ordered parity sets of (n, N).

To proceed, we form a new N-tuple $\mathbf{T} = <t_1, t_2, t_3, \ldots, t_N>_N$ by the rule

$t_i = r_i$ if r_i is even; or if both r_i and s_i are odd.

$= s_i$ if r_i is odd and s_i is even.

Each of **R** and **S** have n even numbers in their tuples. However the positioning of the even numbers in their N-tuples must be different somewhere since **R** and **S** come from different ordered parity sets. The definition of **T** is designed to collect all the even numbers from **R** and **S** such that **T** has at least one more even number than **R** and **S**. For example, if **R** = <1, 1, 2, 2> and **S** = <1, 2, 1, 2>, then **T** = <1, 2, 2, 2>. That is, **T** belongs to an unordered parity set, (n', N), where $n' > n$.

The label permutation f maps **T** as

$$\mathbf{T} = \langle t_1, t_2, t_3, \ldots, t_N \rangle_N \text{ maps to } f(\mathbf{T}) = \langle f(t_1), f(t_2), f(t_3), \ldots, f(n_N) \rangle_N$$

Each $f(t_i)$ is either $f(r_i)$ or $f(s_i)$. Since $f(\mathbf{R})$ and $f(\mathbf{S})$ both are members of the same ordered parity set of (m, N), it follows that $f(\mathbf{T})$ is a member of the same ordered parity set (m, N). That is, the label permutation f, maps an N-tuple **T** in (n', N), where $n' > n$, to an N-tuple $f(\mathbf{T})$ in (m, N). Since a label permutation is invertible, it follows that there is no N-tuple in (n, N) that the label permutation maps to $f(\mathbf{T})$. This mapping of **T** contradicts the initial assumption that the label permutation maps (n, N) to (m, N) and completes the *reductio* needed to establish the theorem.

References

Bartha, Paul (2004) "Countable Additivity and the de Finetti Lottery," *The British Journal for the Philosophy of Science*, **55**, pp. 301-321.

Benci, Vieri, Horsten, Leon and Wenmackers, Sylvia (2013) "Non-Archimedean Probability," *Milan Journal of Mathematics,* **81**, pp. 121-51.

Blackwell, David and Diaconis, Persi (1996) "A Non-Measurable Tail Set," pp. 1-5 in T. S Ferguson, L. S. Shapley and J. B. McQueen, eds., *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell.* Hayaward, CA: Institute of Mathematical Statistics.

de Finetti, Bruno (1974) *Theory of Probability: A Critical Introductory Treatment*. Vol. 1. Chichester: John Wiley & Sons.

de Finetti, Bruno (1972) *Probability, Induction and Statistics*. London: John Wiley & Sons.

Guth, Alan (2007) "Eternal Inflation and its Implications," https://arxiv.org/abs/hep-th/0702178v1

Jaynes, Edwin T. (2003) *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

Kadane, Joseph B,. Schervish, Mark J. and Seidenfeld, Teddy (1986) "Statistical Implications of Finitely Additive Probability" in *Bayesian Inference and Decision Techniques.* Amsterdam: Elsevier Science Publishers, pp. 59-76.

Kolmogorv, Andrey N. (1950) *Foundations of the Theory of Probability.* Trans. N. Morrison. New York: Chelsea.

Norton, John D. (2010)"Deductively Definable Logics of Induction." *Journal of Philosophical Logic*, 39 (2010), pp. 617-654.

Norton, John D. (2011) "Challenges to Bayesian Confirmation Theory," *Philosophy of Statistics, Vol. 7: Handbook of the Philosophy of Science.* Prasanta S. Bandyopadhyay and Malcolm R. Forster (eds.) Elsevier.

Norton, John D. (2018) "How to Build an Infinite Lottery Machine," *European Journal for Philosophy of Science*. **8**, pp. 71-95.

Norton, John D. and Pruss, Alexander R. "Correction to John D. Norton 'How to Build an Infinite Lottery Machine,'" *European Journal for Philosophy of Science*. **8**, pp. 143-44.

Norton, John D. (manuscript) "Eternal Inflation: When Probabilities Fail," Prepared for special edition "Reasoning in Physics," *Synthese*, eds. Ben Eva and Stephan Hartmann. http:www.pitt.edu/~jdnorton/homepage/cv.html or http://philsci-archive.pitt.edu/14401/

Norton, John D. (manuscript a)"How NOT to Build an Infinite Lottery Machine."

Pruss, Alexander R. (2014) "Infinitesimals are Too Small for Countably Infinite Fair Lotteries," *Synthese*, **19**, pp. 1051-57.

Shafer, Glenn (2008) "The Game-Theoretic Framework for Probability," Ch. 1 (pp. 3-15) in B. Bouchon-Meunier, C. Marsala, M. Rifqi and R. R. Yager, *Uncertainty and Intelligent Information Systems.* Singapore: World Scientific.

Steinhardt, Paul J. (2011) "The Inflation Debate: Is the Theory as the Heart of Modern Cosmology Deeply Flawed?" *Scientific American*, April 2011, pp. 36-43.

Weintraub, Ruth (2008) "How Probable Is an Infinite Sequence of Heads? A Reply to Williamson," *Analysis*, **68.3**, pp. 247-50.

Wenmackers, Sylvia and Horsten, Leon and (2013) "Fair Infinite Lotteries," *Synthese*, **190**, pp. 37-61.

Williamson, Timothy (2007) "How Probable is an Infinite Sequence of Heads," *Analysis*, **67.3**, pp.173-80

November 8, 9, 2017; June 10, 21, November 8, 2018; January 19, 2019.

Chapter 14

Uncountable Problems[193]

1. Introduction

The previous chapter examined the inductive logic applicable to an infinite lottery machine. Such a machine generates a countably infinite set of outcomes, that is, there are as many outcomes as natural numbers, 1, 2, 3, … We found there that, if the lottery machine is to operate without favoring any particular outcome, the inductive logic native to the system is not probabilistic. A countably infinite set is the smallest in the hierarchy of infinities. The next routinely considered is a continuum-sized set, such as given by the set of all real numbers or even just by the set of all real numbers in some interval, from, say, 0 to 1.

It is easy to fall into thinking that the problems of inductive inference with countably infinite sets do not arise for outcome sets of continuum size. For a familiar structure in probability theory is the uniform distribution of probabilities over some interval of real numbers. One might think that this probability distribution provides a logic that treats each outcome in a continuum-sized set equally, thereby doing what no probability distribution could do for a countably infinite set. That would be a mistake. A continuum-sized set is literally infinitely more complicated than a countably infinite set. If we simply ask that each outcome in a continuum-sized set be treated equally in the inductive logic, then just about every problem that arose with the countably infinite case reappears; and then more.

This chapter will explore what sorts of inductive logics can implement uniformity of chance over an outcome set of continuum size. The notion of uniformity used is label independence, as already developed in the previous chapter. To start, we will presume the outcome set is "bare," that is, it has no further structure beyond its continuum size. Then, in Section 2 below, we shall see that label independence imposes on it an inductive logic something like the infinite lottery machine inductive logic, but with more sectors. This is an unfamiliar logic, remote from a probabilistic logic.

---

[193] My thanks to Jeremy Butterfield for a close reading of this chapter that led to many corrections.

If we seek a sense of uniformity of chance compatible with a probabilistic logic, we must weaken the requirement of label independence. It will be weakened in successive sections in three stages. In Section 3, the unrestricted requirement of label independence is weakened by requiring that the independence holds only for permutations that preserve a σ-field of subsets of a continuum-sized outcome set. This is a natural first step, since probability measures in continuum sized outcome sets are standardly only defined over such subsets. We will find that this weakening is insufficient. A probability measure fails to conform with the weakened requirement of label independence. The failure is not remedied by a further weakening that only allows permutations that are involutions. The applicable logic turns out to be akin to that of the completely neutral support of Chapter 9.

In Section 4, label independence will be further weakened by assuming that the continuum outcome set has its own metrical structure, commonly the metrical geometry of a space. The permutations of label independence are restricted to those that preserve areas or volumes of this metrical geometry. This weakened version of label independence is, finally, compatible with a probabilistic logic: it is one that matches probabilities with the space's areas or volumes.

The success, however, proves limited. For if the metrical space is infinite in area or volume, a probabilistic logic cannot provide uniformity of chances. It is easy to see that a metrically adapted label independence requires that this uniformity be expressed by the same inductive logic that applies to the infinite lottery machine. This inductive logic is the one that applies to the stochastic process of continuous creation of matter in Bondi, Gold and Hoyle's steady state cosmology. Its application to this case is teased out in enough detail to return some curious results.

That this last inductive logic is applicable is demonstrated by decomposing the space into infinitely many parts. The parts are then reassembled in a way that respects the background metrical structure of the space, but precludes an additive measure. This construction is one of the simplest of a corner of mathematics that explores "paradoxical decompositions." This literature is introduced in Section 5. It has explored more thoroughly the difficulties faced when we seek to use additive measures to gauge the size of sets in a metrical space. The construction of Section 4 employed a decomposition into infinitely many parts. If our space had hyperbolic geometry, then a remarkable construction reported by Wagon (1994) shows that similar results can be achieved by decomposing the space into just three parts each of infinite measure.

This literature in paradoxical decompositions is the locus of nonmeasurable sets. These are sets in a metrical space to which no area or volume can be assigned consistently. While the difficulties for probability measures have so far arisen only in metrical spaces of infinite area or volume, these nonmeasurable sets become problematic for probability measures that match the

areas and volumes of spaces with finite total area of volume. For such a probability measure will fail to assign a value to these nonmeasurable sets. Since these nonmeasurable sets impose a fundamental limitation on the use of probability measures in such spaces, they will be pursued in the remainder of the chapter.

Section 6 will review the simplest example, a Vitali set. Since a Vitali set is metrically nonmeasurable, it is beyond the reach of a probability measure adapted to the spatial metric. Instead, the chance that some outcome of a random process will be found in a Vitali set is shown to follow a familiar inductive logic, that of the infinite lottery machine. This section also discusses the awkwardness that nonmeasurable sets are not constructible by the means normally employed in set theory. Rather their existence is posited by the axiom of choice.

Finally in Section 7, I recount a nonmeasurable set described by Blackwell and Diaconis (1996) that comes closer to the sorts of systems commonly treated in accounts of inductive inference. It is a probabilistically nonmeasurable outcome set that arises with infinitely many coin tosses. In Section 8, I show that there is a weak inductive logic native to the example that I call an "ultrafilter logic."

Overall, this investigation shows that, in many cases for a continuum sized outcome set, a probabilistic logic fails to apply. Other, non-probabilistic logics do apply locally to the specific problem posed. To recount them, they appear as:

Section 3.6, Section 4.2, Section 6.2: variations on an infinite lottery machine logic.

Section 8: an ultrafilter logic.

2. The Inductive Logic of Uniform Chances in a Bare Continuum

How might an inductive logic provide equal support or equal chances to every outcome in a space of continuum size? To answer, we need to specify the applicable notion of equality or uniformity of chances. That condition was developed in the previous chapter. An infinite lottery machine selects among a countable infinity of numbers fairly, that is, without favoring any. Each of the infinity of outcomes was assigned a unique number label. The fairness of the lottery was expressed in the condition:[194]

[194] Here and below, a permutation is a one-one map on the label set or, correspondingly, on the outcome set. In the previous chapter, these sets were countable. In conformity with modern usage, the term "permutation" will continue to be used when the label of outcome set is continuum sized. The term is synonymous with bijection.

Label independence

All true statements pertinent to the chances of different outcomes remain true when the labels are arbitrarily permuted.

That individual outcomes have equal chance is secured through propositions like:

Outcomes numbered "37" and "18" have the same chance.

The statement remains true no matter how we redistribute number labels across the outcomes. This indifference to the labels assigned to individual outcomes can only come about if all outcomes have the same chance. It is otherwise with statements like:

Outcome number "37" has greater chance than outcome number "18."

This statement cannot remain true under a relabeling that switches labels "37" and "18," assuming that the relation of "greater chance" is asymmetric.

The same applies to sets of outcomes:

The odd numbered set of outcomes has the same chance as the even numbered set of outcomes.

This statement remains true no matter how we may permute the number labels over the outcomes. Once again, this indifference of the sets to the numbers that label their elements can only come about if the two sets have the same chance. From similar statements, it follows that two sets of outcomes have the same chance just in case there is a permutation of the number labels that reassigns the numbers labeling the first set to the second set.

We now apply label independence to an outcome set of continuum size. We saw in the previous chapter that the chance values assigned to sets of outcomes of an infinite lottery machine drawing were divided into two sectors, the finite sector and the infinite sector. Replicating the procedure of the previous chapter for the new case of a continuum-sized outcome set, we find a similar, but more complicated structure, with three sectors. In the continuum case, the chance of an outcome in various outcome sets has the indicated values and associated informal interpretation:

Finite set of outcomes of size n:

A countable infinity of values, $V(n)$, $n = 1, 2, 3, \ldots$; "very unlikely."

Countably infinite set of outcomes:

one value only, V(countably infinite); "unlikely."

Continuum-sized infinite set of outcomes:

For an outcome set of continuum size and whose complement is continuum sized,
one value only, V(continuum-co-continuum); "as likely as not."

For an outcome set of continuum size and whose complement is countably infinite,
one value only, V(continuum-co-countable); "likely."

For an outcome set of continuum size and whose complement is finite,

V(continuum-co-finite *n*), *n* = 1, 2, 3, …; "very likely."

The strength of support grows as we move down this list. The distance between the sectors is very great since we step up the hierarchy of infinites. We could, presumably, find many results that match those of the infinite lottery machine logic and many more that are not in it, because of its extra structure. However I will pass over this exercise. What matters for our purposes is that the fullest implementation of uniformity in a continuum-sized outcome set leads to a logic that is quite different from a probabilistic logic.

3. Uniformity over a σ-Field of Outcomes

3.1 A Uniform Probability Distribution

The logic of the last section is very different from a probabilistic logic. We were driven to this logic by the requirement of label independence. If we are to find conditions more conducive to a probabilistic logic, we will need to weaken this requirement. To map a pathway for the weakening, we need to see our goal: a uniform probability distribution over a continuum-sized outcome set. Take the especially hospitable case[195] of outcomes labeled by real numbers in the interval [0,1], that is the set of real numbers x, such that $0 \leq x \leq 1$. The uniform probability distribution over this interval is derived from a probability density function

$$p(x) = 1 \tag{1}$$

and it is plotted in Figure 1.

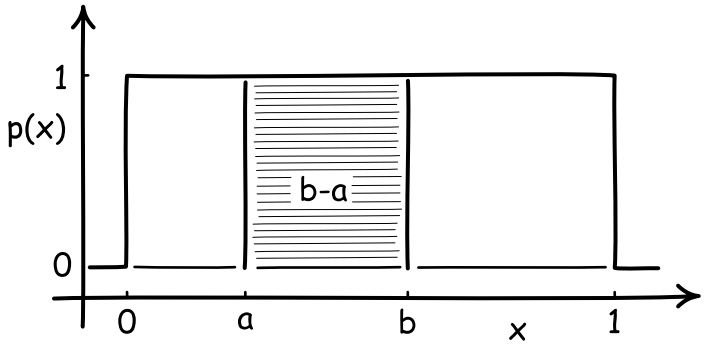


Figure 1. Uniform Probability distribution

[195] It is hospitable since, otherwise, if either end of the interval extends to infinity, a uniform non-zero probability density over the interval integrates to an infinite probability over the whole interval.

We extract probabilities from this probability density for sets of outcomes by computing the corresponding areas under the curve. The probability of an outcome labeled by a real number in the interval [*a*, *b*], where $0 \le a \le b \le 1$, is the area shown in the figure and, of course, is equal to *b*-*a*.

This distribution certainly *looks* like it is choosing without favor among the continuum sized outcome labeled by [0,1]. The curve in Figure 1 is flat. It is also free of a problem facing a uniform probability distribution over a countably infinite outcome space: there is no countably additive, uniform probability distribution over the set. For such a distribution, each outcome would have to be assigned the same probability. If that value is zero, then their countably infinite sum is also zero, in contradiction with the requirement that the probabilities of all mutually exclusive outcomes must sum to unity. In contrast, the probability density (1) can assign zero probability to each of its continuum many outcomes without a corresponding difficulty. The summation of an uncountable infinity of zeroes is not a well-defined operation in standard probability theory.

In spite of these encouraging signs, the uniform probability distribution fails to implement the requirement of label independence. Take statements such as

(Eq) The probability of events labeled by real numbers in [0, 0.5] is the same as the probability of events labeled by real numbers in [0.5, 1],

Since the permutations admissible under label independence are entirely unrestricted and can scatter the labels about in all imaginable ways, it is easy to see that this and many other statements like it fail to remain true when the number labels permuted. Some restriction on the permutations is needed if label independence is to apply.

3.2 The σ-Field

One of the founding results of modern measure theory is that an additive measure, such as a probability measure, cannot assign a measure to all subsets of points in a space if the space is sufficiently large. Then there are many nonmeasurable sets. In Section 6 below, we shall see the standard example that arises in the interval [0,1] of real numbers, a Vitali set. As a result, probabilities can be defined only for a preferred subset of all the subsets of real numbers in [0,1]. The resulting restriction on the scope of probability measures has been built into the modern mathematical formalism from the outset. Kolmogorov (1950), the *locus classicus* of the modern tradition, introduces the distinction in his definitions. A probability measure is defined in the context of a set of "elementary events." (p. 2, Ch. II) It is, for example, the set of outcomes labeled by real numbers in [0,1]. However a probability is not automatically defined for all subsets of this set. Rather, at the outset, probabilities are defined only for some of these subsets. These are the "random events" that form a field or algebra of sets. That is, the field or algebra is, by definition, closed under the finite union, finite intersection and complement of its members.

When the set of elementary events is infinite, the fields or algebras are required to be σ-fields or σ-algebras. That is, they are closed under countably infinite unions and intersections.

Since a probability measure can assign probabilities only to some of the subsets of elementary event labeled by real numbers in [0,1], those sets have to be identified if the probability measure is to be adequately defined. The standard procedure is to work backwards from those probabilities that we cannot forego. In forming the probability distribution associated with (1), we expect that, whatever else, the probability assigned to all intervals of the form [*a*,*b*] above is *b*-*a*. So we include in the σ-field all intervals of the closed form [*a*,*b*] as well as half-open [*a*,*b*), (*a*,*b*] and and open (*a*,*b*).[196] We then require that the σ-field associated with the uniform distribution be one that contains all these intervals and is closed under all countable unions and intersections. It is not obvious that such a field should exist or, if so, that it is unique. Both are assured by the Extension Theorem (Kolmogorov, 1950, p. 17).[197]

3.3 σ-Field Adaptation

The uniform distribution does not assign probabilities to all subsets of the elementary events labeled by real numbers in [0,1]. It follows that the truth of statements concerning subsets of elementary events cannot be preserved under an arbitrary permutation of the numbering of the elementary events used in the statement. The permutation may take a set for which a probability is defined to one that is nonmeasurable. What is a true statement for the original set about its probability may fail to be true when those same number labels are applied to a nonmeasurable set, for the latter set has no probability. Thus the subsets in the σ-field are favored in the sense that a probability is defined for them only. Label independence fails.

If a probability density (1) is to conform with label independence, we need to weaken label independence. A first step in this weakening is to restrict the permutations so that they only map sets of events in the σ-field to sets of events in the σ-field.

σ-Field Adapted Label independence

All true statements pertinent to the chances of different outcomes remain true when the labels are permuted by all permutations that preserve the sets of the σ-field.

---

[196] By the usual convention [*a*,*b*) contains all *x* for which $a \le x < b$, *etc.*

[197] See Rosenthal (2006, Ch. 2) for a more expansive introduction to this result of great foundational importance.

A consequence is that sets of elementary events labeled by some open, half-open or closed interval of real numbers, always remain labeled by such intervals under all permutations to be considered.

3.4 Failure

While σ-field adaptation is a necessary adaptation if the uniform probability density (1) is to be compatible with label independence, it turns out not to be sufficient. The uniform probability density (1) still does not conform with the weakened requirement. The permutations of the weakened requirement are continuous functions on x that invertibly map the interval [0,1] back to [0,1]. The condition of invertibility is essential. Otherwise the function would be redistributing the number labels in such a way that one elementary event is assigned more than one new number label. There are of course *very* many such invertible functions. Label independence requires that all of them leave the probability distribution unchanged. The trouble is that virtually all of them do not leave it unchanged.

One example illustrates the general behavior. We start with two events consisting of elementary events labeled by real numbers x in the intervals [0, 0.5] and in [0.5, 1.0]. The probability density (1) assigns equal probability of 0.5 to each event. As we saw above in (Eq), label independence requires that this statement remain true when we permute the numbers that label the elementary events. We use an invertible, continuous function to carry out the permutation. Let that function map each real number x in [0,1] to a new value y in [0,1] according to:

$$y = f(x) = \sqrt{1 - x^2} \tag{2}$$

To use the function as a permutation of labels, we take the elementary event that was originally labeled y and assign it the new real number label x. The number x is "carried along" by the function. Under this permutation, as shown in Figure 2 on the left, the two events originally labeled with real numbers in the intervals [0, 0.5] and in [0.5, 1.0] are mapped to the events originally labeled with real numbers in the intervals [0.8666, 1] and in [0, 0.8666], respectively. These last events are now assigned the new, carried along number labels in the intervals [0, 0.5] and in [0.5, 1.0] respectively.

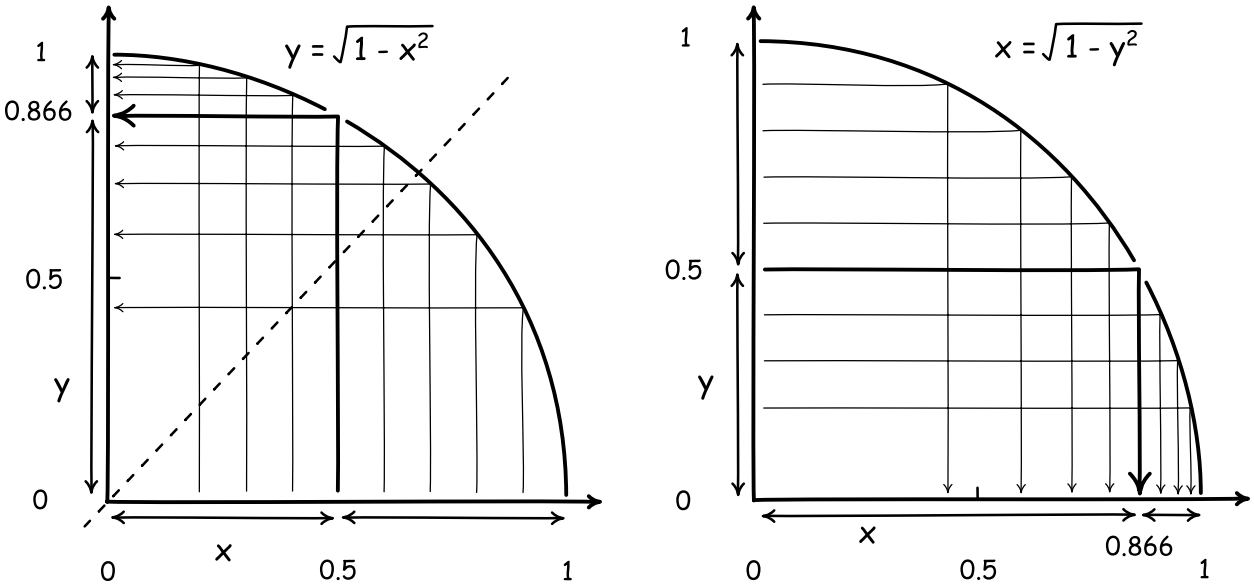


Figure 2. Uniformity of Probability Not Preserved under Permutation

These two intervals have unequal probabilities under the probability density (1): the probabilities are 1-0.8666 = 0.1333 and 0.8666 respectively. The permutation (2), however, assigns them new number labels in the intervals [0, 0.5] and in [0.5, 1.0] respectively. Statement (Eq) is false if we use the permuted number labels. Label independence is violated.

What would it take for label independence to be preserved? The condition needed is simple. A permutation like (2) can "carry along" the probabilities assigned to the origin set to the destination set. The key condition is that this carried along probability must match that originally assigned to the destination set. That is what failed for the permutation (2) above.

We can give this condition a general formulation as follows. The probability assigned to some small interval x to $x+dx$ is approximated by $p(x)dx$. Under the permutation, the number labels in the interval x to $x+dx$ are now reassigned to events originally labeled by numbers in the interval y to $y+dy$. These events were originally assigned a probability approximated by $p(y)dy$. The condition that this original probability and the carried along probability agree is:

$p(y)dy = p(x)dx$, in case dx and dy have the same sign; or

$p(y)dy = -p(x)dx$, in case they differ in sign.

Taking the limit of dx and dy to zero, we have[198]

[198] The absolute norm in $|dx/dy|$ keeps $p(y)$ positive in both cases above. Note that $|dx/dy|$ is either always positive or always negative, since the conditions of continuity and invertibility requires $x(y)$ to be everywhere increasing or everywhere decreasing.

$$p(y) = p(x)\left|\frac{dx}{dy}\right| \tag{3}$$

Here $p(y)$ is the new probability density induced by the carrying along of the original probability density by the permutation, expressed in the original number labels.

A short calculation shows that the carried along probability density of (3) when computed for the permutation (2) and the source probability density (1) is

$$p(y) = \frac{y}{\sqrt{1-y^2}}$$

This induced probability density is no longer uniform over its argument, y. Thus, statement (Eq) will turn from true to false under permutation (3), violating label independence.

These last considerations lead directly to the general condition that must be satisfied by all permutations if label independence is to be preserved. It is simply

$$p(y) = p(x) \tag{4}$$

Comparing (3) and (4), we see that this equality of probability densities can only be secured if $\left|\frac{dx}{dy}\right| = 1$. This last condition is violated by almost every permutation of the number labels. For $y(x)$ a continuous, differentiable function of x, it is satisfied only by two cases $y=x$ and $y=1-x$.

The outcome is that the probability density (1) does not distribute the chances over a continuum set of elementary events indifferently, in the sense captured by the requirement of σ-field adapted label independence. For there are just two "right" ways to apply the numbering. That suggests that there is more structure hidden in the example than merely a continuum-sized set and its σ-field of subsets.

3.5 Involutions

Before proceeding, we should visit briefly with a tempting escape from the problems just developed. Might we propose that some x is the "right" labeling to use; that it has some property intrinsic to the problem; and that a permutation y is somehow ill-suited, since it takes us to another labeling that lacks the property?

The particular function (2) above was chosen with just this possibility in mind. For it is an involution, which means it has the characteristic property that a double application of the function returns the original argument. That is $x = f(f(x))$. This means that there is a perfect symmetry in the relationship between x and y. Exactly the same functional form as (2) takes us back from y to x:

$$x = f(y) = \sqrt{1-y^2}$$

Figure 2 on the right shows the inverse mapping of the interval y in [0, 0.5] to the interval x in [0.8666, 1]. The graph of an involution has the distinctive property of symmetry around the diagonal axis of the dashed line, $y=x$, shown in Figure 2. Clearly there are very many more involutions since this symmetry is all that is required.

The use of an involution responds directly to the idea that some labeling might be the "right" one. For it follows from the symmetry that, for any property that x bears with respect to y, there is a corresponding property that y bears with respect to x. Thus any decision that one of x or y is somehow favored cannot be derived from properties intrinsic to the parameters. For whatever case we make for favoring x based on the intrinsic properties of x, there is a corresponding case that can be made for y. What results is a further weakening of label independence:

σ-Field, Involution Adapted Label independence

All true statements pertinent to the chances of different outcomes remain true when the labels are permuted by all involutions that preserve the sets of the σ-field.

The existence of many involutions then shows that this proposal for escape fails. There is no intrinsic property of one labeling x that distinguishes it. A preference for x must be imposed by us externally by fiat. Such an external imposition breaks label independence. We may, however, find an external basis for the imposition, as we shall see in Section 4 below.

3.6 The Natural Inductive Logic on [0,1]

What if we forego the idea that inductive support must be represented probabilistically?[199] What inductive logic over the intervals of [0,1] conforms with these two weakened requirements of label independence? Even with these weakenings, it turns out that the only inductive logic admissible is akin to the infinite lottery machine logic.[200] The logic assigns the same neutral value I to any interval[201] (a, b), where $0 \le a < b \le 1$ in [0,1], excepting $(a, b)=(0,1)$:

$$\text{support}((a, b)) = I \qquad (5)$$

That this is the unique inductive logic conforming with the weakened label independence follows from two statements:

---

[199] For comparison, the transformational behavior of probability measures under involutions has been explored in greater detail in Norton (2008).

[200] As with the infinite lottery machine logic, different supports are assigned to sets of outcomes of finite size or countably infinite size.

[201] For simplicity of exposition, I consider only open intervals (a,b). The same results apply to half open and closed intervals.

(i) In some real number labeling of the elementary events, all intervals (a, b) of equal size $|b-a|$ accrue the same support: support((0, 0.1)) = support((0.1, 0.2)) = support((0.2, 0.3)) =… etc.

(ii) For any $0<a<1$, $0<b<1$, there exists an involution on [0,1] that maps the interval (0, a) to the interval (b, 1). By label independence, they have the same support.[202]

Take any two intervals in the scope of (5): (a, b) and (c, d). By (i), they have the same support as (0, b-a) and as (1-(d-c), 1), respectively. Through (ii), label invariance entails that the intervals (0, b-a) and (1-(d-c), 1) have equal support. Hence all intervals in (5) have the same support, which we label as "I".

In this analysis, (i) is an assumption that amounts to requiring that there is at least some numbering that is naturally adapted to the equalities of support.[203] Statement (ii) is derived from the properties of involutions. Readers who are satisfied that the statement is correct might like to skip over the details that follow.

Statement (ii) can be demonstrated though two families of involutions that are jointly dense in the unit square, as displayed in Figure 3.

[202] For the statement "Events labeled by (0, a) have support X." must be true also of events labeled by (b, 1) since this second set of elementary events can be relabeled through the involution by numbers in (0, a).

[203] Almost all of (5) can be derived with constructions like those of (ii). However no continuous involution can map all the equalities needed. None can map, say (0,0.5) to (0.1, 0.6). Something like assumption (i) is needed to complete derivation of (5).

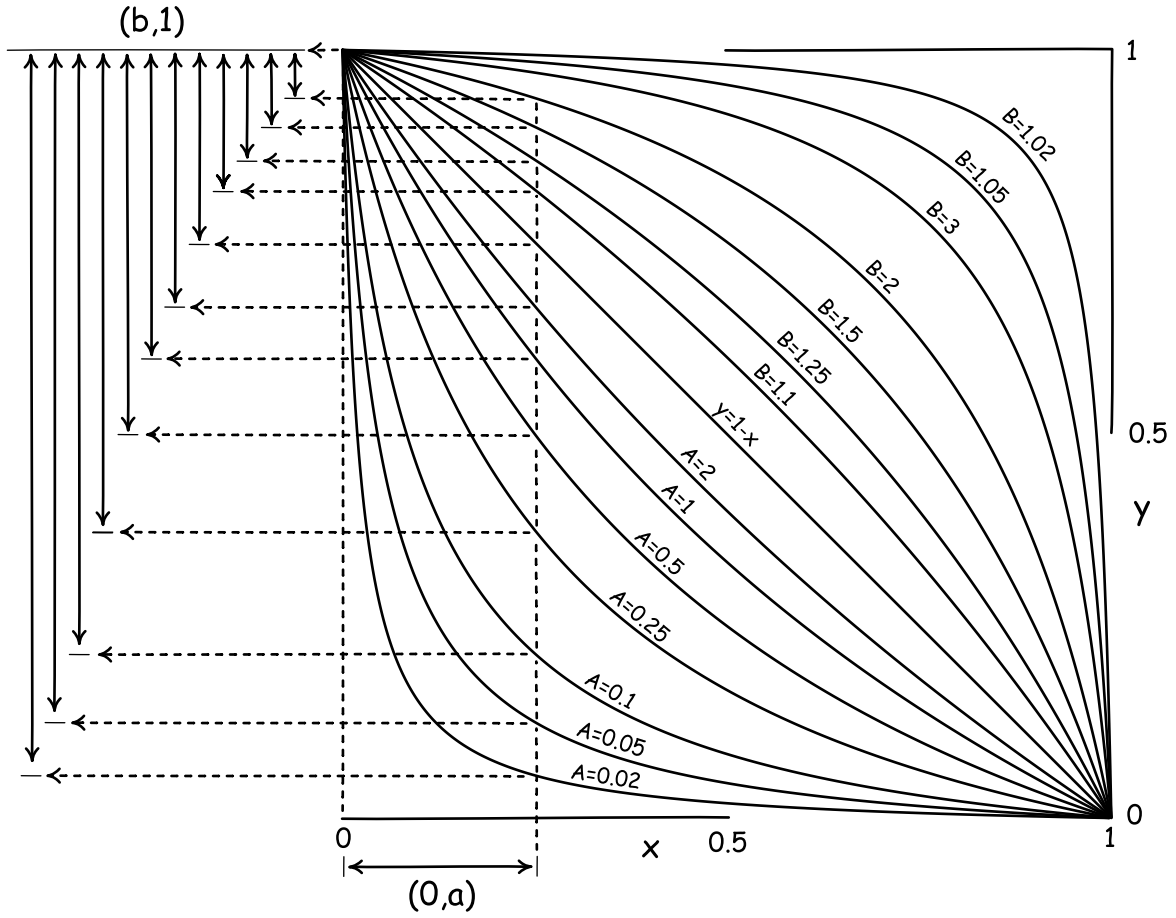


Figure 3. Two Families of Involutions on [0,1]

These involutions derive from the formulae:

$$y = \frac{A^2 + A}{x + A} - A \text{ , all } A>0, \text{ and } y = \frac{B^2 - B}{x - B} + B \text{ , all } B>1.$$

That they are involutions can be seen by rearranging each to give

$$(x+A)(y+A) = A^2+A \quad \text{and} \quad (x-B)(y-B) = B^2-B$$

Since x and y enter symmetrically into these rearranged formulae, it follows that, in each case, y has the same functional dependency on x as x does on y.

Consider the interval $(0, a)$ of (ii) for any $0<a<1$. It follows from the density of the involutions that there always exists one involution that maps $(0, a)$ to $(b,1)$ for any $0<b<1$. As Figure 3 shows, the A family of involutions, maps $(0, a)$ to $(b,1)$, where $0<b<1-a$. The B family of involutions maps $(0, a)$ to $(b,1)$, where $1-a<b<1$. The involution $y=1-x$, intermediate between the two families, covers the intermediate case of $b=1-a$, in which $(0, a)$ is mapped to $(1-a,1)$

4. Uniformity from Metrical Lengths, Areas and Volumes

4.1 Metrical Adaptation

If the uniform probability density (1) is to conform with label independence, we will need to weaken the requirement still further. In many important cases, a continuum-sized outcome set has further structure: a spatial metrical structure, to which the probability distribution is required to be adapted. Metrical structure assigns lengths in one-dimensional continua, areas in two-dimensional continua and volumes in three and higher dimensional continua.

When metrical structure is present, we often require adaptation of the chances to it. That means that sets of outcomes that are equal in length, area or volume have equal chances. These cases arise when, in accord with the material theory of induction, background facts warrant it. Here are some examples. A very long steel beam has defects randomly distributed through it. If it is stressed uniformly, this fact ensures that fracture is equally probable in portions of equal length. A dart is thrown at a dart board. Assuming disturbances from sufficiently many random factors, it is equally likely to strike regions of equal area. Under the physical principle of the maximization of thermodynamic entropy, a molecule of an ideal gas, free of external fields, is equally likely to be in parts of the containing vessel of equal volume.

This adaptation of chances to metrical structure can be implemented by restricting the set of the permutations in the requirement of label independence:

> *Metrically Adapted Label independence*
>
> All true statements pertinent to the chances of different outcomes remain true when the labels are permuted by all permutations that preserve the metrical measures of outcome sets.

A permutation preserves metrical measure just when labels identifying some metrically measurable set of outcomes are permuted to a new set of outcomes that has exactly the same metrical measure. In generic cases, such a permutation can switch any region with any other of the same metrical measure. In these cases, it follows from this weakened version of label independence that the chance of some outcome depends only on the length, area or volume associated with it. The statement "outcome A has chance such and such" must remain true when the labels identifying outcome A are relocated to any other part of the space under a metrical measure preserving permutation. The relocated outcome must have the same length, area or volume as the original, no matter how they may differ in their other properties.

These metrical measure-preserving permutations are allowed to preserve metrical measure patchwise. That is, they can divide up the space into patches and rearrange them, as long as the rearrangement preserves the measure of each patch. This last patchwise construction

is a mainstay of traditional geometry. It is the standard method of proving equality of areas and volumes. Here is a rather pretty example that uses area-preserving permutations to prove Pythagoras's theorem. It is due to Rufus Isaacs (1975). The square on the left of Figure 4 shows four right angles triangles, each with sides of length *a*, *b* and hypotenuse *c*. They enclose a central square of area c^2, which is the "square on the hypotenuse" of Pythagoras' theorem. The area associated with this square is redistributed under a permutation shown in two steps in the central two squares. First two triangles are permuted so that their positions are moved down the figure. Then two of the triangles are moved together up the figure. The result, shown in the square on the right, is that the region forming the square of area c^2 has been relocated to a new region consisting of two squares, one of area a^2 and another of area b^2. These are the "squares on the other two sides." They are shown by this construction to be equal to the square on the hypotenuse.

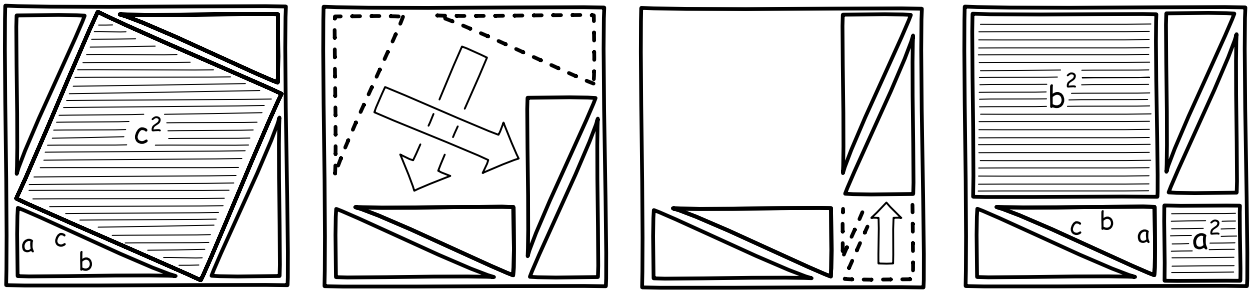


Figure 4. A Metric Preserving Permutation Proves Pythagoras' Theorem

If the chances are expressed by probabilities, metrically adapted label independence requires equal lengths, areas and volumes to be equally probable. Familiar cases work just as we would expect. These successful applications of the probability calculus arrive easily. It is because an additive metrical structure is already present in the physical assumption that the spatial continua have lengths, areas or volumes native to them. Chances acquire that additive structure upon adaptation to the metrical structure. Disjoint volumes add to give the combined volume, so the chances of outcomes in them add also to give the disjoined chances. Since the total system length, area or volume may have an arbitrary magnitude, all that remains is to normalize the adapted chances to unity to recover probability measures. If the total area of dart board is 144 square inches, then the probability of the dart striking any nominated square inch area 1/144.

4.2 The Infinite Lottery Machine Logic, Again

We can now see which will be the troublesome cases: those in which the lengths, areas or volumes of the total system are infinite. For then normalization over a uniform measure is no longer possible. If the dartboard is infinite in area, then the probability of the dart striking any nominated square inch is $0 = 1/\infty$. Since the infinite area is a countable infinity of unit areas, the chance relations among them turn out to be the same as in the infinite lottery. That is, the requirement of metrically adapted label independence leads us to the same inductive logic as applies to an infinite lottery machine.

An easy way to see this is to continue with the infinite dart board, that is, the example of areas on an infinite Euclidean plane. A process identifies a point in the plane in such a way that its chances conform with metrically adapted label invariance. We can divide this plane into infinitely many tiles of equal, finite area. For convenience, let us pick square tiles. We consider the outcome that the point selected is in one or more of these tiles. Each will have an equal chance. Infinitely many real number pairs label each square uniquely. Since there are a countable infinity of tiles, we can relabel them with a single natural number, 1, 2, 3, …. The resulting relabeling will now conform with the original, unrestricted requirement of label independence. Since the labels are natural numbers, the arguments of the previous chapter apply. The chances of outcomes in various sets of the tiles conform with the infinite lottery logic.

It now follows that all areas consisting of finitely many, n, tiles have the same chance and, as with the infinite lottery, they are assigned the chance value V_n. Since the areas of the tiles are additive, we have the further property of the additivity of these chance values. For all finite m and n,

$$V_{m+n} = V_m + V_n$$

These finite cases can be developed further in obvious ways. The more interesting cases, however, are outcomes in parts of the plane of infinite area. Crudely, under metrical adaptation, we expect trouble, since all infinite areas are equal. Using arguments carried over from the analysis of the infinite lottery machine, we will find that the chances of outcomes in all infinite-co-infinite regions have the same value, called V_∞ in the infinite lottery case.

To see this, divide the infinite plane into four quadrants, I, II, III and IV. We can then reproduce the argument concerning the sets *one*, *two*, *three* and *four* of the infinite lottery machine. We first number the tiles in the quadrant I with the numbers in the set

$$one = \{1, 5, 9, 13, ...\}$$

and then continue for quadrants II, III and IV with the numbers in the sets

$$two = \{2, 6, 10, 14, ...\}$$
$$three = \{3, 7, 11, 15, ...\}$$

$$four = \{4, 8, 12, 16, \ldots\}$$

respectively, as shown on the left in Figure 5.

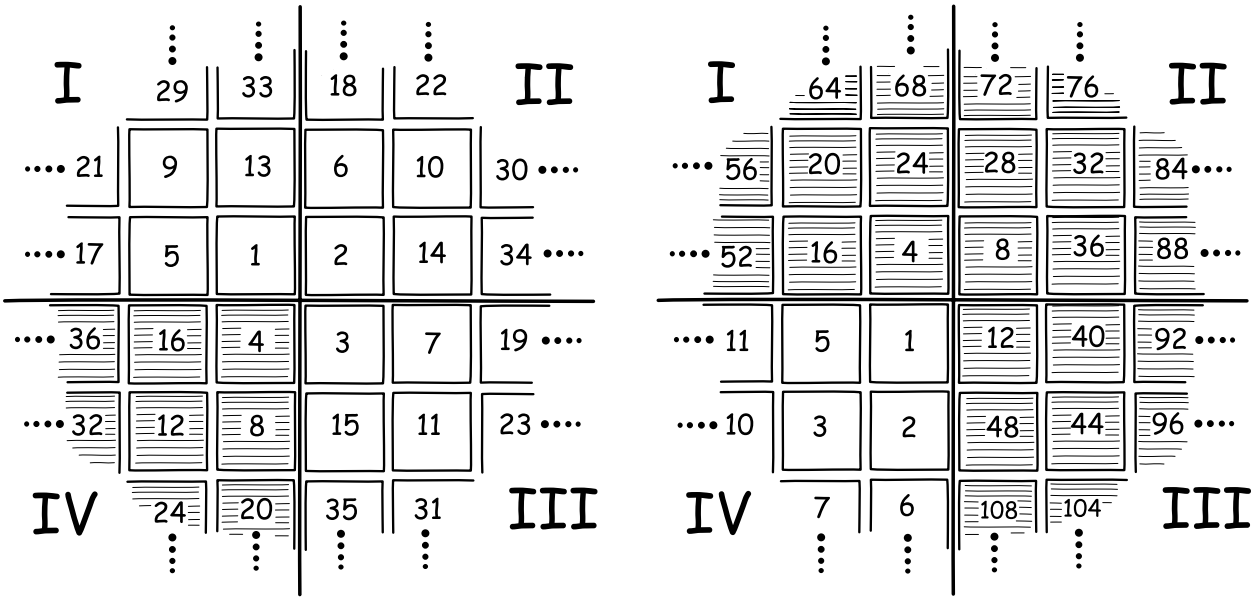


Figure 5. Rearranging Tiles over the Quadrants of an Infinite Plane

Since each quadrant contains a countable infinity of tiles, we can proceed just as we did with the infinite lottery machine. We can rearrange the tiles so that all those in quadrant I fill both quadrants I and III, while those formerly in quadrants II, III and IV fill just quadrants II and IV. Or we can rearrange the tiles so that those in quadrant IV fills quadrants I, II, and III, while those formerly in quadrants I, II and III just fill quadrant IV. This rearrangement is shown on the right in Figure 5. Since the rearrangement of tiles is merely a permutation of the labeling, it preserves chances. With further similar permutations, we can conclude:

$$\begin{aligned} &\text{Ch (I)} = \text{Ch (I or II)} = \text{Ch(I or III)} = \ldots = \text{Ch (III or IV)} \\ &= \text{Ch(I or II or III)} = \ldots = \text{Ch (II or III or IV)} = V_\infty \end{aligned} \tag{6}$$

where "Ch(I)" designates the chance of an outcome in quadrant I.

Since this inductive logic has been elaborated more fully in the previous chapter, there is no need to duplicate the analysis here. Similar manipulations can show that this same inductive logic applies to one-dimensional continua with length and three- and higher dimensional continua with volume, if the chance processes in them conform with metrically adapted label independence. The next section provides an illustration in a science of this logic in a three dimensional space.

4.3 Continuous Creation of Matter in Steady State Cosmology

The steady state cosmology of Bondi, Gold and Hoyle enjoyed considerable attention with its initial formulation of 1948, until it eventually succumbed to several empirical problems. The most notable was an enduring difficulty in explaining naturally the cosmic background radiation observed by Penzias and Wilson in 1964. The cosmology is based on the "perfect cosmological principle." It goes beyond the more familiar cosmological principle in asserting that the universe presents the same average aspect to us not just at all positions in space, but at all times as well.

We know from measurements of the velocities of distant galaxies that the matter of the universe is everywhere expanding. That would normally entail that the average density of matter is everywhere decreasing, so it is lesser at later times. This decrease would violate the perfect cosmological principle. So steady state cosmology posits the continual creation of matter at just the right rate to maintain a constant, average matter density through time. Since ordinary matter is particulate in nature, this continual creation must be a discrete process with particles popping into existence stochastically. In Bondi and Gold's (1948) original proposal, the rate of creation was (p. 256):[204]

> The required rate of creation … can be estimated as at most one particle of proton mass per litre per 10^9 years.

By the time of writing of Bondi (1960), the requisite creation rate was updated with new astronomical measurements of the rate of expansion of the universe. Bondi now estimated it as (1960, p. 143)[205]

> …on an average the mass of a hydrogen atom is created in each litre of volume every 5×10^{11} years.

The difference between creation of a particle of proton mass and of hydrogen atom mass is inconsequential. A hydrogen atom consists of a proton and an electron and the proton comprises roughly 99.9% of the atom's mass.

For our purposes, the delicate question is just what stochastic rules govern the creation of these particles. The theorists ruled out the initially plausible possibility of matter creation within stars. Insufficient newly created matter could escape from stars to form new galaxies. (Bondi and Gold, 1948, p. 266; Bondi, 1960, p. 149). On grounds of simplicity, the theorists proposed creation processes uniformly distributed through space. From Bondi and Gold (1948, p. 268):

---

[204] This corresponds to a mass creation rate of approximately 10^{-43} g/sec cm.^{3} (p. 265).

[205] This corresponds to a mass creation rate of approximately 10^{-46} g/sec cm^3 (p. 143).

> According to this view the probability of creation taking place in any particular four-dimensional element of volume (spatial volume element x element of time) is simply proportional to its (four-dimensional) volume, the factor of proportionality being a function of position. By our argument in 4.1 this factor cannot vary very much from point to point.

From Bondi (1960, p. 151):

> It seems simplest to suppose that the probability of creation in any small four-dimensional element of space-time is simply proportional to its four-dimensional volume.

On the strength of these remarks, we shall proceed in assuming the following stochastic model. In some fixed interval of cosmic time, there is an equal chance of creation of a hydrogen atom in each region of space of the same volume. Creation events are independent of each other.

Bondi and Gold assume that chance in this model can be probabilistic. They are mistaken. Since the space of steady state cosmology is Euclidean and, thus, infinite, this stochastic model conforms with metrically adapted label independence and will be governed by the infinite lottery machine inductive logic. As a result, the process of continual creation that they describe will not proceed quite according to normal expectation.

To explore the application of this logic to continual creation, imagine the Euclidean space of the cosmology divided into two infinite parts, "left" and "right" by some infinite plane. We will ask after the distribution of new particle creation events on the two sides of the plane in the course of a year. Since the average creation rate per unit volume of space is assumed non-zero, infinitely many particles will be created in each side over the year. Is this creation rate the same in both sides? That is, in the long run, are one in two creation events on the left side?

It is tempting to give the quick answer that the rate is infinitely many particles per year in both. Therefore, they are equal. This equality is something less than it seems. It does not support the further conclusion that one in two creation events are, in the long run, on the left side. Take the case in which the rate of particle creation per unit volume per year on the left side is 1,000 times greater than on the right side. Since both volumes are infinite, this case too yields a creation rate of infinitely many particles per year on both sides. Yet we do not expect one in two of them to be in this left side in the long run. It seems that a more refined means of comparing the rates of creation is needed.

In the course of a year, infinitely many particles will be created, but it will be a countable infinity. (There are a countable infinite of equal volumes of space. In each at most a finite number of particles will be created, usually zero or one.) If we track these creation events one by one, we can form the ratio of left side particle creation events to the total number. Among N particle creation events, there will be N_L creation events in the left side.

Since left and right are equally favored, our expectation is that the ratio of N_L/N will stabilize towards one half as we let N go to infinity. This expectation is not supported by the infinite lottery inductive logic. This case is isomorphic to the frequency of even numbers in repeated drawings from an infinite lottery machine. We saw in the previous chapter (§10.7) that the relative frequency of even numbers among all drawn does not stabilize to any definite value.

This result may seem to contradict the symmetry of right and left. Surely half of all creation events must happen on the left in the long run; and half must happen on the right? That expectation depends on the tacit assumption that there is an average in the long run to the fraction of creation events. We now see that there is not. The symmetry of left and right is preserved in the sense that no stable fraction arises in the long run for *both* left and right.

This result arises from tracking creation events in infinite volumes of space. If we restrict our consideration to finite volumes of space, then the normal probabilistic analysis succeeds. Over time, constant mass is preserved on average in each finite volume of space, as required by steady state cosmology.

Finally, as a minor point, this analysis involves a technical complication. It requires an enumeration of the particle creation events in the year by 1, 2, 3, 4, …, N, … so that the limit of the ratio N_L/N can be formed. Such an enumeration is possible since there are only a countable infinity of creation events. However the enumeration must be dictated by a rule that is independent of whether the event is in left region or right region. The simplest such rule is to number the creation events by the their time order. We would number the temporally first event 1, the second 2, and so on. The difficulty is that there may be no first event if the creation times have an accumulation point towards the past. That arises if, for example, the creation events happen at times (in years) 1/100, 1/101, 1/102, 1/103, … There can be multiple such accumulation points. If there are accumulation points towards the future, then the enumeration can never pass them.

I believe the following rule will solve the problem. Divide the year into 1/10ths and assign 1, 2, 3, … to the first event in each 1/10th, if there is one in each 1/10th. Next divide the year in 1/100ths and assign the next numbers to the first unnumbered events in each 1/100th, if there is one in each 1/100th. Continue for 1/1000ths, 1/10000th... If several events have *exactly* the same creation time, assign them the same number and increment both N and N_L in one step.[206]

[206] This method will fail if infinitely many events have exactly the same time of creation. I presume this is not expected to happen.

5 Paradoxical Decompositions

5.1 What They Are

The construction of Section 4.2 above is just the first of many that yield results troublesome to additive measures. It is one of the simplest instantiations of what is known as a paradoxical decomposition. Their specification is rather general. Following Wagon (1994, Ch. 1), such decompositions arise in the context of a set E that can be partitioned into a countable collection of pairwise disjoint subsets, $A_1, A_2, A_3, \ldots, B_1, B_2, B_3, \ldots$ [207]

$$E = A_1 \cup A_2 \cup A_3 \cup \ldots \cup B_1 \cup B_2 \cup B_3 \cup \ldots$$

There must also be a group G that acts on the set E. Its elements map these subsets to other subsets of E. The original set E admits a paradoxical decomposition if elements of the group can map the A-sets of the partition to sets whose union exhaust E; and correspondingly for the B-sets. That is, there are elements of G, $g_1, g_2, g_3, \ldots$ and $h_1, h_2, h_3, \ldots$, such that we have

$$E = g_1(A_1) \cup g_2(A_2) \cup g_3(A_3) \cup \ldots$$

$$E = h_1(B_1) \cup h_2(B_2) \cup h_3(B_3) \cup \ldots$$

The standard definitions (Wagon, 1994, Def. 1.1, p. 4; p.7) do not explicitly allow for a common and important case: the mapping of the disjoint A-sets and B-sets onto E can be inverted. That is, a partition of the entire set E can be mapped back to either the A-sets or B-sets by elements of G.[208] When this inversion is possible, then elements of the group G can map the A-subsets onto the B-subsets; and conversely.

The construction of Section 4.2 above conforms to the conditions of paradoxical decomposition. Quadrant IV might correspond to the A-sets and the union of quadrants I, II and III might correspond to the B-sets. The group is the group of isometries of a Euclidean space. These are the maps on the space that preserve metrical distance and thus also areas. They comprise translations, rotations and reflections. Moving a tile from one part of the space to another, while preserving its area, corresponds to allowing one of the isometries to act on it. In this case, it is a translation.

The conditions for a paradoxical decomposition are realized since a rearrangement of the tiles in quadrant IV can cover the whole space; and the same is true of the tiles in the union of

---

[207] In the case that the A-subsets and the B-subsets each are finite in number, they do *not* need to be the same number.

[208] This inversion can fail if, for example, the image sets $g_1(A_1), g_2(A_2), g_3(A_3), \ldots$ are not disjoint.

quadrants I, II and III. The case that concerned us, however, was the further case in which inversions are possible. Then the tiles in quadrant IV can be swapped with those in quadrants I, II and III. The import of several swaps of this type was the non-additive chances (6).

5.2 How They Extend the Analysis

There are two aspects of the argument of Section 4 for these non-additive chances that could be strengthened. First, the argument requires a decomposition into infinitely many subsets that are then rearranged to give the final result. One might worry that there is some trickery peculiar to the infinitude of the decomposition.

(i) Can the construction still proceed if the decomposition is into finitely many parts only?

Second, the total area of the Euclidean plane involved in the paradoxical decomposition is infinite.

(ii) Are paradoxical decompositions possible if we require the total area or, more generally, the total volume of the space to be finite?

The literature in paradoxical decompositions has provided affirmative answers to both questions.

A paradoxical decomposition with finitely many subsets and using the group of isometries is not possible in the Euclidean plane. It is possible, however, if we move to non-Euclidean geometries. After the geometry of Euclid, the next simplest geometries are the spaces of constant positive and negative curvature. The second case of constant negative curvature is a hyperbolic geometry. It is a space of infinite area. In it Euclid's axiom of the parallels fails in this way: There is more than one straight line through a point, parallel to a given straight line elsewhere in the space. It can be visualized, piecewise, as the geometry induced on a saddle shaped surface in a higher dimensional Euclidean space.

Wagon (1994, pp. 61-68) shows that it is possible to divide up a two-dimensional hyperbolic space into three disjoint parts whose union exhausts the space and provides a paradoxical decomposition, using the isometry group. (See Wapner (2005, pp. 45-48) for a simplified and engaging development.) Call the disjoint parts A, B and C. If we choose a suitable axis of rotation, Wagon shows that it is possible to rotate A by 120º so that it now coincides with B. A further rotation by 120º then leaves A coincident with C. These rotations are isometries, so they preserve the areas of the parts rotated.

We might pause at this moment and imagine that a point is chosen randomly in the space such that metrically adapted label independence is respected. These rotations by 120º are metrically adapted permutations that can swap the labeling among the three sets A, B and C. Thus they have equal chances. If we assign probabilities to the chosen point being in A or in B or in C, we must then have

$$P(A) = P(B) = P(C) = 1/3$$

so that $P(A) + P(B) + P(C) = 1$.

The trouble is that rotations around a different point in the space lead to different results. With a different, suitably chosen axis of rotation, a rotation of A by 180º leaves it coincident with the union of B and C. Applying the same reasoning, we now arrive at probability assignments

$$P(A) = P(B \cup C) = P(B) + P(C) = 1/2$$

They are incompatible with the first set of probability assignments. Once again we find that these chances cannot be represented by probabilities.

A curious sidelight is that this case of a hyperbolic space could almost be applied directly to the example of steady state cosmology of Section 4.3. The spacetime of steady state cosmology is a de Sitter spacetime. Bondi, Gold and Hoyle introduced a cosmic time that slices the spacetime into spaces at different instants of cosmic time. They chose a slicing that yields Euclidean spaces. A de Sitter spacetime is rich in symmetries. It turns out that there are other ways of slicing it that admit different cosmic times. In another choice of cosmic time, the spaces at each cosmic instant are hyperbolic in their geometry. If we ask for matter to be created continuously by some stochastic process that is uniform in the hyperbolic space, the construction just sketched, promoted to a three dimensional space, shows that this uniformity cannot be represented probabilistically. The demonstration does not require decomposition into infinitely many parts, but just the three indicated. However the cogency of this more elegant construction is lessened by the fact that a slicing of a de Sitter spacetime into hyperbolic spaces is uncongenial to steady state cosmology. For in this slicing, the radius of curvature of the space would vary with cosmic time. (See Bondi, 1960, p. 145.) While this variant slicing is simply another way of displaying the spacetime structure of the steady state cosmology, its associated cosmic time is not one in which the perfect cosmological principle can be expressed.

The areas A, B and C of this construction are not as simple geometrically as the quadrants of Euclidean space used in Section 4.2. Each consists of infinitely may parts, with the parts touching only at points, as shown in the rather pretty diagrams in the references above. However decomposition of the hyperbolic space into these three parts is notable in one aspect: *it does not require the axiom of choice*. The significance of this statement will be clarified below.

The hyperbolic space is infinite in area and the three parts A, B and C are also infinite in area. That infinity allows them to be rotated into one another in ways that preclude a finite, additive measure for the areas. For when areas are infinite, we can write all of the following:

$$\text{Area}(A) = \text{Area}\ (B) = \text{Area}\ (C) = \infty$$

$$\text{Area}(A) = \text{Area}\ (B \cup C) = \text{Area}\ (B) + \text{Area}\ (C) = \infty$$

Since these equalities cannot all be satisfied if the areas of the parts are finite, one might expect that a paradoxical decomposition of a space of finite area or volume is not possible.

That expectation proves incorrect. There are paradoxical decompositions of spaces of finite volume. The celebrated example is the Banach-Tarski paradox. It has been discussed in such detail elsewhere, that it needs only the barest statement here. See Wapner (2005, Ch.5) for a very clear development; and Wagon (1994) for a mathematically more thorough treatment. The basic result is that a sphere in three-dimensional Euclidean space can be decomposed into five parts. The parts are then rearranged in space, where the rearrangement employs only volume preserving isometries. The result is two spheres, each with the same volume as the original sphere.

The air of paradox reflected in the name derives from the apparent impossibility of the process. We decompose a sphere into parts that can be recombined into two spheres whose total volume is double that of the original sphere, where all the rearrangments are isometries. The air of paradox is dispelled, however, once we find that four of the five parts in the standard decomposition are nonmeasurable in the background metric of Euclidean space. They are not simple volumes of the type normally encountered in geometry. They are scatterings of infinitely many points that defy simple geometric description. No volume can be consistently assigned to them.[209] Thus the constructions are revealed as very fancy versions of a more familiar decomposition. We can take a countable infinity of entities labeled 1, 2, 3, … and divide them into the set of odd labeled entities and the set of even labeled entities. If we now relabel the entities in each set with 1, 2, 3, … and 1, 2, 3, …, we have doubled the set of entities, or at least that is what the labeling indicates.

While Banach-Tarski like constructions have proven enormously stimulating to mathematical inquiry,[210] the most important contribution to our concerns here arises at the outset.

---

[209] A point to which we will shortly return: the axiom of choice is needed to arrive at their existence.

[210] When one first encounters these constructions, one might be quite amazed that a mortal mathematician could discover them. Or at least that was my reaction. What I found very helpful was the recognition that the more complicated constructions derive from a simple piece of group theory. The elements of the free group with two generators *a* and *b* consist of finite strings of symbols like $abba^{-1}b^{-1}$a of arbitrary, but always finite length. It is easy to see that a paradoxical decomposition is possible in this set of group elements. Any good treatment shows it. All that remains is to realize the generators in some geometrical setting, for example as rotations in space, and in a way that preserves the free group properties. Banach-Tarski like paradoxes then appear and they require three dimensions of space, since in two dimensions the two generators *a*

It is that there are nonmeasurable sets. Their existence represents some sort of obstacle to the universal applicability of additive probability measures in inductive inference. The next section turns to look at how these nonmeasurable sets arise.

6. A Nonmeasurable Set

6.1 A Vitali Set

The simplest example of a nonmeasurable set, used almost universally as in introduction to the general idea, is a Vitali set. (See Kharazishvili, 2004, ch. 1; Wagon, 1994, pp. 7-8; Wapner, 2005, pp. 132-35) The version to be developed here will be a subset of the interval of real numbers [0,1), that is all real numbers x such that $0 \leq x < 1$. These real numbers will be the angular coordinates that cover a circle as shown in Figure 6.

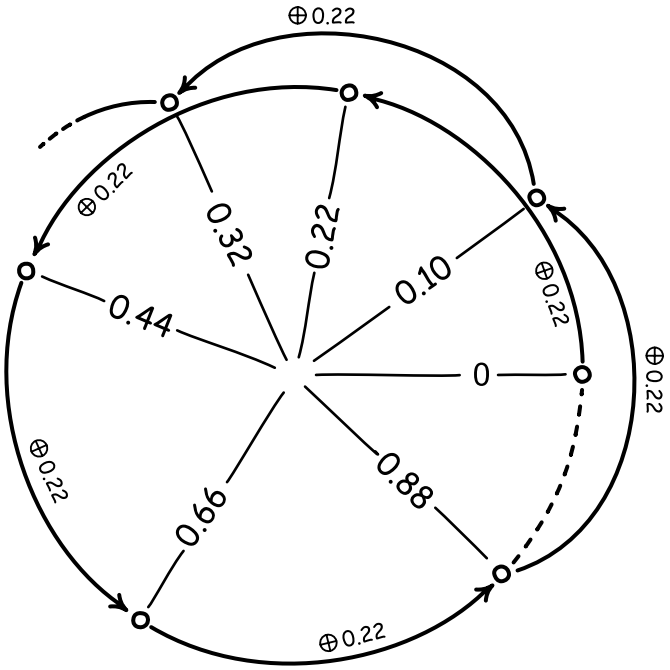


Figure 6. Equivalent numbers used in the construction of a Vitali Set

Two real numbers are defined to be equivalent under the relation "~" if they differ only by a rational number. That is $x \sim y$ just in case there is a rational number r such that $y = x \oplus r$. Addition "⊕" is modulus 1 addition. To compute it, the numbers x and r are added by ordinary arithmetic. If the resultant exceeds one, one is subtracted. If it is negative, one is added. This

---

and b cannot be realized. The complications of the geometry of the rotations mask the constructions' simple origins.

modular rule ensures that the sum shown always remains in the interval [0,1). Figuratively, addition by *r* just steps us repeatedly round the circle of Figure 6. This figure shows points equivalent under successive addition of the rational number 0.22 = 11/50, that is 0, 0.22, 0.44, 0.66, 0.88, 0.10, 0.32, …

Since the relation is an equivalence relation, it divides all the real numbers in [0,1) into disjoint equivalence classes. They are distinguished by a number that, as I shall say, "seeds" them. The rational number 0 seeds an equivalence class that contains all the rational numbers in [0,1). This shows immediately that each equivalence class has infinitely many seeds: every rational number in [0,1) seeds the same class. Irrational numbers seed other classes. The irrational $1/\sqrt{2} = 0.7071\ldots$ seeds a class that contains $(\sqrt{2}-1)/2 = 0.2071\ldots$ since

$$\frac{1}{\sqrt{2}} - \frac{1}{2} = \frac{\sqrt{2}-1}{2}$$

The simple graphic of Figure 7 displays the partition of [0,1) into the equivalence classes.

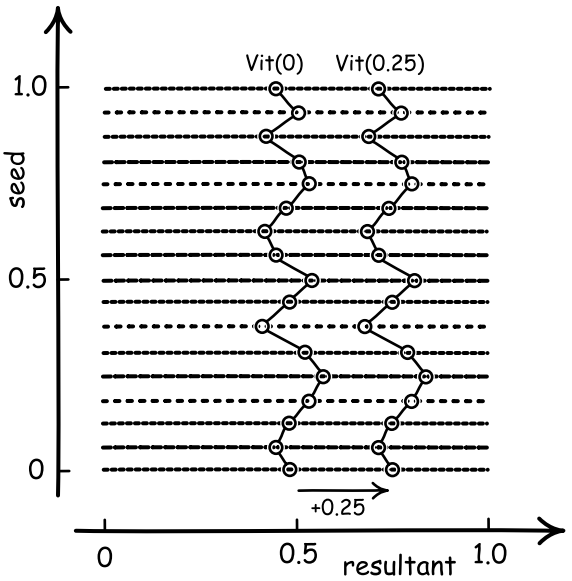


Figure 7. Choices that Form a Vitali Set

The points in the square are all the real numbers in [0,1). Each is uniquely picked out by the seed of the equivalence class to which is belongs and the rational increment added to the seed to arrive at it. The vertical axis shows the seeds used to create each equivalence class. The axis has many gaps in it since all duplicated seeds are eliminated. Its seeds include only one rational number and only one of $1/\sqrt{2}$ and $(\sqrt{2}-1)/2$. The horizontal axis shows the values in [0,1) that the various members of each equivalence class can take after all the rationals are added to

the seed of the equivalence class. Each equivalence class is represented by a single horizontal line.

A Vitali set is formed by taking just one number from each equivalence class. This means that the difference between two elements in the set cannot be a rational number. Forming the set amounts to taking a vertical section in the square shown in Figure 7. It seems obvious at this point that such a section can be taken. (This is a point to which we will return shortly.) Moreover there are very many ways that this section can be taken, so very many sets can be Vitali sets. We just need to settle on one to proceed. Call it $Vit(0)$.

To demonstrate that this is a nonmeasurable set, we need a measure, for a set can be nonmeasurable only with respect to some specified measure. We take the uniform distribution (1) over [0,1) as that measure. Its uniformity gives it the property of translation invariance. That is, if the probability density assigns some probability $P(A)$ to a subset A of [0,1), then it assigns the same probability to the set A_x produced by translating all numbers in A by the same amount x:[211] $P(A_x) = P(A)$. Applying a uniform translation by r to all the numbers in the Vitali set $Vit(0)$, we form the translated set $Vit(r)$. Figure 7 shows $Vit(0.25)$.

It is easy to see that the set of translated Vitali sets, for all rational numbers r, partition the interval [0,1), as shown in Figure 8.

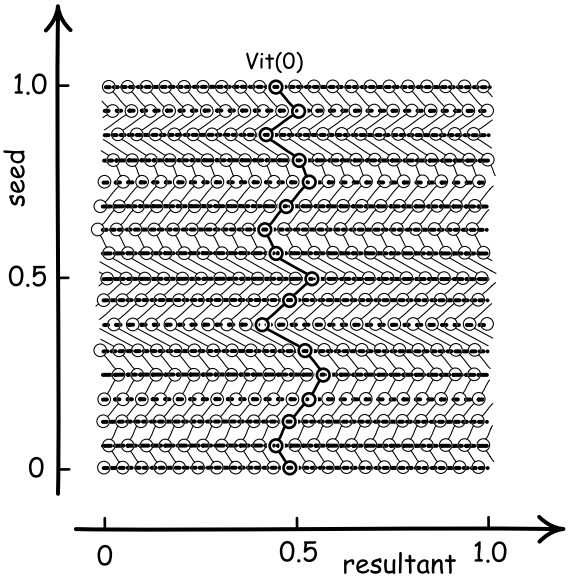


Figure 8. Vitali Sets Partition [0,1)

[211] That is, A_x is $\{y \oplus x: y \in A\}$.

That is their union is [0,1) and the translated sets are pairwise disjoint. The first follows by construction, since every number in [0,1) is either in $Vit(0)$ or arrived at from an element of $Vit(0)$ by adding a rational r to it, which means that it is a member of $Vit(r)$. Two translated Vitali sets $V(r)$ and $V(s)$ are disjoint for unequal rational numbers r and s. For otherwise they share a common element of the form $x \oplus r = y \oplus s$, where both x and y are elements of $V(0)$. However this last equation entails that x and y differ by a rational number. This cannot be true of any two distinct elements of $V(0)$, since each is drawn from a distinct equivalence class.

Assume for purposes of a reductio argument, that the Vitali set is measurable under the uniform density (1) and that is has a probability P. Since the probability density is invariant under translation, it follows that all uniformly translated Vitali sets $Vit(r)$ have the same probability. The set of rational numbers is countable.[212] Therefore there are countably many translated Vitali sets. The countable sum of their probabilities must be unity. That is, the summation of a countable infinity of probabilities P must be unity. No real number P can satisfy this condition. If P is zero, the countably infinite sum is zero. If P is greater than zero, no matter how small, the countably infinite sum is infinite. We have arrived at a contradiction. The Vitali set $Vit(0)$ is not measurable under the uniform density (1).

6.2 The Infinite Lottery Machine Logic, Again

How does the existence of nonmeasurable sets like a Vitali set affect inductive inference? We can set up an inductive inference problem that uses this Vitali set by assuming that a real number has been chosen in the interval [0,1). We will assume that the choice is uniform in the sense that the chance of selection in any set, if defined, is unchanged by translations of the set. It follows that the distribution of chances in the space conforms with metrically adapted label independence, where the permutations are translations that preserve the metric associated with the probability density (1). It now follows that each of the translated Vitali sets $Vit(r)$ must have equal chances. For any pair of Vitali sets, $Vit(r)$ and $Vit(s)$, a translation by s-r shifts the labels on the first set to the second.

The inductive problem is to determine the chances that the point selected lies in one of the Vitali sets, or in some union of them. The probability measures derived from the uniform density (1) cannot supply chances for these outcomes, for it is not defined on them. Rather, the applicable logic is the infinite lottery machine logic of the previous chapter. To see this note that the countably many Vitali sets $Vit(r)$ can be relabeled by the natural numbers 1, 2, 3, … Each Vitali set $V(1)$, $V(2)$, … has the same chance and, under the new labeling, conforms with the

[212] For each rational can be represented by the ratio p/q of natural numbers p and q. The pair can then be mapped one-one to an infinite subset of the natural numbers by the formula 2^p3^q.

original, unrestricted requirement of label independence. These are just the conditions to which an infinite lottery machine conforms. By repeating the arguments concerning it, we can infer that:

Chance that the point chosen is in some finite set of Vitali sets of size N is V_N.

Chance that the point chosen is in some infinite-co-infinite set of Vitali sets is V_∞.

Chance that the point chosen is in some infinite-co-finite set of Vitali sets, where the complement is of size N, is V_{-N}.

The familiar results now follow. There is the same chance that the point chosen is in the infinite set of Vitali sets that have even numbered labels, in those with odd numbered labels, in those with labels that are powers of ten: 1, 10, 100, 1000, … etc. On many repetitions there is no stabilization of frequencies such as would conform with a probability measure. We do not stabilize with roughly half the points selected in the odd numbered set and half in the even numbered set.

6.3 The Axiom of Choice

This last analysis assumes that a logic of induction should accommodate outcomes in nonmeasurable sets like the Vitali sets. However these nonmeasurable sets have a disputed status in mathematics. The difficulty derives from a key step in the analysis. The Vitali set $V(0)$ was formed by selecting just one element from each of the equivalence classes above. It was simply assumed that such a selection is possible. To see that matters are not quite so simple, one should reflect on just how we are to make the selection. Might we choose the smallest or largest element in each equivalence class? That fails since there might be no smallest or largest element. Might we choose that element that is the median value, that is, the one that comes half way through? Since the equivalence classes are infinite, "half way through" is ill-defined. Might we choose the element whose value coincides with the mean of all members in the equivalence class? That fails since there may be so such element.

We might suspect that all these failures derive from a poor imagination. There is some recipe, we might hope, even if very complicated, that lets us specify which set is our Vitali set $V(0)$. However it turns out that no one has been able to find a constructive formula that can specify the uncountable infinity of choices needed. There are formal results that suggest but do not prove that no such constructive formula is possible. Rather, the best we can now do is simply to assume that there does exist a set comprised of just one element from each equivalence class. At first glance, the existence of such sets seems so straightforward that it can hardly be doubted. But then we find reasons for doubt. Since a Vitali set results from an uncountable infinity of selections of numbers from an uncountable infinity of equivalence classes, if there are any Vitali sets, then there are very many of them. Yet when we try positively to specify just one, we can

find no way to do it. If they exist, all we can say is that they are somewhere in very great numbers in the mathematical universe. We just cannot specify precisely where.

These last considerations have been codified into more precise mathematics. The standard treatment of sets is the Zermelo-Fraenkel set theory.[213] Its axioms were developed to rescue set theory after Russell's paradox showed its naïve foundations fatally flawed. In the naïve set theory, we assume that a set can be formed as those things that satisfy any condition we can specify. Famously Russell used this rule to create the set of all sets that do not contain themselves as elements. The set is contradictory in that it can be member of itself if and only if it is not a member of itself.

To avoid this problem, Zermelo-Fraenkel set theory is restrained in just what sets it allows to exist. Its axioms do provide cautiously for the existence and behaviors of certain sets and include what amount to principles of set construction. The axiom schema of subsets tells us that we can always create a new set as a subset from another by placing some restrictive condition on elements in the original set. This replaces the problematic naïve rule with a benign rule, since its set delineating condition can only carve off a set from an already existing set. It does not permit formation of Russell's set. Other axioms assert the existence of the null set; of the union of two sets that are already elements of another set; of the power set of all subsets of a set; and of an infinite set constructed by specific conditions.

Constructive axioms of this type proved able to recover much of set theory. However they are not rich enough to provide for the sets like the Vitali sets of the last section. It turned out that their existence could only be secured by introducing a new, non-constructive axiom that merely asserted the existence of certain sets, but gave no recipe for their construction. That axiom is the "axiom of choice," or something equivalent to it. That axiom amounts to the assertion that, if we have a set of member sets that are pairwise disjoint, then there exists another set comprised of just one element from each of the member sets. The Vitali set $Vit(0)$ formed above is just such a set. The presumption that it exists amounts to applying the axiom of choice.

The axiom of choice has been surrounded by an air of uncertainty. A major motivation for the uncertainty was the discovery of the Banach-Tarski paradox, for the formation of the sets in the paradox require the axiom. As a result, treatments of the paradox routinely include labored discussions of the cogency of the axiom. See, for example the ominously numbered Chapter 13 of Wagon (1994). As far as I can see the question of the admissibility of the axiom and thus of nonmeasurable sets remains open simply in virtue of the lack of any well-principled means to decide for or against it.

---

[213] For an easier introduction, see Stoll (1979, Ch. 7).

The original basis for arguments against it was the intuitive inadmissibility of results like the Banach-Tarski paradox. To block the paradox, one had to overturn something in the foundations of set theory. The axiom of choice stood out as the easiest target because of its non-constructive character. But if one reconciles to the Banach-Tarski paradox so that it becomes the more benignly labeled Banach-Tarski theorem, then this basis for rejecting the axiom of choice is lost. Other reasons for rejecting it are hard to find. Its truth is not empirically decidable. There is no physical test we can perform to detect the existence of nonmeasurable sets of points specifically in some physical space. The axiom has been shown to be consistent with the other axioms of the Zermelo-Fraenkel set theory, so there no problem in logic in adding it to the axioms of the theory.

Correspondingly, however, there seems to be no decisive grounds for adding the axiom of choice to the other axioms of Zermelo-Fraenkel set theory. Just as there is no empirical way to falsify the axiom, there is no empirical way to demonstrate it. Rather the principal motivation for employing it seems to be pragmatic: much useful mathematics depends upon it. For example, Zorn's lemma, which is equivalent to the axiom of choice, is needed to demonstrate that every vector space has a basis.[214]

This pragmatic attitude is perhaps not so different from a simpler one. No measurement can distinguish whether a physical magnitude is an irrational real number or some nearby rational number. Any measurement has some inexactness. We can never affirm by direct measurement that the hypotenuse of a right-angled triangle with unit sides is exactly the irrational number $\sqrt{2}$ = 1.41421… as opposed to the nearby rational numbers 14/10 or 141/100 or 1414/1000 and so on. However if we forego the possibility of irrational lengths in space, we forego the right-angled triangles of Pythagoras' theorem. Instead the best we would have would be many triangles, all with sides of rational length, that come arbitrarily close to the side lengths of Pythagoras' theorem. We may congratulate ourselves on the purity of our prudence in restricting ourselves to the observationally more secure. Our reward would be mathematical complexities that would propagate pain and misery through the entirety of our physical theories.

Our question here is not simply that of the admissibility of the axiom choice. It is a slightly different one. Should an account of inductive inference be responsible for relations among propositions that pertain to nonmeasurable sets? To forego exploring these relations would require positive reasons for precluding nonmeasurable sets. I do not see them unless we are prepared to entertain anthropocentric perspectives on the world. That might happen if we are so committed a subjectivist that we reduce the scope of inductive inference to relations among

[214] See Brunner et al. (1996) for an extended analysis of the role of the axiom of choice in the mathematics of quantum theory.

things that we can construct. That attitude seems quite presumptuous to me. That nonmeasurable sets outstrip our constructive prescriptions seems to me quite reasonably explained by the weakness of those prescriptions. They are weak, as we have repeatedly learned. We would like a finite axiom system whose theorems would include all the truths of arithmetic. Goedel's famous theorem shows that no finite axiom system can do this. It tells us that our arithmetic axiomatic methods are weak in their reach. If finite prescriptions are essential to us, we run into trouble at the very start of mathematics. There is an uncountable infinity of real numbers in [0,1]. Yet our language admits of only countably many sentences for describing them. Most real numbers outstrip our descriptive reach. Return now to nonmeasurable sets. Are there, we might ask, nonmeasurable sets of points in our physical space? Whether there are or not is a physical fact about space and true whether our finite constructive devices allow us to give a precise description of them.

In my view, as long as the status of these sets remains open, we should consider what an inductive logic must do to accommodate them. For a general understanding of the nature of inductive inference must be expansive enough include these accommodations. To do otherwise is to prejudge the status of nonmeasurable sets and artificially restrict the scope of inductive inference. It is in that permissive spirit that the explorations of this chapter are undertaken.

7. Blackwell and Diaconis' Nonmeasurable Coin Toss Event

Most instances of nonmeasurable sets arise in the esoteric realm of abstract mathematics. When we use the sets to specify chancy events, that makes the events seem distant from the concerns of an inductive logic that may apply to real science. It would help to reduce that distance if we could find nonmeasurable events that arise in the archetypal probabilistic problem of sequential coin tosses. Blackwell and Diaconis (1996) have described such events. An account of them will be given in this section. An interesting bonus is that the apparatus needed to describe the events enables specification of another inductive logic that, while very weak, applies to events that are otherwise probabilistically nonmeasurable.

7.1 Tail Events

Blackwell and Diaconis' event arises in the case of infinitely many coin tosses. Each toss has a probability of 1/2 of a head H or a tail T; and the tosses are all probabilistically independent. Our elementary events will be infinite sequences of heads and tails. If we let variables a_1 = H or T, a_2 = H or T, a_3 = H or T, …, then such an infinite sequence is represented by the infinite tuple $\mathbf{a} = < a_1, a_2, a_3, \ldots >$. The nonmeasurable event will be one of what is called a "tail set," or, as I shall call them here, "tail event." These are events whose properties (such as

the probability, if defined) depend only on the long term behavior of the infinite sequence, that is, on its tail.

Such events are familiar and important. For example, elementary events like

$$\langle H, T, H, T, H, T, H, T, \ldots\rangle \text{ and } \langle H, H, T, T, H, H, T, T, \ldots\rangle$$

are distinctive in that the limiting relative frequency of heads H is 1/2. This distinctive property is shared by many other elementary events that differ in finitely many of the individual coin tosses. For example

$$\langle H, H, H, H, H, H, H, H, H, H, H, T, H, T, H, T, H, T, \ldots\rangle$$

differs from <H, T, H, T, H, T, H, T, …> only in its first few tosses. It will still return a limiting relative frequency of heads H of 1/2. The heavy weighting towards H in the early tosses is eventually and inexorably swamped by the later tosses.

Each of these elementary events has a probability given by the infinite product 1/2 x 1/2 x 1/2 x …. That is, each has probability zero. There are infinitely many elementary events that return this limiting relative frequency. We combine[215] them disjunctively to form the event "*half*": that the infinitely many coin tosses return a limiting relative frequency of heads H of 1/2. Since the individual tosses are probabilistically independent and each of probability 1/2, we can apply the strong law of large numbers to conclude that the event *half* will occur with probability one, $P(half) = 1$.

This last paragraph describes the distinctive property of a tail event: its probability is unaffected by whatever may happen in finitely many of the tosses that comprise it. More precisely:

Tail event characterization 1: a tail event is probabilistically independent of the outcome of any finite set of tosses.

Recall that two events A and B are probabilistically independent just if $P(A \,\&\, B) = P(A).P(B)$. This defining property means that *half* is independent of the conjunction $(a_1 = H) \,\&\, (a_2 = H) \,\&\, (a_7 = H) \,\&\, (a_{63} = H)$, so that:

$$P(half \,\&\, (a_1 = H) \,\&\, (a_2 = H) \,\&\, (a_7 = H) \,\&\, (a_{63} = H))$$
$$= P(half) \,.\, P((a_1 = H) \,\&\, (a_2 = H) \,\&\, (a_7 = H) \,\&\, (a_{63} = H))$$

and similarly for any other finite set of tosses.

There are many other tail events. For example:

quarter: the limiting relative frequency of heads H is 1/4. $P(quarter) = 0$.

three-quarters: the limiting relative frequency of heads H is 3/4. $P(three\text{-}quarters) = 0$.

---

[215] If we think of the events as propositions, then we are "or"ing them together. If we think of them as elements of a set, we are collecting them into a set.

interval-no: the limiting relative frequency of heads H lies in some interval of reals that does *not* contain 1/2: $P(\textit{interval-no}) = 0$.

interval-yes: the limiting relative frequency of heads H lies in some interval of reals that *does* contain 1/2: $P(\textit{interval-yes}) = 1$.

even-H: an infinite number of even numbered tosses are head H. $P(\textit{even-H}) = 1$

Tolstoy: the infinite sequence contains, infinitely often, the entirety of Tolstoy's *War and Peace*, encoded in binary using H and T, as well as every variant of the same length created by all possible typographical errors. $P(\textit{Tolstoy}) = 1$.

It may at first seem that this list of examples is uncreative in the sense that every probability is a zero or a one. Those zeroes and ones are unavoidable however. The Kolmogorov (1950, pp. 69-70) Zero-One Law asserts that all tail events to which probability can be assigned are of probability zero or one only.

The proof of the law involves some mathematical complications. Rosenthal (2006, §3.5) gives a serviceable formulation as well as a helpful account of tail events. The basic idea behind the proof, however, is so simple and striking as to bear mention. As we saw above, the defining characteristic of a tail event we shall call "*tail*" in infinitely many coin tosses is that it is probabilistically independent of any event formed from only finitely many coin tosses, such as one we will here call "*finite*." That means

$$P(\textit{tail} \ \& \ \textit{finite}) = P(\textit{tail}) \ . \ P(\textit{finite})$$

for all possible *finite*. The unusual circumstance is that the event *tail* is a member of the infinite set of events formed from all possible instantiations of *finite*, when closed under finite and countable unions and intersections.[216] This leads eventually to the curious result that *tail* is independent of itself! Substituting *tail* for *finite* in this last equation and noting that *tail* & *tail* = *tail*, we have

$$P(\textit{tail}) = P(\textit{tail} \ \& \ \textit{tail}) = P(\textit{tail}) \ . \ P(\textit{tail})$$

This equation admits only two solutions

$$P(\textit{tail}) = 0 \quad \text{and} \quad P(\textit{tail}) = 1$$

Since we will shortly be dealing with nonmeasurable events, we will need another characterization of tail events that does not explicitly invoke probability measures. That condition is simply that

Tail event characterization 2: if $\mathbf{a} = < a_1, a_2, a_3, \ldots >$ is an elementary event within some tail event and $\mathbf{b} = < b_1, b_2, b_3, \ldots >$ is any elementary event that differs from it in only finitely many tosses, then $\mathbf{b}$ is also in the tail event.

---

[216] That is, the σ-algebra formed from all instantiations of *finite*.

This new characterization entails the original one above in case the events concerned have well defined probabilities. To see this, pick any finite set, such as a_1 and a_3. Let us say that

$$\mathbf{a}_{\text{H.H...}} = < a_1 = \text{H}, a_2, a_3 = \text{H}, a_4, a_5, a_6, \ldots >$$

is an elementary event in some tail event where $a_2, a_4, a_5, a_6, \ldots$ have some values that are kept fixed in what follows here. The new condition requires that all combinations of alternative values of a_1 and a_3 appear in other elementary events in the tail event. These additional events are

$$\mathbf{a}_{\text{H.T...}} = < a_1 = \text{H}, a_2, a_3 = \text{T}, a_4, a_5, a_6, \ldots >$$

$$\mathbf{a}_{\text{T.H...}} = < a_1 = \text{T}, a_2, a_3 = \text{H}, a_4, a_5, a_6, \ldots >$$

$$\mathbf{a}_{\text{T.T...}} = < a_1 = \text{T}, a_2, a_3 = \text{H}, a_4, a_5, a_6, \ldots >$$

The probabilistic contribution to the tail event by these four elementary events is

$$P(\mathbf{a}_{\text{H.H...}} \text{ v } \mathbf{a}_{\text{H.T...}} \text{ v } \mathbf{a}_{\text{T.H...}} \text{ v } \mathbf{a}_{\text{T.T...}}) = P(\mathbf{a}_{\text{H.H...}}) + P(\mathbf{a}_{\text{H.T...}}) + P(\mathbf{a}_{\text{T.H...}}) + P(\mathbf{a}_{\text{T.T...}})$$

$$\begin{aligned} &= P(a_1 = \text{H}) \,.\, P(a_3 = \text{H}).\, P(<a_2, a_4, a_5, a_6, \ldots>) + \\ &\quad P(a_1 = \text{H}) \,.\, P(a_3 = \text{T}).\, P(<a_2, a_4, a_5, a_6, \ldots>) + \\ &\quad P(a_1 = \text{T}) \,.\, P(a_3 = \text{H}).\, P(<a_2, a_4, a_5, a_6, \ldots>) + \\ &\quad P(a_1 = \text{T}) \,.\, P(a_3 = \text{T}).\, P(<a_2, a_4, a_5, a_6, \ldots>) \\ &= P(<a_2, a_4, a_5, a_6, \ldots>). \end{aligned}$$

This is just the probabilistic contribution to the tail arising when tosses a_1 and a_3 are excluded, which shows the probability is independent of the tosses a_1 and a_3. Repeating for all other finite combinations of tosses, we see that the probability of the tail event is independent of any of these finite combinations, which is the first characterization of tail event above.

7.2 An Intermediate Tail Event E[217]

We can start with a tail event of probability zero. By adding new elementary events to it, we can expand it to a tail event of probability one. For example, we might start with the tail event *interval-no* that is defined by the limiting relative frequency of heads lying in the interval 0.9 to 1.0. Since 0.5=1/2 is not in that interval, this tail event has zero probability. We continuously expand the interval by adding more elementary events until the interval becomes 0.4 to 1.0. At the moment when the interval expands to include the limiting relative frequency of heads of 0.5, its probability will flip from zero to one. Writing "rf" for the limiting relative frequency of heads and assuming that the intervals include their end points, we have

P(rf in 0.6 to 1.0) = 0
P(rf in 0.55 to 1.0) = 0
P(rf in 0.51 to 1.0) = 0
P(rf in 0.50001 to 1.0) = 0
P(rf in 0.5 to 1.0) = 1
P(rf in 0.49999 to 1.0) = 1
P(rf in 0.45 to 1.0) = 1

This last example suggests that, as we assemble sets of elementary events into events, we find no tail events intermediate between events with probability zero and those with probability one. Certainly there are none in the sequence just considered. However that last sequence included by construction only tail events with well-defined probabilities. What Blackwell and Diaconis demonstrate is that there are very many tail events, intermediate between events with zero and one probability, and that these tail events are probabilistically nonmeasureable. No probability can be assigned to each of them.

---

[217] Alex Pruss has pointed out another way that a nonmeasurable tail event may be formed in this coin tossing example. Each elementary event has a reversed event in which every H is replaced by T and every T by H. We form maximal equivalence classes of elementary events, such that two events in the same class differ only in finitely many of the individual coin toss outcomes. For each such equivalence class U there is reversed class U^r consisting of the reversals of the elementary events in U. The entire outcome set is partitioned by an infinity of (unordered) pairs of such classes: $\{U, U^r\}, \{V, V^r\}, \ldots$ Using the axiom of choice for collections of two-membered sets, we choose one equivalence class from each pair. Their union is the tail event N. The entire outcome set is partitioned by N and its reversal N^r. The event N satisfies conditions (a) and (b) of Section 7.3 and thus is nonmeasurable. See http://alexanderpruss.blogspot.com/2017/11/heres-simple-construction-of-non.html

We begin assembling Blackwell and Diaconis' event "*E*" as a set of elementary events, making our focus the presence of H toss outcomes. The first elementary event in *E* is just one that consists of all H:

$$\mathbf{a}_{\text{all-H}} = \langle H, H, H, H, H, H, H, H, \ldots \rangle$$

We now add to *E* all elementary events that differ from $\mathbf{a}_{\text{all-H}}$ in only finitely many tosses. They include:

$\langle T, H, H, H, H, H, H, H, \ldots \rangle$
$\langle H, T, H, H, H, H, H, H, \ldots \rangle$
$\langle H, H, T, H, H, H, H, H, \ldots \rangle$
…
$\langle T, T, H, H, H, H, H, H, \ldots \rangle$
$\langle T, H, T, H, H, H, H, H, \ldots \rangle$
…

Call them "infinite H, finite T" elementary events. There are as many of these elementary outcomes as there are subsets of the natural numbers. That is, there is a higher order of infinity of them. Nonetheless, the probability of the event just formed is zero. It is a tail event characterized by a limiting relative frequency of heads of one. Our starting point is essentially the same as the growing intervals of tail events above.

We will add many, many more elementary events to *E* but in a way that avoids the flipping of probability from zero to one. We achieve this by adding elementary events in a way that conforms with a specific set of rules. To express them, we need to define the intersection operation $\cap$ on elementary events. The intersection of elementary events $\mathbf{a}$ and $\mathbf{b}$ is the elementary event $\mathbf{a} \cap \mathbf{b}$ that has H in every position that has H in both $\mathbf{a}$ and $\mathbf{b}$ and T otherwise. For example:

$$\mathbf{a} = \langle H, T, H, T, H, T, H, T, \ldots \rangle$$
$$\mathbf{b} = \langle H, H, T, T, H, H, T, T, \ldots \rangle$$
$$\mathbf{a} \cap \mathbf{b} = \langle H, T, T, T, H, T, T, T, \ldots \rangle$$

The complement $\mathbf{a}^c$ of an elementary event is just that same event $\mathbf{a}$ with each occurrence of H switched to T and each occurrence of T switched to H. For example:

$$\mathbf{a} = \langle H, T, H, T, H, T, H, T, \ldots \rangle$$
$$\mathbf{a}^c = \langle T, H, T, H, T, H, T, H, \ldots \rangle$$

The event *E* is a set of elementary events, where we write elementary event $\mathbf{a}$ is a member of *E* as $\mathbf{a} \in E$.

The rules for forming E are that the following conditions are respected as the elementary events are added:

I. The "no-H" elementary event $\mathbf{a}_{\text{no-H}} = \langle T, T, T, T, \ldots \rangle$ is not in E.

$\mathbf{a}_{\text{no-H}} \notin E$.

II. ("containment") If $\mathbf{a} \in E$ and $\mathbf{b}$ arises by replacing some T in $\mathbf{a}$ by H, then $\mathbf{b}$ is also in E.

If $\mathbf{a} \in E$ and $\mathbf{a} \cap \mathbf{b} = \mathbf{a}$, then $\mathbf{b} \in E$.

III. ("intersection") The intersections of elementary events in E are also in E.

If $\mathbf{a} \in E$ and $\mathbf{b} \in E$, then $\mathbf{a} \cap \mathbf{b} \in E$.

IV. ("exhaustion") For every element $\mathbf{a}$, either $\mathbf{a}$ or its complement $\mathbf{a}^c$ is in E.

For all $\mathbf{a}$, either $\mathbf{a} \in E$ or $\mathbf{a}^c \in E$

V. ("free") The infinite intersection of all elementary events in E is the "no-H" event.

$$\bigcap_{\mathbf{a} \in E} \mathbf{a} = \mathbf{a}_{no-H}$$

Those with mathematical interests will recognize these five conditions as defining a free ultrafilter. The first three specify a filter. The fourth makes the filter an ultrafilter; and the fifth makes it a free ultrafilter.[218]

These conditions impose a definite structure on the elementary events that comprise E. From III. and I., we have that every intersection of elementary events in E must have some H toss outcomes. Thus, for all elementary events $\mathbf{a}$, just one of $\mathbf{a}$ or its complement $\mathbf{a}^c$ can be included in E. Condition V. ensures that every elementary event in E must contain infinitely many H toss outcomes.[219]

---

[218] Blackwell and Diaconis do not implement the ultrafilter structure directly on the tuples that form the elementary events. Rather they form sets of indices of the locations of H in the tuples. For example, <H, T, H, T, H, T, …> yields the set of odd numbers {1, 3, 5, 7, …}. The ultrafilter is implemented in the set of all these subsets of the natural numbers.

[219] To see this, assume otherwise that there is an elementary event $\mathbf{fin}(n)$ in E that has finitely many H—say, n of them. If $n>1$, then there is an elementary event $\mathbf{a}$ such that $\mathbf{a} \cap \mathbf{fin}(n)$ and $\mathbf{a}^c \cap \mathbf{fin}(n)$ each have one or more H, but each is strictly fewer than n. Since just one of $\mathbf{a}$ and $\mathbf{a}^c$ is in E, it follows from the intersection condition III that there is another elementary event in E with fewer H than n. Iterating, it follows that, if there is an elementary event in E with finitely many H, then there is an elementary event in E with just one H. This elementary event $\mathbf{fin}(1)$ must be contained in every elementary event in E. Otherwise the intersection of $\mathbf{fin}(1)$ with some

The set of "infinite H, finite T" elementary events along with $\mathbf{a}_{\text{all-H}}$ satisfies all these conditions, excepting IV.[220] While we have not fully specified the content of *E*, we can already see at this stage that any possible set *E* must include this set. This follows from II and the fact that I requires that some H must be present in all the events of any possible set *E*.

To satisfy exhaustion IV, we need to add further events. We have many choices over which to add. For example, we must add one of the elementary events in *half*

$$\mathbf{a} = \langle H, T, H, T, H, T, H, T, \ldots \rangle$$

or its complement

$$\mathbf{a}^c = \langle T, H, T, H, T, H, T, H, \ldots \rangle$$

But we cannot add both. Next, we must choose among

$$\mathbf{b} = \langle H, H, T, T, H, H, T, T, \ldots \rangle$$

$$\mathbf{b}^c = \langle T, T, H, H, T, T, H, H, \ldots \rangle$$

Adding the tail event *half* flipped the probability of the continuously growing set of tails events above from zero to one. We now see that this tail event cannot be a subset of *E*. For all four of $\mathbf{a}$, $\mathbf{a}^c$, $\mathbf{b}$ and $\mathbf{b}^c$ are included in *half*. It also suggests that no tail event with a relative frequency in the vicinity of 0.5 can be in *E*. That these tail events are precluded from *E* gives the first indication that our path leads away from events with well defined probabilities. We may avoid the flipping of probability from zero to one by including only parts of these tail events in *E*.

We need to make many, many, many decisions of this type. We get a rough estimate of the number by noting that there are as many elementary events as there are members of the power set of the natural numbers, that is the set of all subsets of the natural numbers.[221] We then make about that many choices of inclusion between each elementary event and its complement. This suggests that the number of ways of forming distinct *E*s is two orders of infinity higher than the natural numbers:[222] it has the cardinality of the power set(power set (natural numbers)). There are *very* many possible events *E*!

The supposition, here, is that, if we persist in adding elementary events to *E* prudently, we will arrive at a set conforming with all the conditions. In particular, exhaustion IV will be

---

elementary event in *E* would be $\mathbf{a}_{\text{no-H}}$ so that $\mathbf{a}_{\text{no-H}}$ must also be in *E* by III, which then violates I. But if $\mathbf{a} \cap \mathbf{fin}(1) = \mathbf{fin}(1)$ for all $\mathbf{a}$ in *E*, then the free condition V is violated.

[220] They are equivalent to a Fréchet filter.

[221] Each subset of the natural numbers corresponds to an elementary event. The odd numbers {1, 3, 5, …} corresponds to <H, T, H, T, H, …>.

[222] A more precise analysis shows that this is the cardinality of the set of ultrafilters on the natural numbers. See Comfort and Negrepontis (1974, p. 147).

satisfied. This is an apparently innocent supposition and essential to the formation of E. It is, however, a non-constructive assumption of existence. We have not specified just which elementary events can be added to satisfy exhaustion IV and, were we to try, our efforts to do so would fail. The existence assumption turns out to be of a similar character to the axiom of choice described above. More precisely, the existence of E is proved by the ultrafilter theorem. Its proof commonly employs Zorn's lemma, which is equivalent to the axiom of choice. However, the ultrafilter theorem is logically weaker than the axiom of choice, as displayed in Herrlich (2006, p. 18).

Nonetheless all the vacillations that surround the earlier construction of the Vitali sets arise again here. As reported above, my view is that we should persist in exploring these systems. To do otherwise is to prejudge the admissibility of axioms like the axioms of choice and thus to restrict artificially the scope of our inductive logics.

7.3 Event E is Probabilistically Nonmeasurable

We can now prove that any event E conforming with the conditions I.-V. is nonmeasurable. For purposes of a reductio argument, assume that event E is measurable; and thus so also is its complement event E^c, the set of all elementary events not included in E. We will find that

(a) from a symmetry, $P(E) = P(E^c) = 0.5$; and

(b) since E is a tail event, by the Kolmogorov Zero-One Law, $P(E) = 0$ or 1.

Since (a) and (b) contradict, the reductio is completed. The set E is not measurable.

To see (a), note that there is a one-one correspondence between elementary events in E and those in E^c: each $\mathbf{a} \in E$ corresponds to $\mathbf{a}^c \in E^c$. To implement the correspondence, we just flip H to T and T to H in each elementary event $\mathbf{a}$. It follows that each set of elementary events $\mathbf{a}$ in E is mapped to a corresponding set in E^c with a mirror image structure, under the flipping of H and T. Thus, if a probability is defined for the first set, then the corresponding set has the same probability. An easy way to see this is to note that we turn some set of elementary events in E into the corresponding set in E^c, without making any changes to the physical tosses, merely by imagining that the labels on each of the coins is switched from H to T or T to H. If follows that, if E is probabilistically measurable, then so is E^c; and they have the same probability. Since $P(E) + P(E^c) = 1$, we infer that $P(E) = P(E^c) = 0.5$.

To see (b), consider some elementary event $\mathbf{a} \in E$. Let $\mathbf{b}$ be any elementary event that differs from $\mathbf{a}$ in finitely many of its toss outcomes. From exhaustion IV, we have that one of $\mathbf{b}$ or $\mathbf{b}^c$ is in E. If $\mathbf{b}^c$ is in E, then so must $\mathbf{a} \cap \mathbf{b}^c$. But since $\mathbf{a}$ and $\mathbf{b}^c$ agree only on finitely many

toss outcomes, it follows that $\mathbf{a} \cap \mathbf{b}^c$ has only finitely many H. We saw above that all elementary events in E have infinitely many H. Therefore $\mathbf{b}$ is in E. That is, for every elementary event in E, the event E also contains every other elementary event that differs from it in only finitely many toss outcomes. Recalling *Tail event characterization 2* above, it now follows that E is a tail event. By the Kolmogorov Zero-One Law, it has probability zero or one.

8. The Ultrafilter Logic

The analysis above shows that probabilistic reasoning over the outcomes of infinitely many coin tosses cannot proceed if our considerations include the very many nonmeasurable events of type E. The probability calculus falls silent over them.[223]

There are so very many elementary events arising with infinitely many coin tosses that we run into problems with standard methods even prior to attempting probabilistic analysis. For example, we might try to characterize the event consisting of all elementary events in which there are (in some sense) more heads than tails. One natural approach employs limits. We consider a finite sequence of coin tosses and compute the ratio of the number of heads to the number of tails. The event of interest consists of all elementary events in which that ratio is greater than one. We then take the limit as the number of coin tosses goes to infinity. The event that results will be something less than what we sought. For it is easy to contrive elementary events for which the ratio in question has no limit. All of these will be omitted from the event.

Should we despair of inductive inferences that encompass all the elementary events of the infinite coin toss? It turns out that, if we are willing to consider rather weak systems of inductive logic, we can find one that applies. It is embodied in the conditions I.-V. of the last section that characterizes an ultrafilter. A popular way of explaining the import of an ultrafilter is that it is a specification of which sets are large. In this case, a set of elementary events satisfying the conditions I.-V. contain a large number of H; all the rest do not. What makes this a natural understanding is that these conditions admit only elementary events with infinitely many H; and condition II explicitly continued to populate E with all those elementary events with more H in them. The notion of "large" at issue here is, in intuitive terms, vague. Let us simply turn this around and assert that what we mean by "large" is membership in some set E that conforms with I.-V.

---

[223] We can get no help from upper and lower probabilities. Blackwell and Diaconis (1996) also show that the lower probabilities of both E and E^c are zero. Thus the correspondingly intervals are maximally large.

What results is a two-valued inductive logic that responds to the evidence that the actual outcome of infinitely tosses contains many H. The elementary events in *E* are "supported" (one value) by the evidence as having many H. The remainder are "not supported" (the other value). The axioms of the logic are the conditions I.-V. above. They play the same role as the Kolmogorov axioms of probability theory.

There are infinitely many possible sets *E* of elementary events. This infinity enables the logic to have a dynamics loosely akin to that of conditionalization in probabilistic analysis. We start out with the choice of applicable *E* left entirely open. This is as evidentially neutral a starting point as the logic admits. We can then carry out the analog of conditionalization by restricting the admissible sets *E* to those with some particular elementary event or some set of elementary events. Loosely speaking this restriction introduces the new information that something in these elementary events is close to the actual outcome. More precisely, to conditionalize on some elementary event **a** in this way is say that some infinite subsequence of **a** must be common to all elementary events in *E*. For axiom III., in conjunction with the other axioms, requires that every elementary event in *E* have an intersection with **a** that has infinitely many H in it.

As with the probability calculus, there are restrictions on the events on which we can conditionalize. In the ordinary probability calculus, we cannot conditionalize on events with zero probability. Correspondingly, if we have conditionalized on a set of elementary events containing

$$\mathbf{a} = \langle H, H, H, H, T, H, H, H, H, T, H, H, H, H, T, \ldots \rangle$$

we cannot then conditionalize on a set containing its complement

$$\mathbf{a}^c = \langle T, T, T, T, H, T, T, T, T, H, T, T, T, T, H, \ldots \rangle$$

For the axioms preclude membership of both in *E*.

The logic is weak. It is merely two-valued and, as a practical matter, no finitely specifiable set of evidence will lead to complete determination of the membership of *E*. For, as we have seen, the existence of ultrafilters must be assumed without a finite recipe for the construction of any one of them. If we exclude highly contrived fantasies, I cannot now think of a factual scenario whose background facts would require axioms I.-V. to govern our inductive inferences.

The value of the logic lies in reminding us that many logics of inductive inference are possible. If we infer probabilistically over outcomes of infinitely many coin tosses, we do arrive at many strong results. However their cost is all these inferences fall silent over the nonmeasurable events. If we are prepared to accept a weaker inductive logic, then we see that there is a logic native to the mathematical structure that does embrace all events.

9. Conclusion

The considerations of this chapter have been wide-ranging. They are, however, unified by a single question. How might an inductive logic represent the uniformity of chances over an outcome set of continuum size? It might have seemed that this is an easy case for a probabilistic logic. Is it not realized by a uniform probability density over some continuum-sized set such as the interval [0,1]? That proves not to be the case. If we define the uniformity of chances through the requirement of label independence, the inductive logic that arises is very far from a probabilistic logic.

The bulk of the chapter has tried to find how we may alter the requirement of uniformity until it matches what the probability calculus can provide. These alterations were introduced by weakening the requirement of label independence until we arrived at a version adapted to a background spatial metric. Even this weakening and the addition of background metrical structure met with limited success. For the inductive logic adapted to spaces of infinite area or volume is not probabilistic. Further, nonmeasurable sets arise in spaces of finite area and volume. They escape the reach of a probability measure if its probabilities are to match the spatial areas and volumes. The only escape from this last problem seems to be to find reasons to ignore these sets. That they are non-constructible is a tempting way to banish them from our consideration. However this escape comes at the cost of supposing that all that exists in mathematics and in the physical world described by mathematics is what we can construct by our meager, finite methods.

References

Blackwell, David and Diaconis, Persi (1996) "A Non-measurable Tail Set," Statistics, Probability and Game Theory: IMS Lecture Notes—Monograph Series, **30**, pp. 1- 5.

Brunner, Norbert; Svozil, Karl; Baaz, Matthias (1996) "The Axiom of Choice in Quantum Theory," *Mathematical Logic Quarterly*, **42**, pp. 319-40

Bondi, Hermann (1960) *Cosmology*. 2nd ed. Cambridge: Cambridge University Press.

Bondi, Hermann and Gold, Thomas (1948) "The Steady-State Theory of the Expanding Universe," *Monthly Notices of the Royal Astronomical Society*, **108**, pp. 252- 70.

Comfort, W. Wistar and Negrepontis, Stylianos (1974) *The Theory of Ultrafilters.* Berlin: Springer-Verlag.

Herrlich, Horst (2006) *Axiom of Choice*. Berlin: Springer-Verlag.

Kharazishvili, A. B. (2004) *Nonmeasurable Sets and Functions.* Amsterdam: Elsevier.

Kolmogorov, Andrei N. (1950) *Foundations of the Theory of Probability*. Trans N. Morrison. New York: Chelsea Publishing Company.

Norton, John D. (2008) "Ignorance and Indifference." *Philosophy of Science,* **75**, pp. 45-68.

Stoll, Robert R. (1979) *Set Theory and Logic*. New York: Dover.

Rosenthal, Jeffrey S. (2006) *A First Look at Rigorous Probability Theory*. 2nd ed. Singapore: World Scientific.

Isaacs, Rufus (1975) "Two Mathematical Papers without Words," *Mathematics Magazine*, **48**, p. 198.

Wagon, Stan (1994) *The Banach-Tarski Paradox*. Cambridge: Cambridge University Press.

Wapner, Leonard M. (2005) *The Pea and the Sun*. Wellesley, MA: A. K. Peters, Ltd.

December 3, 2016; May 15, 2017
References to Ch 10 added August 7, 2017

Field that initiates equation numbering:

Chapter 15

Indeterministic Physical Systems

1. Introduction

The indeterministic systems to be investigated in this chapter share the common characteristic that determining one aspect of the system leaves others open. The most familiar cases are ones in which the present state of the system fails to fix its future state. We shall see several such systems here in Section 3. The most important are systems with infinitely many degrees of freedom, for this sort of determinism is generic amongst them. Rather than delve into the details of the physics of such systems, the mechanism that generates the indeterminism will be illustrated by the simplified system of the infinite domino cascade.

A different sort of indeterministic system will be explored in Section 4. At the risk of abusing the term, I will also describe as indeterministic systems in which, at the same moment of time, one component fails to fix others, contrary to normal expectations. The examples will be drawn from Newtonian gravitation theory.

Each instance of indeterminism poses a problem in inductive inference. From the known aspect, what strengths of inductive support are provided to the remaining underdetermined aspects? Given this present, what support is provided to the various possible futures? Given this mass distribution, what support is given to the various possible Newtonian potential fields? As explained in Section 5, each of the problems has been chosen so that the complete background physics is transparent and transparently provides no probabilities over the various underetermined possibilities. The problem for inductive analysis is to find the strengths of inductive support for the different possibilities, without altering or adding to this physics. For to do otherwise is to change the problem posed.

We shall see in Section 6 that probabilities can only be assigned as strengths of inductive support if we add to the background facts. Normalization of a probability measure, for example, requires that the probabilities of different times of spontaneous excitation in a temporally indeterministic system diminish to zero as the times grow large. This diminution must happen at

some rate: quickly or slowly; and fitting a probability measure to the process requires that some speed be chosen. To make that choice, however, is to add to the physics provided.

This is just the first of a series of problems that preclude the use of probabilities as strengths of support. The final example requires the adaptation of a uniform probability measure to an infinite dimensional space of Newtonian potentials. The infinity of the dimensions present especially intractable problems.

Section 7 then describes how the material theory of induction solves the inductive problems. We are to look to the background physical facts to provide the strengths of inductive support. By design, these facts provide very little. They allow us to say of various processes or components that they are necessary, possible and impossible. These three evaluations become the values of a spare, three-valued inductive logic. Its strengths of support coincide with those of "completely neutral support" described elsewhere, including Chapter 10 here. This completely neutral support can be fixed by certain invariances in space of possibilities; and we shall see that they are realized in this case as well.

We proceed first with a preliminary in Section 2 on the project now undertaken.

2. Why Take Simple, Unrealistic Physical Systems Seriously?

The illustrations to come involve simple, physically unrealistic systems that, mostly,[224] we will not encounter in the ordinary practice of science. So why pay any special attention to them in investigations of inductive inference? There is a simple pragmatic reason for considering them. If the analysis of the warranting relations is to be transparent, we need simple systems. We need systems in which the full set of background facts is easy to comprehend, so that their full import can be seen clearly and unequivocally.

This pragmatic reason, however, is not the principal one. The deeper reason for taking these simple systems seriously pertains to the range of applicability of inductive inference. We do not balk at reasoning deductively about fictitious systems, not matter how bizarre we may find them. Correspondingly, I see no reason to prohibit inductive inference over such systems. There is no guarantee, of course, that every system will admit rich inductive inferences. Just what is possible inductively will be determined by the background facts that obtain, as the material theory of induction asserts. When we ask which inductive inferences are warranted in the simple systems below, we will find that their strengths of inductive support cannot be probability measures. That is, we will find through counterexamples that the probability calculus does not provide a universally applicable logic of induction.

[224] The exception is the example of the quantum spin of electrons.

It may be tempting to block the counterexamples by insisting that the scope of inductive inference is limited to ordinary physical systems of the type we normally encounter in science. This would be an unnecessary restriction on the reach of inductive methods. Worse, it would be of no help in protecting the probability calculus as the universally applicable logic of inductive inference. For the restriction to ordinary systems gives up universal applicability at the outset. Moreover the restriction itself would conform with the material theory of induction, for the range of applicability of probabilistic inductive logic would be circumscribed by the factual restriction to ordinary systems.[225]

3. Temporally Indeterministic Systems

The general idea of determinism is that the fixing of one aspect of a system fixes some other. This section will address the case of temporal aspects. In a (temporally) deterministic physical system, the present state of the system determines its future states. With the notable exception of quantum measurement, physical systems are generally assumed to be deterministic. The present state of the planetary system fixes the future movements of the planets and whether there will be an eclipse on any nominated time.

Systems that violate temporal determinism have attracted considerable attention in recent decades in philosophy of physics, with the modern era marked by the publication of John Earman's *Primer* (1986). Once we start to look for indeterministic systems, we find them in many places.

3.1 The Dome

One of the simplest indeterministic systems is the "dome." Since it has been discussed extensively elsewhere (Norton 2003, §3; 2008), it needs only a brief recapitulation. A unit point mass slides frictionless over the surface of a dome in a vertical gravitational field with acceleration due to gravity g, as shown in Figure 1.

[225] I set aside here the further problem of delineating just what will count as "ordinary." Many of the systems ordinarily considered in science are highly idealized and thus highly unrealistic.

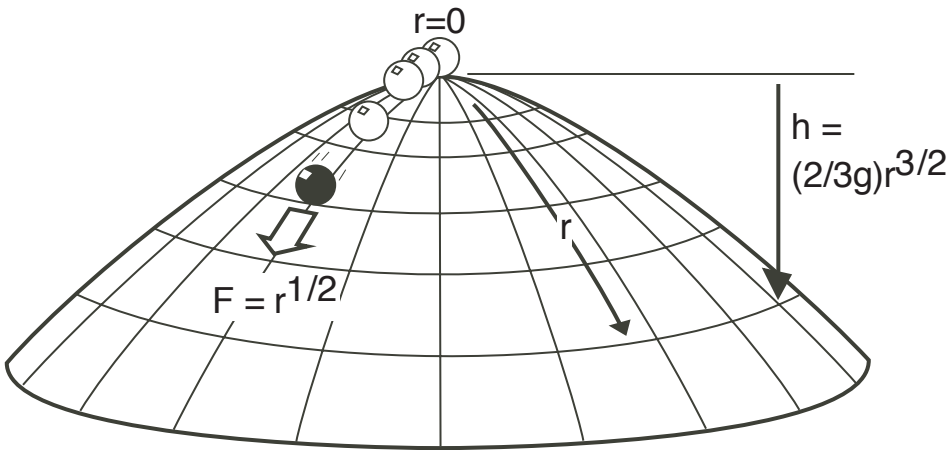


Figure 1. The Dome

The dome has a vertical axis of rotational symmetry about its apex and the surface is depressed below the apex by a (negative) height $h = (2/3g)r^{3/2}$, where r is the radial distance to the point from the apex along the surface. The force F on the point mass along the surface of the dome is

$$F = (\mathrm{d}/\mathrm{d}r)\, gh = r^{1/2}$$

and is directed outward from the apex. The motion of the point mass is governed by the equation of motion

$$\frac{d^2r}{dt^2} = r^{1/2} \tag{1}$$

where t is time. Initially, at time t=0, the point mass is located at the apex r=0 at rest. Since the force at the apex is $F = 0^{1/2} = 0$, one solution to the equation of motion is that the mass remains at the apex for all time:

$$r(t) = 0 \qquad \text{all } t$$

However there is a second family of solutions, in which the particle moves spontaneously at time t=T for any time T≥0:

$$r(t) = (1/144)\,(t - T)^4 \qquad \text{all } t \geq T$$
$$= 0 \qquad \text{all } t \leq T$$

In this second solution, the particle remains quiescent up to and including time t=T. Then it moves away from the apex in any direction.

This spontaneous excitation results entirely from the equation of motion. There is no hidden triggering event, such as a slight bump to the dome that may dislodge the point mass from the apex. If there is no spontantoues motion, it is so because the equations of motion allow it. If there is spontaneous motion at time T, it happens just because the equation of motion also allow it.

The dome is a Newtonian system with only finitely many degrees of freedom. That is, its state can be specified fully just by specifying a finite list of magnitudes: the position of the particle on the dome, its speed and its direction of motion. The dome is unusual in its indeterminism in that, generally, Newtonian systems with finitely many degrees of freedom are deterministic. It was devised originally to display an unusual exception to this generality. Because of its exceptional character, the indeterminism of the dome is highly sensitive to changes in the physical system and its indeterminism can be eliminated by small adjustments to it.

3.2 Masses and Springs

Matters change, however, once we consider Newtonian systems with infinitely many degrees of freedom. An important example is a system of infinitely many interacting particles. It has infinitely many degrees of freedom since its state can only be specified by specifying infinitely many magnitudes, such as a the mass, position and velocity of each particle. Such systems are generically indeterministic. While circumstances need to be specially contrived to induce indeterminism among the systems with finitely many degrees of freedom, indeterminism is simply the standard, generic behavior of these systems with infinitely many degrees of freedom. There are many examples in the literature. Often they arise in the supertask literature, as reviewed in Manchak and Roberts (2016).

The masses and springs example consists of an infinite chain of mass-spring-mass-spring-… shown in Figure 2.

Figure 2. Masses and Springs

Its temporal behavior is recovered from an application of Newton's laws along with Hooke's laws for the springs. If the system is set initially in equilibrium with all the masses at rest and the springs unextended or uncompressed, then the system can remain in this quiescent state indefinitely. However, at any later moment, it can spontaneously self-excite with all the masses set in motion. The system is noteworthy for the ease with which a full mathematical description can be given and for what is represents physically. It is a standard model of a one-dimensional crystal, extended to infinite size. It indicates that more complex solids, such as infinite three-dimensional crystals, will exhibit similar indeterminism.[226]

[226] I have argued in Norton (2012) that this fact ensures that the infinite component, thermodynamic limit of thermal physics cannot involve examination of a system that consists of infinitely components. Through their indeterminism, such infinite systems have qualitatively

In all these systems, the infinity of the number of degrees of freedom is essential. A finite system, no matter how large, will not manifest the indeterministic behavior as freely. A finite chain of mass-spring-mass-spring-…, once quiescent, remains so for all time, no matter how large it is.

3.3 The Infinite Domino Cascade

Rather than work through the technical details of the examples, I will display a toy example, shown in Figure 3., that illustrates the mechanism that brings about indeterminism in all these infinite cases. In a domino cascade, dominoes or slender tiles are set on their edges in a row, such that when one falls, it strikes another, leading it to fall; that falling domino strikes yet another, leading it to fall; and so on down the row.

Consider a very large row of dominoes, finite in number. We assume no external perturbing effects. There are no slight vibrations from passing trucks, no thermal agitation from air molecules, and so on. If it is set up at rest initially, it will remain so indefinitely.

Consider an infinite row of dominoes with the same provisions. As with the finite case, it *can* remain at rest indefinitely. However, it is also possible for it to be set into motion spontaneously. The final stages of this spontaneous motion are:

- the first domino falls, because it was struck by the second domino that started falling earlier;
- the second domino fell, because it was struck by the third domino that started falling earlier;
- the third domino fell, because it was struck by the fourth domino that started falling earlier;

and so on.

As we proceed through the falling of the first, second, third, … dominoes, we trace the process back through time and eventually consider the falling of all infinity of the dominoes.

---

different properties from the real target of analysis, systems with many, but finitely many, components.

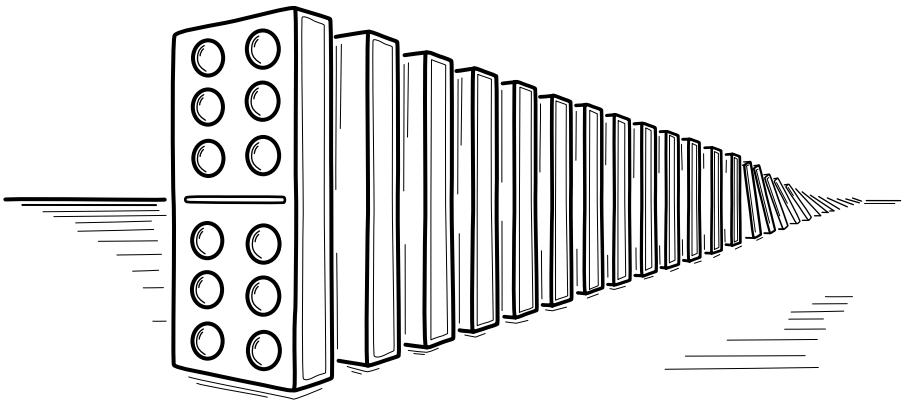


Figure 3. Infinite Domino Cascade

This cascade of falls could not happen spontaneously if there were finitely many dominoes. For, as we trace back through the finite cascade, we would eventually come to the last domino. It would not fall because there are no further dominoes to fall on it. There is nothing to start the cascade. In the infinite case, we never come to the end of the cascade. For any domino, there is always a next domino to fall on it. So every domino falls. There is no first fall to initiate the cascade and no need for one.

All that remains now is to close a loophole. If each domino takes the same amount of time to fall onto the next, then the infinity of domino falls needed to complete the cascade requires an infinite time. That does not make the process impossible. Rather it makes it uninteresting for our purposes, for it is simply a process that has been underway for all of an infinite past time. If each fall takes one second, then the Nth domino fell N seconds ago; and so on for N indefinitely large.

We close the loophole by contriving the geometry of dominoes such that each time of fall is successively shorter as we proceed along the cascade. If the successive dominoes require 1/2, 1/4, 1/8, 1/16, … seconds to fall, then all infinity of them will have fallen after 1/2 + 1/4 + 1/8 + 1/16 + … = 1 second. To an observer, the motion would appear as follows. The initially quiescent dominoes remain so for some time. Then, off in the distance of the infinite end of the row of dominoes, at the moment of spontaneous excitation, there is a disturbance that rapidly propagates towards the beginning of the row and leaves all the dominoes toppled.

Some delicacy is needed to arrange all the dominoes so that they can behave this way. The time each takes to fall on the next will depend on how hard it is struck and how close is the next domino. Under plausible assumptions, computed in Appendix A, the time each domino needs to fall onto the next scales in direct proportion to the distance between the dominoes. Thus we secure the above schedule of acceleration of the falls by shrinking the distance between the dominoes in proportion to the times 1/2, 1/4, 1/8, 1/16, … If we assume that the widths of the

dominoes are scaled similarly, then the cascade can be completed in finite time just if the length of the domino row is finite.

One outcome of this scaling is that the dominoes will become arbitrarily thin. One might imagine that this means that the dominoes become pseudostable rather like a pencil balanced on its infinitely sharpened tip. However none of the dominoes will be pseudostable, since a pseudostable system is one which is toppled by an arbitrarily small perturbation. Each domino will have a finite width, even if small, which forms a stable base. Toppling it requires some non-zero work to lift its center of mass past its edge.

This is a toy model. However it illustrates how indeterminism arises generically in systems with infinitely many degrees of freedom. In such systems there are many cascades of excitation processes that cannot arise spontaneously in finite systems, since the finite system requires some initiating event to get the process started. In a system with infinitely many of degrees of freedom, these processes can happen spontaneously without need of some initiating event, for they are comprised of infinite cascades of events that have no first member.

These general remarks can be made more precise. For a synopsis of the analysis for a more general case and for the quantitative analysis of the masses and springs example specifically, see Norton (2012, Appendix).

4. Indeterminism Among Components of a System

In the indeterminism of the last section, the present state of the system fails to fix its future state. It may also happen that, at the same time, the state of some components of a system may fail to fix the state of other components, contrary to our expectations. The problem in inductive inference is then to determine the strengths of support afforded to these incompletely determined components.

4.1 Gauge Systems

There is a simple recipe for generating many problems of this type by injecting a small fiction into physics. Modern physical theories are replete with gauge freedoms. They arise when one has two descriptions that appear to be of distinct physical systems, but it turns out that the differences are merely artifacts of the descriptions used. It is "the Eiffel tower" and "la tour Eiffel." The two systems are the same physically. They just differ in their names.

Imagine, however, that through some novel physics we do find a way to distinguish the two. Then we would have a difference that makes a difference; and learning which is the correct one would become a problem in inductive inference. Since there are many gauge freedoms in

modern physics, this stratagem can create many new inductive inference problems of just the type sought here.

Sometimes fact can mimic fiction. The gauge field associated with magnetism is the vector potential $\boldsymbol{A}$. In classical physics, it is merely a useful adjunct in computing magnetic field strengths, but not a physically significant quantity in its own right. The coming of quantum theory initially showed promise of changing this circumstance. Bohm and Aharonov (1959) found a quantum effect that arose when there was an $\boldsymbol{A}$ field present, but no magnetic field. They initially offered it as evidence that the $\boldsymbol{A}$ field is physically significant. Later analysis showed the situation to be more complicated.

For concreteness, I will elaborate one of the simplest gauge freedoms. In ordinary Newtonian gravitation theory, the physically significant quantity is the gravitational force on a unit test mass and the associated quantities of work. The distribution of all such possible forces over all space is the Newtonian gravitational force field $\boldsymbol{f}$. For the case of the sun, the force field is given by the familiar inverse square law

$$f(r) = \frac{GM}{r^2} \tag{2}$$

where a force of magnitude $f(r)$ on a unit test mass is directed towards the center of the sun. M is the mass of the sun, r the radial distance from the center of the sun to the test mass and G the universal constant of gravitation.

The Newtonian gravitational potential field $\varphi(r)$ is defined through the work $W(r_0, r_1)$ needed to be performed against this force field when we move a unit test mass from one position r_0 to another r_1. That is, the potential fields $\varphi(r_0)$ and $\varphi(r_1)$ are related by

$$W(r_0, r_1) = \varphi(r_1) - \varphi(r_0) = \int_{r=r_0}^{r_1} f\, dr = \int_{r=r_0}^{r_1} \frac{GM dr}{r^2} = -\frac{GM}{r_1} - \left(-\frac{GM}{r_0}\right) \tag{3}$$

We usually infer from (3) that $\varphi(r) = -\frac{GM}{r}$. However we are really only authorized to infer to something weaker:

$$\varphi(r) = -\frac{GM}{r} + K \tag{4}$$

where K can be any number, positive or negative, large or small.

The choice of K leaves the physically significant quantities unaltered. That is, for all K we end up with the same work term $W(r_0, r_1)$ in (3) since

$$\left(-\frac{GM}{r_1} + K\right) - \left(-\frac{GM}{r_0} + K\right) = \left(-\frac{GM}{r_1}\right) - \left(-\frac{GM}{r_0}\right)$$

and the same force field $f(r)$ in (2) since

$$f(r) = -\frac{d}{dr}\left(-\frac{GM}{r} + K\right) = \frac{GM}{r^2}$$

The freedom in selection of different *K's* is a gauge freedom and transforming between different, physically equivalent expressions for $\varphi(r)$ by changing the value of K is a gauge transformation.

The inductive inference problem posed is this. We introduce the fiction that some new physics will enable us to detect and distinguish among the gravitational potentials of (4). Given the gravitational force field $f(r)$ of the sun (2), what is the inductive strength of support given to the gravitational potential fields $\varphi(r)$ of (4) with different values of K?

4.2 Newtonian Cosmology

Indeterminism among components in a physical theory can arise without need for any fictitious physics. A simple example, inspired by Wallace (2016), arises in Newtonian gravitation theory. We expect that the specification of the position and masses of all bodies in the universe will fix the gravitational force on a test body and the gravitation potential field at any point in space.

That things are not simple precipitated an acute problem in Newtonian cosmology in the late 19th and early 20th century. Newtonian cosmology assumes that infinite Euclidean space is filled with a uniform matter distribution of constant density ρ. The expectation is that there is a unique gravitational force acting on any test body in such a universe. That force is calculated by summing all the gravitational forces acting on the test body from the uniformly distributed cosmic matter. The trouble is that there are many ways to sum these forces. Pick any resultant force you like and there is a way to carry out the sum so that the net force on the test body is just that force. For a survey of this period and for an example of the simple calculations that lead to the multiplicity of forces, see Norton (1999a).

In retrospect, the difficulty is all too easy to see. Contrary to expectations, the cosmic matter distribution does not fix the net gravitational force on the test body. Many fields are compatible with the one matter distribution and thus we can compute many forces on the test body simply by drawing quantities from different possible fields.

At the time, however, this possibility was overlooked since the loss of uniqueness of the force was unthinkable. Instead, many physicists found it obvious and even compelling that the symmetries of the problem must force a unique solution: there can be no preferred directions in a homogeneous, isotropic cosmology. So the net force can point in no direction. Hence there is no net force on the test body and, as a result, the gravitational potential field must everywhere be a constant.

We shall return below to this risky idea that physical intuition can override what well-established equations say. Before we do, it is interesting to note that a favored resolution was to modify Newton's inverse square law of gravity until it returned the expected constant gravitational potential. This computation was used by Einstein in 1917 as a foil to motivate his introduction of the cosmological constant into general relativity.

We can develop the difficulty as follows. A curious result of Newtonian gravitation theory concerns an infinite flat plate of matter of density ρ and thickness Δx. The gravitational force exerted by this plate on a test body turns out to be independent of the distance from the plate. It is just

$$f = 2\pi G\rho\, \Delta x. \tag{5}$$

directed along the line of shortest distance to the plate. (See Appendix B for a justification and demonstration of this result and further analysis of this example.) We can use this result to determine the gravitational force on a test body in a Newtonian cosmos. We divide the uniform matter distribution into infinitely many flat plates of thickness Δx and infinite area, arranged parallel to the y and z axes of a Cartesian coordinate system (x, y, z).

Consider a unit test mass at some fixed x-coordinate position, say $x'=x$. We can divide the matter distribution that acts gravitationally on it into two parts. As shown in Figure 4, the first consists of all those infinite plates between x'=-x and $x'=x$. The second consists of all the remaining infinite plates.

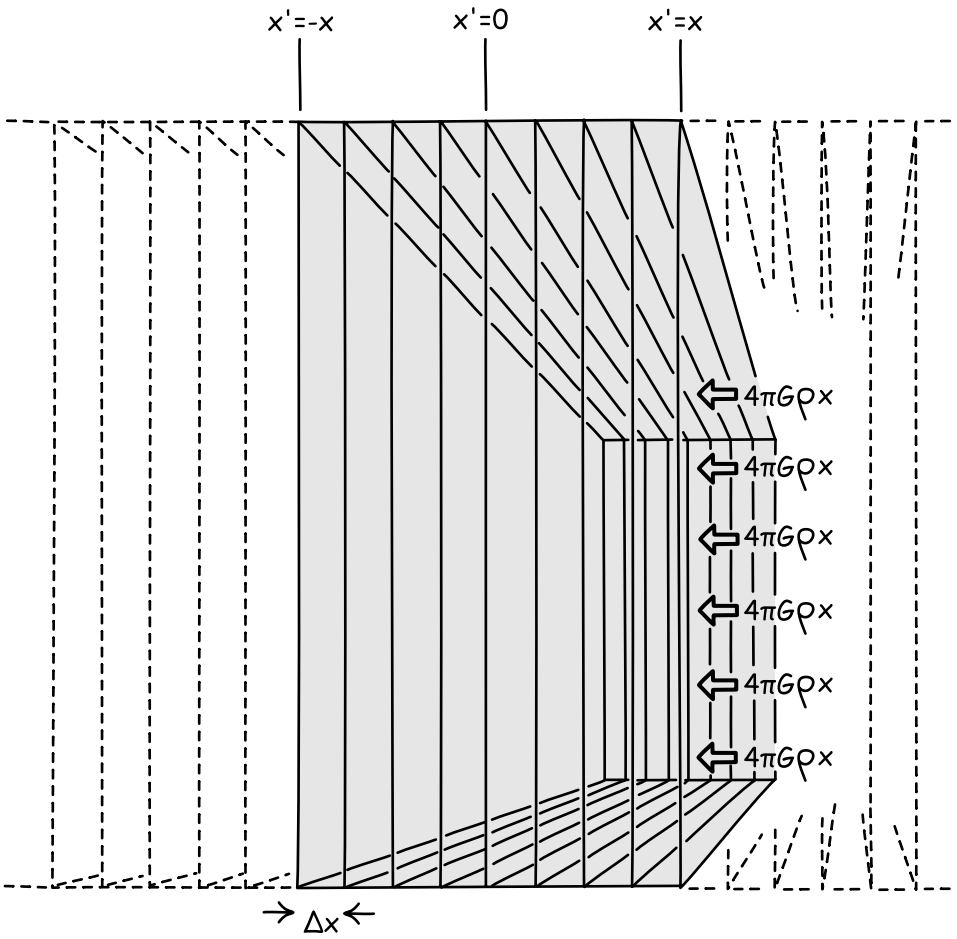


Figure 4. Unbalanced Forces in Newtonian Cosmology

We have from (5) that each plate of thickness Δx contributes force $2\pi G\rho\,\Delta x$. Hence the force on the test body from the plates between x'=-x and $x'=x$ is just their sum

$$f(x\ y, z) = 4\pi G\rho x. \tag{6}$$

and is directed along the x-axis towards x=0. The remaining plates each exert the force $2\pi G\rho\,\Delta x$ on the text body. The force will be in the $+x$ direction if the plate is located at $x'>x$ and it will be in the $-x$ direction if the plate is located at $x'<-x$. Hence we can pair up the plates at coordinate positions $+x'$ and $-x'$, matching one that exerts a force in the $+x$ direction with one that exerts a force in the $-x$ direction, so the net force from the pair is zero. This pairing exhausts all the matter of the second part, as shown in Figure 5. The net result is that the force on the test body is given by (6).

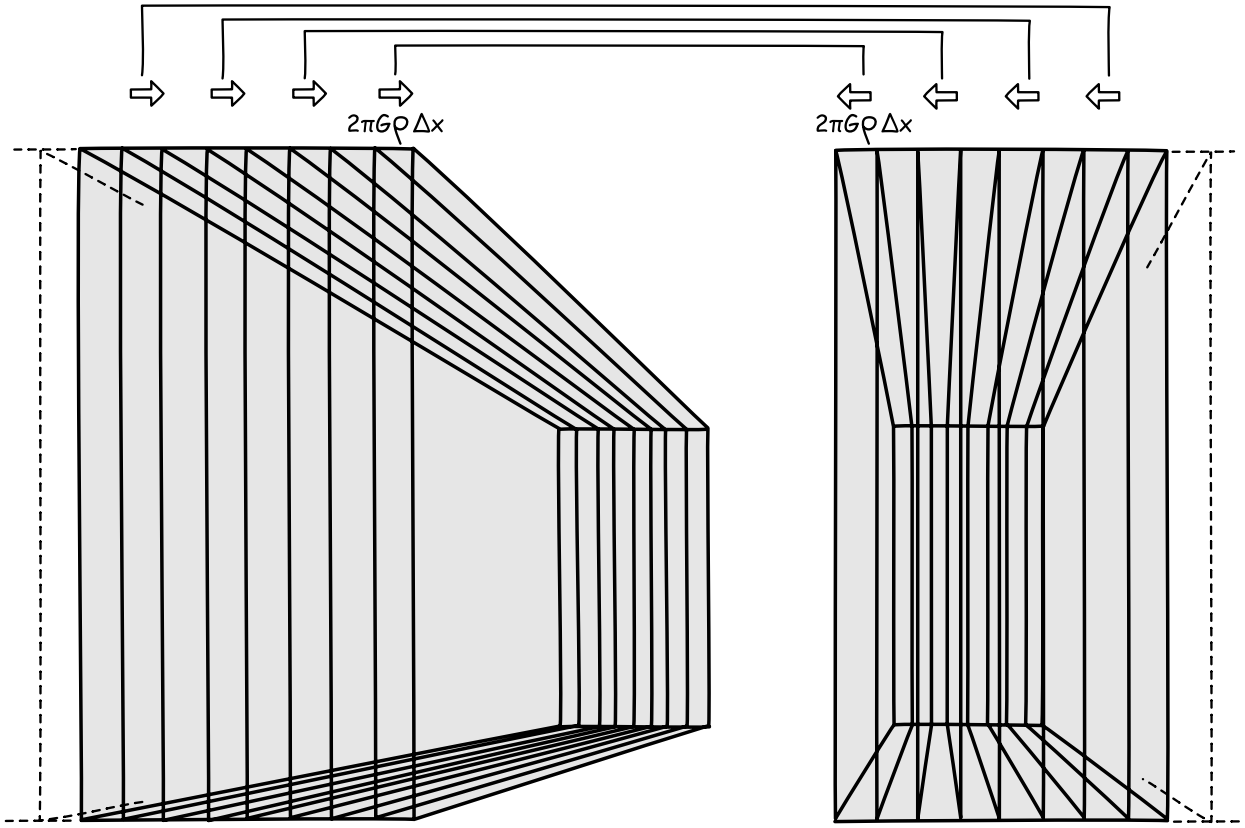


Figure 5. Balanced Forces in Newtonian Cosmology

We can repeat this construction for every point in space, so that the expression (6) represents the gravitational force field due to the cosmic matter. This force field induces a gravitational potential through a relation analogous to (3) as

$$\varphi_x(x,y,z) = \int_{x'=0}^{x'=x} f(x',y,z)\,dx' = 2\pi G\rho x^2 \tag{7a}$$

The problem should now be obvious. The division of the cosmic matter into plates perpendicular to the *x* axis was arbitrary. We could also have divided it into plates perpendicular to the *y* or the *z* axes. We could then replicate the above analysis and recover two distinct potential fields[227]

$$\varphi_y(x, y, z) = 2\pi G\rho\, y^2 \tag{7b}$$

$$\varphi_z(x, y, z) = 2\pi G\rho\, z^2 \tag{7c}$$

We can generate still further potential fields. Another arbitrary choice was to locate the center of the plates of the first part at *x*-coordinate 0. We could equally have chosen any *x*-coordinate, such as x_0. We would then have arrived at the gravitational potential fields

[227] For experts: the potentials (15a, b, c) derive from physically distinct gravitational systems and not gauge equivalent along the lines of Malament (1995). For more, see Appendix B.

$$\varphi_{x,x_O}(x, y, z) = 2\pi G\rho\,(x - x_0)^2 \tag{8a}$$

$$\varphi_{y,y_O}(x, y, z) = 2\pi G\rho\,(y - y_0)^2 \tag{8b}$$

$$\varphi_{z,z_O}(x, y, z) = 2\pi G\rho\,(z - z_0)^2 \tag{8c}$$

Taken together, we have many potentials compatible with the cosmic matter distribution. One might well suspect at this point, quite correctly, that we have only begun to explore the gravitational potential fields compatible with the cosmic matter distribution.

These potential fields form a large space and we can navigate through them by the following artifice. We start with any admissible potential, such as (7a). We arrive at another simply by adding a "harmonic function" to it. (A harmonic function is one that satisfies Laplace's equation $\nabla^2\Phi = 0$. For more, see Appendix B.) It turns out that

$$\Phi = 2\pi G\rho\,(y^2 - x^2)$$

is a harmonic function. Adding it to (7a) moves us to (7b):

$$\varphi_x(x, y, z) + \Phi = 2\pi G\rho\, x^2 + 2\pi G\rho\,(y^2 - x^2) = 2\pi G\rho\, y^2 = \varphi_y(x, y, z)$$

Another harmonic function is

$$\Phi = 2\pi G\rho\,((z-z_0)^2 - x^2)$$

Adding it to (7a) moves us to (8c).

The remarkable fact is that there are infinitely many harmonic functions and they are linearly independent. That means that we cannot reduce the set by expressing some as linear combinations of others. If we represent an infinite set of linearly independent harmonic functions as $\Phi_1, \Phi_2, \Phi_3, \ldots$, then adding any linear combination of them to an admissible potential produces another. Thus we arrive at an infinite dimensioned space of gravitational potentials

$$\varphi_x + a_1\Phi_1 + a_2\Phi_2 + a_3\Phi_3 + \ldots \tag{9}$$

where the space is parameterized by infinitely many parameters $a_1, a_2, a_3, \ldots$ which can each independently take on all values, positive and negative, large and small. The potentials of (7a, b, c) and (8a, b, c) are just some of the simplest potentials in the space.

The inductive problem to be addressed shortly is to determine the support for each of the solutions in the space of potentials defined by (9), given the spatial geometry and matter distribution of Newtonian cosmology.

Since both the spatial geometry and the matter distribution are isotropic and homogeneous, it is natural to assume that the gravitational potential will share some or all of these symmetries. One may even have a strong intuition, as did the physicists of the past, that the potential must share these symmetries. Imposing them would have the effect of greatly reducing

the size of the space of potentials (9). While the resulting reduced problem is interesting its own right, it is not the one to be addressed here. We do not assume homogeneity and isotropy of the potential field, for there is no compulsion to assume either. It is not an assumption that can derived from the corresponding symmetries of the geometry and the matter distribution and, as the viablility of the potentials (9) show, it is not enforced on individual potentials of Newtonian gravitational theory.

5. Inductive Analysis of Temporally Indeterministic Systems

The indeterministic systems of Sections 3 and 4 above each pose a problem in inductive inference. Take certain fixed aspects of a system: its present state or certain of its components. Find the strength of inductive support that aspect provides to some other aspect: the system's future state or certain others of its components. The systems have been chosen so that all share the following two properties:

- The physics described is an exhaustive account of the totality of background facts. There are no further hidden background facts.
- The physics leaves one aspect of the system underdetermined, but provides no probabilities for the different possibilities.

An essential condition to be placed on the inductive analysis is that it merely extracts and displays the relations of inductive support already present in the fully specified systems. That is, setting off the controlling idea for emphasis:

The analysis may not impose new physics.

For to impose new physics is to introduce new facts that alter the problem posed. What would result might well be a cogent analysis of some problem, but it would not be an analysis of the problem originally posed.

6 A Probabilistic Analysis

Let us attempt to represent the strengths of inductive support as probabilities. We shall see that this analysis inevitably imposes new physical facts on the systems.

6.1 Temporally Indeterministic Systems[228]

The temporally indeterministic systems of Section 3 all involve systems that remain quiescent until some time $t=T$ of spontaneous excitation. The inductive problem is to determine

[228] The analysis of this section draws on Norton (2010a).

the strengths of support for various times T. Initially, this looks like a problem tailor-made for probabilistic analysis, for it is similar to the problem of radioactive decay: a radioactive atom remains quiescent until the moment of decay. This moment is governed by the familiar law of radioactive decay. The probability $P(T)$ of decay in the time interval from 0 to T is

$$P(T) = 1 - \exp(-T/\tau) \tag{10}$$

where the time constant τ of the decay is related to the empirically determined half-life of the element by $T_{1/2} = \tau \ln 2$.

This law of radioactive decay is the natural probabilistic law adapted to these cases, for it is the unique law with "no memory" of what happened in the past. That is, whether the atom will decay in the moments immediately to come is independent of how long the atom has survived so far, without decaying. It has no memory of whether that past survival was long or short.

If we write $Q(T) = 1 - P(T)$ for the probability that the atom does *not* decay in the initial time T, then this no memory property is expressible as

$$Q(T+u) = Q(T).Q(u) \tag{11}$$

That is, the probability that the atom survives undecayed for a total time $T+u$ is given by the probability that it survives first for time T and then, given no decay, that it then survives for a further time u. The no memory property says that these last two probabilities are independent, so the probability of the conjunction of their outcomes is just the product of (11). This relation entails the exponential decay law (10).[229]

The probability distribution (10) expresses a physical chance. It is immediately and naturally converted into a logic of induction through the conditional probabilities it induces on pairs of hypotheses concerning the time of decay. For example, write:

$H(T_1, T_2)$: the hypothesis that the time T of spontaneous excitation occurs in the interval $T_1 \leq T < T_2$

If we take as our background B the physical description of the radioactive atom, then the support accrued to the hypothesis from B that the atom will decay sometime up to time T is just given by

$$P(H(0, T)|B) = P(T) = 1 - \exp(-T/\tau)$$

---

[229] Differentiate (10) with respect to u and find $\frac{dQ(T+u)}{d(T+u)} \cdot \frac{d(T+u)}{du} = \frac{dQ(T+u)}{d(T+u)} = Q(T) \cdot \frac{dQ(u)}{du}$.

Evaluate this expression at u=0 and recover $dQ(T)/dT = k\, Q(T)$, where $k = dQ(u)/du|_{u=0}$ is a constant independent of T. The solution is $Q(T) = \text{constant} \cdot \exp(kT)$. Since the atom must eventually decay, $P(T) = 1 - Q(T)$ must go to unity as T goes to infinity. Hence we must have "constant" = 1 and $k = -1/\tau$, for any $\tau > 0$.

The support for the hypothesis of decay between T_1 and T_2, from the evidence that decay happens by time $T > T_2 > T_1$ is

$$P(H(T_1,T_2) \mid H(0,T)) = \frac{\exp(-T_1/\tau) - \exp(-T_2/\tau)}{1 - \exp(-T/\tau)}$$

All this is unremarkable and it seems to be the natural analysis to apply to the spontaneous excitations of Section 3. Here, however, our familiarity with radioactive decay is leading us astray. For the probabilistic law (10) includes a time constant τ. The magnitude of the time constant has a profound effect on the dynamics, as shown in Figure 6.

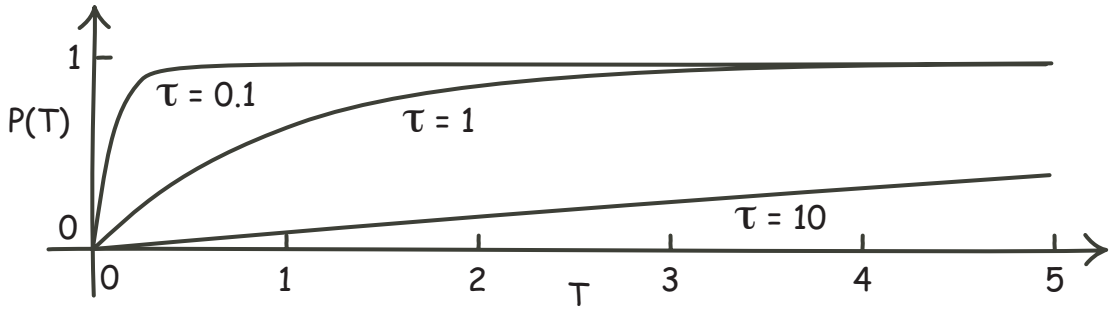


Figure 6. Effect of Different Time Constants τ on the Probability of Spontaneous Motion

A small time constant entails that spontaneous excitation is all but sure to happen soon. If τ is one millisecond, then there is a probability of 0.999 of spontaneous excitation in time τ ln 1000 = 6.91τ = 6.91 milliseconds.[230] A large time constant entails that spontaneous excitation is very unlikely to happen soon. If τ is one thousand years, then there is a probability of only 0.001 of spontaneous excitation in τ ln 1.001 = 0.001τ = one year.

Since use of the probabilistic law (10) requires selection of a time constant τ, it can only be employed if we, in effect, make some judgment about how soon the spontaneous excitation will occur. We already have the complete physics of the systems of Section 3. There is no time scale provided and no judgments of sooner or later. All the physics tells us is that spontaneous excitation is possible.

Thus to apply the probabilistic law (10) is to introduce new physics. That is, it is to change the problem posed to a new one to which probabilistic methods happen to be well-adapted.

---

[230] To arrive at these estimates, invert (7) to recover $T = \tau \ln [1/(1-P)]$.

The analysis above is just a beginning. There are many ways to apply probabilistic analysis to this problem of spontaneous excitation. While some are quite ingenious, none succeed. Here are a few of the possibilities.

The physics is indifferent to which is the time T of spontaneous excitation. So a natural choice is a uniform distribution of probability over all values of T from zero to infinity. The immediate difficulty is that the probabilities of such a uniform distribution cannot sum to unity. We set equal the probability of equal intervals

$$\varepsilon = P(H(0,1)\,|B) = P(H(1,2)\,|B) = P(H(2,3)\,|B) = P(H(3,4)\,|B) = \ldots \qquad (12)$$

Since there are infinitely many of these intervals, the total probability is

$$\varepsilon + \varepsilon + \varepsilon + \varepsilon + \varepsilon + \ldots = \infty \text{ x } \varepsilon = \infty.$$

This is a failure of the probability distribution to normalize: these probabilities should sum to the unit probability required by the axioms of probability theory for the entire outcome space.

While this failure is usually treated as fatal, the normalization condition is sometimes dropped, under the expectation that conditionalization may lead to a normalized probability distribution. However, even if this expectation is sometimes met, the real problem with the distribution (12) is that it still adds to the physical facts. It assures us that, for example, $H(0,2)$ is twice as probable as $H(0,1)$. If we make the usual connections to frequencies, that means that we should expect $H(0,2)$ to arise roughly twice as often as $H(0,1)$ in many repeated trials. The physical facts for these systems include no such provision. They simply allow that any of the times in these hypotheses may be the time of spontaneous excitation; and nothing more.

Another possibility was explored more fully in the earlier Chapter: Infinite Lottery Machines. It is that we drop the requirement of countable additivity that allows us to sum the infinitely many ε's above. Instead, we are allowed to sum finitely many only, that is, we are restricted to finite additivity. The result is that we can set ε=0 in (12) without breaching normalization. All the individual hypotheses of (12) are assigned zero probability

$$0 = P(H(0,1)\,|B) = P(H(1,2)\,|B) = P(H(2,3)\,|B) = P(H(3,4)\,|B) \ldots$$

but their infinite disjunction is assigned unit probability.[231] Finite disjunctions of them are also assigned zero probability

$$P(H(0,3)\,|B) = P((H(0,1) \text{ v } H(1,2) \text{ v } H(2,3))\,|B)$$
$$= P(H(0,1)\,|B) + P(H(1,2)\,|B) + P(H(2,3)\,|B) = 0 + 0 + 0 = 0$$

This is promising initially, since all finite intervals of times are treated equally, even if as zero probability outcomes.

---

[231] Or, more carefully, one less whatever probability is assigned to the hypothesis that there is never a spontaneous excitation.

The difficulty is that the finitely additive measure is still adding significantly to the physics. For even finitely additive measures must assign unit probability to some set of outcomes; and these become privileged as the events we expect to happen. There is no way to assign this privileged set without adding to the physics. For example, the above measure assures us that the time of spontaneous excitation is, with probability 1, greater than or equal to T=1: $P(H(1,\infty)\,|B) = 1$. The physics is equally indifferent to the times of spontaneous excitation as it is to the inverse times of spontaneous excitation, $1/T$. If the finitely additive measure is a reasonable way to represent complete indifference, then it should work equally well when it is applied to the inverse times $1/T$. In that application, by parallel reasoning, we arrive at the result that, with probability one, $P(H(1/1,\infty)\,|B) = 1$. But $H(1/1,\infty) = H(1,\infty) = H(0,1)$,[232] so that we have a contradiction with the earlier probability assignment $P(H(0,1)\,|B) = 0$.

The escape from the contradiction is to decide that only one of the two finitely additive measures may be used. That, however, amounts to selecting a privileged subset of probability one times of excitation: the times between 0 and 1, or between 1 and infinity. The physics makes no such distinction. It is an addition forced on us by the probabilistic measure.

Two further probabilistic embellishments have been treated elsewhere in Norton (2010a) and in earlier chapters. First, one might try to escape the need to select a single time constant τ in (10) by adopting the complete set of measures (10), for all values of τ, as the representation of the strength of support. The motivation is correct in that it seeks a representation weaker than a single probability measure. However it is too indirect in that it seeks to preserve probability measures by using them to simulate a different, non-additive logic. The better approach is simply to write down that logic directly, as is done in Section 7 Below.

Second, one might adopt the measure of (10) as a subjective degree of belief. The earnest but possibly unrealizable hope is that repeated conditionalization will wash away the subjective opinion and leave behind the objective bearing of evidence, or at least some approach to it. Once again, the motivation is good, but the execution poor. Again, the better approach is merely to write down the warranted logic directly.

6.2 Probabilities, Empirically?

While we may not be able to recover probabilities from the physics governing these indeterministic systems, might we introduce them through an empirical artifice? To take a concrete case, imagine that somehow we are able to physically realize a dome. We might then set up very many of them and just observe what happens. Might we find that that the frequencies for different times of spontaneous excitation stabilize towards limiting values? We could then

[232] Aside from the inclusion of T=1 in $H(1/T,\infty)$, but not in $H(0,1)$.

introduce probabilities, set in value to those empirically determined, limiting, relative frequencies.

Dawid (2015) considers an even simpler case in the same spirit. What if we have 100 domes and find that they all excite spontaneously at exactly 16.8 seconds? Might we then infer to a deterministic rule: spontaneous excitation occurs at 16.8 seconds for all domes?

How we treat these proposals will depend on how certain we are of the background, governing physics. Are we certain of the background physics or are we not?

In the first case, we remain certain that the Newtonian physics specified is the totality of the physics governing the processes. That all excitations occur at 16.8 seconds is compatible with the indeterministic physics, but it is not something we could predict from that physics, at the exclusion of many other possibilities. Correspondingly, the background physics authorizes no further predictions, even after we have seen all 100 domes excite at 16.8 seconds. We should remain as uncertain of the next excitation time as we were prior to seeing the first dome in the imagined experiment.

This situation is quite similar to that of a gambler in a casino at a roulette wheel. Neglecting 0 and 00, the chance of a black on a properly functioning wheel is 1/2. Imagine, however, that the gambler steps up to the table with the wheel and finds 20 successive spins to yield black. Assume the gambler is confident of the background theory: the wheel is functioning properly. All the gambler can properly conclude is that an extremely unlikely event has occurred. Twenty successive black outcomes is possible, just improbable.

What the gambler should not now think is that the wheel is on some sort of "streak" so that, contrary to the physical construction of the wheel and the laws of probability, the next outcome is more likely to be black. To think that is to commit a notorious gambler's "streak" fallacy.

It is the same with the dome. As long as we remain convinced that the Newtonian physics described is the totality of the physics that governs the dome, repeated excitations at 16.8 seconds is merely a coincidence. In a similar vein, the indeterministic physics does not support the existence of stable limiting frequencies for different excitation times. Any appearance of such stability is mere coincidence that cannot be expected to persist.

That was the first case. In the second case, we become uncertain that the Newtonian physics described is all that governs the actual domes of our experiment. We suspect that some further or some other physics is at play. What physics it might be is hard to say, since the entire scenario is built from multiple layers of fiction. I leave it to the reader's imagination. Whatever alternative physics we may suspect here is what will guide the inferences.

Once again, the situation is similar to that of the gambler. The probability of 20 black outcomes is exceedingly small: $1/2^{20}$, which is roughly 1/1,000,000. Having seen such an

improbable occurrence, the gambler would reasonably suspect that something odd is afoot. Perhaps the wheel has some ingenious cheating device that is malfunctioning and delivering all black outcomes. If the gambler believes that to be the case and that the cheating device will continue to operate well, the gambler would be well warranted to conclude that the next outcome will be black.

In short, as long as we retain the presumptions made at the outset of the totality of the physics governing the indeterministic systems, any empirically observed regularities of the type suggested will be of no help to us inductively. To expect otherwise is to commit a fallacy analogous to the gambler's "streak" fallacy.

6.3 Systems with Indetermimism Among their Components

The inductive problems posed in Section 4 are to find the inductive strengths of support afforded to underdetermined components of a physical system by those that are fixed by the problem specification. Much of the analysis of Section 6.1 can be carried over to the probabilistic analysis of these problems. Probabilistic analysis fails in the same way. In addition, the infinite dimensionality of the space of underdetermined potentials (9) in Newtonian cosmology raises more problems.

The simplest problem was posed in Section 4.1. We are to choose among the infinitely many gauge equivalent fields of (4). This choice amounted to selection of a value of the constant K, which can take any real value, positive or negative, large or small.

The straightforward approach is to represent strength of inductive support by a probability distribution over K. However, since K has an infinite range, the distribution must be attenuated towards zero for large positive and large negative values of K. Otherwise it will not normalize to unity. Here the difficulty is like that faced by the probabilistic law (10). The rate of attenuation will be represented by some parameter or some characteristic of the distribution that is akin to the selection of the time constant τ in (10). Any choice of a rate of attenuation, however, is an addition to the physics of the gauge system.

One might also try to avoid the problem by employing an unnormalizable probability distribution akin to (12). Once again, this will add to the physics, for it requires us to assign higher probability to larger intervals of K, even through the physics does not authorize it. Finally the difficulties of the finitely additive measure can be replicated here as well.

The still harder case for probabilistic analysis is that of Newtonian cosmology in Section 4.2. For now we are to distribute probabilities uniformly over the space of potentials (9). Its individual solutions are picked out by specifying values for the infinitely many parameters a_1, a_2, a_3, … That is, it is an infinite dimensional space. The familiar problem is that we cannot easily

assign an additive measure over such spaces since the parameter values range from minus infinity to plus infinity. In the examples so far, it is the requirement of normalization of the measure of the full space to unity that forces the problem. The new problem with an infinite dimensional space is there is still no well behaved, uniform measure over this space, even if we drop the requirement of normalization.

To see this, recall that probabilities behave like volumes in space. So, for continuity with familiar notions, let us continue to call them volumes. First consider a space of parameters a_1, a_2, …, a_n of finite dimension n. The set of all points for which $0 < a_i < 2$, all i, forms a cube of side 2. This cube consists of 2^n cubes of unit side. In a three dimensional space, the side 2 cube consists of $2^3 = 8$ unit sided cubes. If we assign unit volume to each unit cube, the side 2 cube just has volume 2^n.

For any finite n this relation is unproblematic. That ceases to be so when we take the case of the infinite dimensional space. For then, the sided 2 cube consists of an uncountable infinity 2^∞ of unit cubes. Since the measure is uniform, all the unit cubes have the same volume. There are two cases: the unit cubes have non-zero volume; and the unit cubes have zero volume.

If the unit cubes have some finite, non-zero volume, then it follows that the side 2 cube must have infinite volume. This follows using only finite additivity of the volumes. For if we suppose any finite volume for the side 2 cube, then we need only sum finitely many of the unit cubes to recover a summed volume greater than it. Of course, if the unit cubes have infinite volume, then so must also the side 2 cube.

The other possibility is that the unit cubes have zero volume. Then the side 2 cube can also have zero volume. However it may also have a finite, non-zero volume or an infinite volume. This may seem odd, since we are supposing the side 2 cube to consist of nothing but zero volume unit cubes. Why not add up all these zeroes and get zero volume? The problem is that there are an *uncountable* infinity 2^∞ of zeroes and adding uncountable infinity of them is an undefined operation.[233] The volume of the side 2 cube must merely be greater than the sum of the volumes of finitely many unit cubes; or (if countable additivity is assumed) of a countable infinity of them. So its volume can be non-zero.[234]

---

[233] This is a familiar result. Each point in the unit intervals of reals is of zero length. Since there there are an uncountable infinity of them, we cannot add them to find the length of the unit interal of reals, which is not zero, but one.

[234] This is an uncommon possibility. In discussions of measures on infinite dimensioned spaces, it is usually assumed that the spaces are separable, which allows that each region can be composed of a countable infinity of equal volume subregions. Separability fails in this case.

These results can be applied to a cube anywhere in the space. Every cube can be decomposed into 2^{∞} half-sided cubes; and every cube is itself a component cube of a doubled-sided cube. What results are three possibilities for the uniform measure. The two simple ones are just that all cubes have either zero volume or infinite volume. The complicated case is that there is some value L such that an L sided cube has finite, non-zero volume. Since the measure is uniform, all cubes of side L will have this volume. It follows by replicating the above reasoning that all smaller cubes that can be compounded to form cube of side L must have zero volume; and all larger cubes that can be built from cubes of side L must have infinite volume.

This third option violates the requirement that we add nothing to the physics, for it singles about quite particular, preferred sets of parameters as just those that reside in the cubes of side L. Since parameter values correspond to gravitational potentials, this is a privileging of certain sets of potentials.

Combining the three possibilities, cubes in this space will almost everywhere have either zero volume or infinite volume. One can see this result informally by noting what happens when we scale up or scale down any region by a factor M. That is, we multiply all the parameter values in the set specifying the region by M. The volume of the region will scale by a factor $M^{\text{dimension of space}} = M^{\infty}$. This factor is zero if $M<1$ and infinity if $M>1$. This suggests that almost all volumes will be zero or infinity. For a finite, non-zero volume cannot stay finite and non-zero under any scaling, either up or down. It becomes an infinite or a zero volume respectively. However employing this factor M^{∞} directly in a more thorough argument is not straightforward since it leads to indeterminate arithmetic forms. For example, scale up a zero volume by an infinite factor M^{∞}, when $M>1$. The new volume is "0 x ∞," which is an expression that cannot be evaluated.

Note that these troubles arise *without* assuming that the volume of the total space normalizes to unity. If we retain countable additivity, the possibilities above admit only two values for the volume of the entire space: zero or infinity.

It might be tempting to drop countable additivity, assign zero volume to any bounded region and unit volume to the whole space. One does not escape the difficulty already developed above for finitely additive measures in the case of spontaneous excitations. Briefly, the measure ought to be indifferent to whether we parameterize the space with the original parameters a_i or their inverses, $1/a_i$. Then we would assign zero volume to the side 2 cube in the inverse parameterization $1/a_i$ for which $0<|1/a_i| < 1$, all i. But this region corresponds to the entirety of the space in the original parameterization, $1 < |a_i| < \infty$, excepting a zero volume cube $0< |a_i| < 1$. In the original parameterization, this region is assigned unit volume.

7. The Inductive Logic Warranted

7.1 The Logic

The material theory of induction directs us to look to the background facts to determine which logic is warranted. In the cases of this chapter, the background facts are, by careful contrivance, such as to support essential no non-trivial inductive inferences at all. They allow us merely to say that certain outcomes are possible but to provide no discriminations of the nature of "more possible" or "less possible." This lack of discrimination can be codified into a formal calculus with three values:[235]

nec = necessary
poss = possible
imp = impossible

These values are assigned to strengths of inductive support, written as "[$A|B$]," where this symbol represents the strength of inductive support afforded to proposition A by proposition B. The little structure these strengths have is induced by deductive relations among the propositions; or, in other terms, by set theoretic containment amongst the sets of possibilities. That is, we have:

$$\begin{aligned}[A|C] &= nec, \text{ if } C \text{ deductively entails } A.\\ &= imp, \text{ if } C \text{ deductively entails not-}A.\\ &= poss, \text{ otherwise.}\end{aligned} \tag{13}$$

The logic is empty until we specify the propositions to which it applies. Many choices are possible here. One convenient choice arises in the context of the spontaneously exciting systems of Section 4.1. The propositions over which this logic is defined are: $H(T_1, T_2)$, as defined in Section 6.1; B: the proposition that describes the background physical facts of the system; and, for completeness, $H(\infty)$: the time of spontaneous excitation $T = \infty$. Proposition $H(\infty)$ corresponds to the case in which there is no spontaneous excitation.

The logic now authorizes us to assign strengths of support such as

$[H(T_1, T_2) \mid B] = poss$, for any $T_2 > T_1$.
$[H(\infty) \mid B] = poss$
$[H(1, 2) \mid H(0, 4)] = poss$
$[H(0, 4) \mid H(1, 2)] = nec$
$[H(0, 4) \mid H(10, 20)] = imp$

[235] This logic has been developed in various forms in Norton (2008a, 2010a and 2010b) and in Chapter 10.

There is a natural and obvious generalization to the systems of Section 4 with indeterminism among the system components.

An important property of this logic is that it is not additive, in contrast with the probability calculus. That is, if A_1 and A_2 are mutually exclusive propositions, such that $[A_1|C] = [A_2|C] = poss$, then it is possible that $[A_1 \vee A_2 \mid C] = poss$. Overall, we violate additivity since

$$[A_1|C] = [A_2|C] = [A_1 \vee A_2 \mid C] \tag{14}$$

The additivity of a probability measure would require in this case that

$$P(A_1|C) + P(A_2|C) = P(A_1 \vee A_2 \mid C)$$

so the probabilities assigned to A_1, A_2 and $A_1 \vee A_2$ cannot be equal unless we have the exceptional case of all probability zero outcomes.

7.2 Invariances

Norton (2008a, 2010b) and Chapter 10 argued that this logic (13) represents the case of completely neutral support; that is, the case in which we have no reason at all to favor any of the contingent propositions in any degree. It was shown that the logic can be derived in two ways from two invariance properties. We shall see below that these invariances are respected to a great extent in these systems. However, do recall that the logic (13) of Section 7.1 was not derived from these invariances, but directly from the possibilities allowed by the background physical facts.

Redescription

The first invariance is invariance under redescription. This invariance is commonly employed in the context of the principle of indifference. It arises when we redescribe a system in a way that preserves our indifferences.

Take, for example, the value of the parameter K in the Newtonian gauge system of Section 4.1. Represent a useful set of hypotheses by:

$$H_K(k_1, k_2)\text{: the parameter } K \text{ lies in the interval } k_1 \leq K < k_2$$

On the basis of the background facts B, we are indifferent to K lying in equal ranges of values, so we have

$$poss = [H_K(0, 1) \mid B] = [H_K(1, 2) \mid B] = [H_K(2, 3) \mid B] = [H_K(3, 4) \mid B]$$
$$= [H_K(4, 5) \mid B] = [H_K(5, 6) \mid B] = [H_K(6, 7) \mid B] = [H_K(7, 8) \mid B]$$

Now replace the parameter K by $L = K^3$. Since L is an equally good parameter to use in (4), we can also write

$$poss = [H_L(0, 1) \mid B] = [H_L(1, 2) \mid B]$$

However $H_L(1, 2) = H_K(1, 2^3) = H_K(1, 8)$

$= H_K(1, 2)$ v $H_K(2, 3)$ v $H_K(3, 4)$ v $H_K(4, 5)$ v $H_K(5, 6)$ v $H_K(6, 7)$ v $H_K(7, 8)$

Combining with $H_L(0, 1) = H_K(0, 1)$ we recover

$$poss = [H_K(1, 2) \mid B] = [H_K(2, 3) \mid B] = [H_K(3, 4) \mid B]$$

$$= [H_K(4, 5) \mid B] = [H_K(5, 6) \mid B] = [H_K(6, 7) \mid B] = [H_K(7, 8) \mid B]$$

$$= [H_K(1, 2) \text{ v } H_K(2, 3) \text{ v } H_K(3, 4) \text{ v } H_K(4, 5) \text{ v } H_K(5, 6) \text{ v } H_K(6, 7) \text{ v } H_K(7, 8) \mid B]$$

This is an example of the failure of additivity of the type of (14).

Negation

The second invariance is invariance under negation. If the support for some proposition A is completely neutral, then we have no grounds to assign it more or less support than its negation not-A. We must assign the two equal support. That is, the strength of support remains unchanged under the negation map that sends hypotheses to their negations.

This negation map can be implemented in the case of systems that can spontaneously excite as follows. Write

$H_T(T_1, T_2)$: the time of spontaneous excitation T lies in the interval $T_1 \leq T < T_2$

Hypothesis $H_T(0, 1)$ says that this time lies in $0 \leq T < 1$. Its negation, not-$H_T(0, 1)$, asserts that that the time of spontaneous excitation lies in $1 < T \leq \infty$. Negation invariance of the strengths of support requires the equality

$$[\text{not-}H_T(0, 1) \mid B] = [H_T(0, 1) \mid B] \tag{15}$$

We can see that this equality obtains according to the rules of (13). For

$$\text{not-}H_T(0, 1) = H_T(1, \infty) \text{ v } H(\infty)$$

and from the rules

$$[\text{not-}H_T(0, 1) \mid B] = [\, H_T(1, \infty) \text{ v } H(\infty) \mid B \,] = poss$$

as well as

$$[H_T(0, 1) \mid B] = poss$$

All these hypotheses accrue equal support *poss* from the background B since none are deductively entailed by B.

We can also derive negation invariance from redescription invariance. Consider the support, not for various times T, but for the inverse times $1/T$. If we are indifferent to the two parameterizations of the time, T and $1/T$, then we would have, under description invariance:

$$[H_T(0, 1) \mid B] = [H_{1/T}(0, 1) \mid B]$$

The interval $1 < T \leq \infty$ is the same $0 \leq 1/T < 1$. That is,

$$\text{not-}H_T(0, 1) = H_{1/T}(0, 1)$$

Combining we infer

$$[H_T(0, 1) \mid B] = [H_{1/T}(0, 1) \mid B] = [\text{not-}H_T(0, 1) \mid B]$$

This is just negation invariance (15).

8. Conclusion

According to the material theory of induction, there is no logic or calculus of inductive inference that applies universally to all problems in inductive inference. It follows that there are problems in inductive inference in which strengths of support cannot properly be represented by probability measures. This chapter illustrates this claim with examples of indeterministic physical systems contrived to be resistant to a representation of strengths of inductive support as probabilities. The contrivance depends on finding simple physical systems in which a full specification of the background physical facts can be given and their burden easily discerned. An inductive analysis must determine strengths of inductive support without requiring alteration of or addition to these background facts. In the examples presented, using probabilities to represent strengths of supports requires just such additions. For this reason their use fails.

The material theory of induction asserts that the applicable logic of induction is determined by these background facts. Their paucity supports a very weak, three-valued inductive logic that happens to coincide with the completely neutral strengths of inductive support elaborated elsewhere.

The inductive problems of this chapter all involve problems of indeterminism in which certain aspects of a system fail to fix certain other aspects. Problems of this sort do arise in recent science. The most obvious involves singularities in general relativity. Singular spacetimes can develop in many ways into the future. The possibilities are not determined and there are no probabilities provided by general relativity to weight the different possibilities.

A white hole is the temporal inverse of a black hole. When systems fall into a black hole, their structures are obliterated by the black hole, whose properties are merely mass, charge and angular momentum. If we now take the time reverse of the falling in, anything that can fall into a black hole can also be ejected by a white hole. The possibilities are not determined.

In relativistic cosmology, the big bang is a spacetime singularity in our common past, out of which the entire universe issued. The long-standing puzzle has been to explain why this singularity issued in a universe that is so nearly spatially homogeneous and isotropic and with spatial curvature very close to zero. Here is a problem in inductive inference. Given the background facts of general relativity and that there is an initial singularity, what support do we have for the various possible cosmologies that may arise? There are very many possible configurations other than the particular one manifested in our universe; and there are no good

reasons provided in pre-inflationary cosmology[236] that we should have just these initial conditions and not others.

It is tempting to convert these last facts into the claim that it is very improbable that we have the initial conditions we do. But such a claim, if read literally, solves the inductive problem by means of a probability measure. Since the background facts listed provide for no probabilities, their introduction is as illicit as in the contrived examples of this chapter.

The moral of the chapter is that we should be prepared for problems in inductive inference in which strengths of support are not well-represented by probability measures. To do otherwise, to persist in representing strengths of inductive support universally as probability measures, risks unwittingly importing new facts that change the problem posed to a new one amenable to probabilistic representation. The outcome is that we will not have solved the problem actually before us but a different one that we wished we had.

Appendix A: Toppling Dominoes

A domino has width W, height H and mass m and is separated from the next domino by an inter-domino distance L. To be toppled, a small impulse is needed to push the domino from its vertical position until it strikes the next domino, as shown in Figure 7.

[236] The once common claim that inflationary cosmology does provide these reasons is now challenged. See for example Holland and Wald (2008).

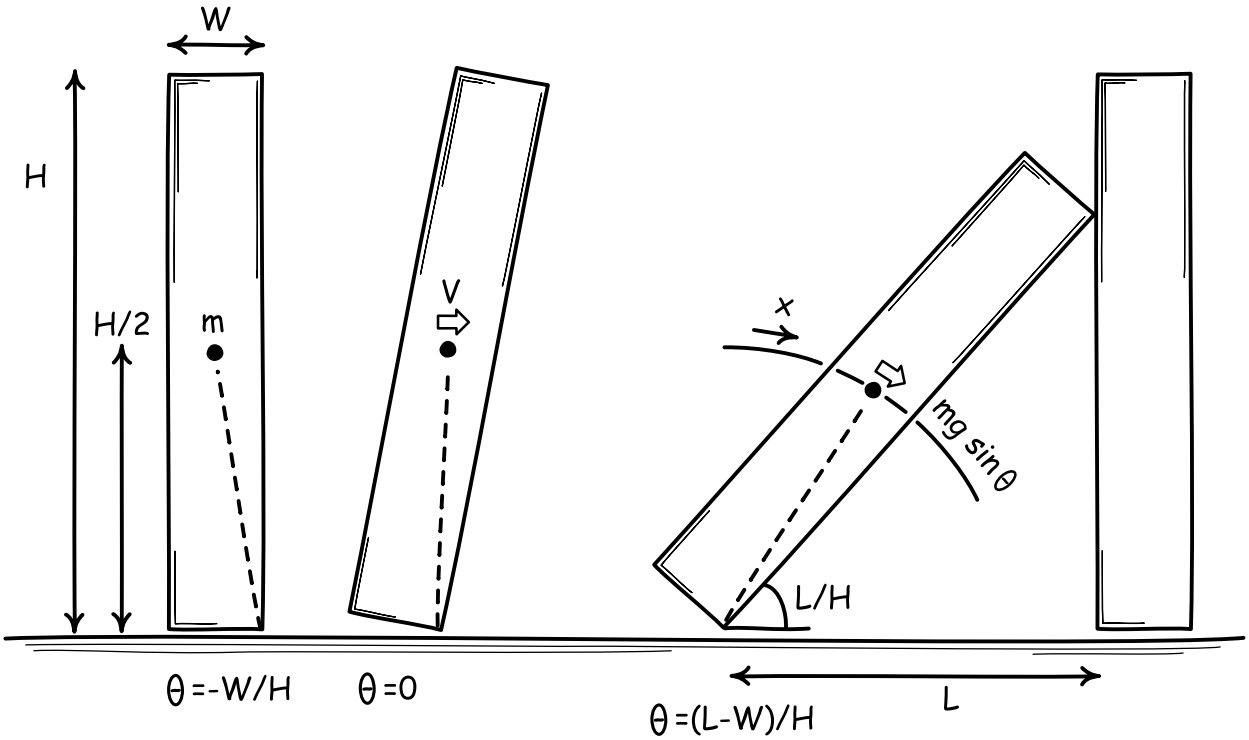


Figure 7. Geometry of a Toppling Domino

As the center of mass of the domino pivots on one edge, if forms an inverted pendulum. Call the angular position of the center of mass θ as it pivots around the edge and set θ=0 when the center of mass is directly over the edge. If the distance along the circular arc traced by the center of mass of the domino is x and the center of mass is located in the geometric center of the domino, then the gravitational force on the center of motion in the direction of the arc is $mg\sin\theta$, for g the acceleration due to gravity. The equation of motion in time t is $m\frac{d^2x}{dt^2} = mg\sin\theta \approx mg\theta$, where sin θ is approximated as θ for the small angles we encounter here. Since $\theta = x/(H/2)$, we have

$$\frac{d^2\theta}{dt^2} = k^2\theta$$

where $k^2 = g/(H/2)$. This inverted pendulum equation of motion admits the general solution $\theta(t) = A\sinh(kt) + B\cosh(kt)$, for undetermined constants A and B. We set $\theta(t) = 0$ when t=0, so that B=0, and arrive at:

$$\theta(t) = A\sinh(kt)$$

In toppling, the center of mass of the domino is first lifted by the rotational pivot about the edge and then falls under gravity once past the edge.

It would be convenient if there were some longest time this motion could take. One might imagine that, if the domino were given just the right, minmal push, it would pivot slowly and its center of mass would momentarily have zero speed as it passes over the edge at the apex of its motion. This cannot happen. A longer computation shows that this motion would require infinite time. (For more, see Norton, 2003, pp. 11-12.)

The best we can secure is that the center of mass, at the moment of passing over the edge, has some small linear speed V. Since the angular speed is $d\theta(t)/dt = Ak \cosh(kt)$, we require $V/(H/2) = d\theta(0)/dt = Ak \cosh(k0) = Ak$. Thus the solution is

$$\theta(t) = V/(kH/2) \sinh(kt) \approx Vt/(H/2)$$

since, for small times, $\sinh(kt) \approx kt$.

The domino center of motion must move from its initial angular position $\theta = -W/H$ to its collision with the next domino at angular position $\theta = (L-W)/H$. Substituting into the last equation for $\theta(t)$, we have $L/H = Vt/(H/2)$ for the time t required by the domino to fall. That is

$$t = L/2V$$

Thus the time t_n for the nth domino to fall is given by $L_n/2V$, where L_n is the distance between dominoes n and $(n-1)$. Thus:

Total time for cascade

$$= \sum_n t_n = (1/2V) \sum_n L_n = (1/2V) \text{ Total distance between dominoes}$$

If we assume that the domino width scales in the same way as the distance between the dominoes, the condition that the cascade completes in finite time reduces to the condition that the domino row be of finite spatial length. (Informally, this condition follows if we imagine that that the falling propagates through the chain at roughly a constant speed V.)

An assumption of this analysis is that each domino has the same speed V as its center of mass passes its apex. One might wonder whether the system can provide each domino sufficient energy. Some qualitative considerations show that this will not be a problem. Each domino by supposition has speed V at its apex and thus kinetic energy $(1/2)mV^2$. Assuming elastic collisions, it will pass this much energy to the next domino as well as the extra energy released when the domino center of mass falls to a lower height overall.

Indeed the problem will not be a lack energy to sustain the cascade, but the danger of a surfeit. For there are infinitely many dominoes of the same mass, each falling through a height in a finite time. If each domino falls to the same prone position, that will result in release of an infinite amount of energy.

Appendix B: Newtonian Cosmology

The force (15) exerted by an infinite, flat plate of density ρ and thickness Δx is independent of the distance to the plate is easy to see qualitatively. Consider the portion of the plate subtended by a very small angle Ω at the location of unit test mass. The volume and thus the mass of this portion is proportional to Ωr^2. However the force exerted by this mass on the test mass diminishes with $1/r^2$. Hence the force is proportional just to Ω and independent of distance.

The full expression for the force is computed as follows. The distance r from the unit test mass to each part of the plate satisfies $r^2 = x^2 + s^2$ where x the shortest distance to the plate and s the distance from the closest point on the plate to the part at issue. A circular ring of width ds at radius s in the plate exerts a force on the unit test mass of

$$\frac{G\rho 2\pi s ds \Delta x}{r^2}\cdot\frac{x}{r} = G\rho 2\pi x \Delta x \frac{sds}{\left(s^2+x^2\right)^{3/2}},$$

where x/r is the cosine of the half angle at the base of the cone subtended by the ring. Integrating over all s, we recover (15) as

$$f = 2\pi G\rho x\Delta x\int_{s=0}^{s=\infty}\frac{sds}{\left(s^2+x^2\right)^{3/2}} = 2\pi G\rho x\Delta x\frac{-1}{\left(s^2+x^2\right)^{1/2}}\Bigg|_{s=0}^{s=\infty} = 2\pi G\rho x\Delta x\frac{1}{x} = 2\pi G\rho\Delta x$$

We can also compute the Newtonian gravitational potential field φ directly from Poisson's equation

$$\nabla^2\varphi = \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\varphi = 4\pi G\rho \tag{16}$$

For constant ρ, the solutions (7a, b, c) and (8a, b, c) follow immediately. For example, we recover (7a) as

$$\nabla^2\varphi_x = \nabla^2\left(2\pi G\rho x^2\right) = \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\left(2\pi G\rho x^2\right) = 2\pi G\rho\frac{\partial^2}{\partial x^2}x^2 = 4\pi G\rho$$

That $\Phi = 2\pi G\rho\,(y^2 - x^2)$ is harmonic follows since

$$\nabla^2(y^2-x^2) = \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)y^2 - \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)x^2 = \frac{\partial^2}{\partial y^2}y^2 - \frac{\partial^2}{\partial x^2}x^2 = 2-2 = 0$$

That adding a harmonic function to a solution of Poisson's equation (16) takes us to another solution follows from the linearity of the operator ∇^2. If Φ is a harmonic function, which

satisfies Laplace's equation $\nabla^2\Phi = 0$, and we add it to an existing solution φ of Poisson's equation (16), their sum $(\varphi + \Phi)$ also satisfies Poisson's equation, for

$$\nabla^2(\varphi + \Phi) = \nabla^2\varphi + \nabla^2\Phi = \nabla^2\varphi + 0 = 4\pi G\rho$$

The full set of harmonic functions is a linearly independent set. There is no simple way to write this set. In spherical coordinates (r, φ, θ), the harmonic functions are

$$\Phi(r, \varphi, \theta) = (A_j r^j + B_j / r^{j+1}) P_j^m(\cos\theta)(a_m \cos m\varphi + b_m \sin m\varphi),$$

for A_j, B_j, a_m, b_m arbitrary constants; $m = -j, -(j\text{-}1), \ldots, (j\text{-}1), j$; and $j = 0, 1, 2, 3, \ldots$;

and $P_j^m(\cos\theta)$ are the associated Legendre functions of $\cos\theta$. (From Bronshtein and Semendyayev, 1985, p. 463, after correction of apparent typographical errors.)

Digression for Experts

Since this problem of Newtonian cosmology has attracted considerable attention in the philosophy of physics literature, I include a short digression for experts.

Among the solutions to (16) is one that is formed as the equally weighted sum of the three solutions $(1/3)\varphi_x + (1/3)\varphi_y + (1/3)\varphi_z$ and is called by Malament (1995) a canonical solution centered at the origin

$$\varphi_{can} = (2/3)\pi G\rho\,(x^2 + y^2 + z^2) = (2/3)\pi G\rho\, r^2 \qquad (17a)$$

where the radial coordinate r satisfies $r^2 = x^2 + y^2 + z^2$. This solution has a special status as a solution with maximum isotropy: it is isotropic about the origin $\boldsymbol{r} = (x, y, z) = \boldsymbol{0}$. That falls well short of the full homogeneity and isotropy that the early physicists expected. It has a preferred center at the origin of coordinates. Infinitely many more, distinct canonical solutions are possible, each centered at different points in space, $\boldsymbol{r}_0 = (x_0, y_0, z_0) \neq \boldsymbol{0}$.

$$\varphi_{can} = (2/3)\pi G\rho\,(r\text{-}r_0)^2 \qquad (17b)$$

Malament showed, however, that the differences among these canonical solutions were only apparent. He adopted the natural assumption that the physically real properties of a Newtonian cosmology manifest in the relative accelerations of point masses in free fall. It turned out that all the canonical solutions give the same relative accelerations. That is, the choice among them was merely the exercising of a gauge freedom. For further motivation for this choice of what is physically significant, see Norton (1995).

Malament's analysis gave a satisfactory answer to this question: which isotropic, homogeneous Newtonian cosmologies are there? The answer is given uniquely by the canonical solutions.

Our present question is a different one. It is: which potential fields are fixed by a uniform matter distribution through Poisson's equation (16). The answer to this question, as has been emphasized by Wallace (2016), is that there are infinitely many such fields and they form the infinite set (9). Only very few of them prove to be physically equivalent after the manner of (17a) and (17b). Solutions (7a), (7b) and (7c) are *not* physically equivalent. It follows from (6) that masses in free fall in (7a) φ_x experience relative accelerations in the *x*-direction but not in the *y*- or *z*-directions. Similarly masses in free fall in φ_y and φ_z experience relative accelerations respectively in the *y*- and *z*-directions only.

A natural way to block this failure of the mass distribution to determine the gravitational potential, as Wallace (2016) has emphasized, is to impose boundary conditions. All but the canonical solutions are eliminated if we require isotropy in the physically significant properties, as do Malament (1995, p. 492, p. 501) and Norton (1995, p. 513, footnote 2). However the imposition of this condition must be understood as a distinct choice we make in order to prune the space of solutions to a subset that happens to interest us. We cannot derive it from the isotropy of space and the matter distribution, for the Poisson equation does not respect this symmetry in its individual solutions.

References

Aharonov, Yakir and Bohm, David (1959) "Significance of Electromagnetic Potentials in the Quantum Theory," *Physical Review*, **115**, pp. 485–491.

Dawid, Richard (2015) "Turning Norton's Dome Against Material Induction," *Foundations of Physics* **45**, pp. 1101-1109.

Earman, John (1986) *A Primer On Determinism*. Dordrecht, Holland: D. Reidel Publishing Company.

Hollands, Stefan and Wald, Robert M. (2008) "An Alternative to Inflation," https://arxiv.org/pdf/gr-qc/0205058.pdf

Malament, David (1995) "Is Newtonian Cosmology Really Inconsistent?" *Philosophy of Science*, **62**, pp. 489-510.

Manchak, John and Roberts, Bryan W., (2016) "Supertasks", *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*, Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/win2016/entries/spacetime-supertasks/>.

Norton, John D., (1995) "The Force of Newtonian Cosmology: Acceleration is Relative" *Philosophy of Science*, **62**, pp. 511-22.

Norton, John D., (1999), "A Quantum Mechanical Supertask", *Foundations of Physics,* **29**, pp. 1265–1302.

Norton, John D., (1999a) "The Cosmological Woes of Newtonian Gravitation Theory," in H. Goenner, J. Renn, J. Ritter and T. Sauer, eds., *The Expanding Worlds of General Relativity: Einstein Studies*, Volume 7, Boston: Birkhäuser, pp. 271-322.

Norton, John D. (2003)"Causation as Folk Science," *Philosophers' Imprint* Vol. 3, No. 4 http://www.philosophersimprint.org/003004/

Norton, John D. (2008)"The Dome: An Unexpectedly Simple Failure of Determinism," *Philosophy of Science*, **75**, pp. 786-98.

Norton, John D. (2008a) "Ignorance and Indifference," *Philosophy of Science,* **75**, pp. 45-68.

Norton, John D. (2010)"Deductively Definable Logics of Induction." *Journal of Philosophical Logic*, 39 (2010), pp. 617-654.

Norton, John D. (2010a) "There are no Universal Rules for Induction," *Philosophy of Science*, **77**, pp. 765-777.

Norton, John D. (2010b) "Cosmic Confusions: Not Supporting versus Supporting Not," *Philosophy of Science*, **77**, pp. 501-523.

Norton, John D. (2011) "Challenges to Bayesian Confirmation Theory," *Philosophy of Statistics, Vol. 7: Handbook of the Philosophy of Science.* Prasanta S. Bandyopadhyay and Malcolm R. Forster (eds.) Elsevier.

Norton, John D. (2012) "Approximation and Idealization: Why the Difference Matters" *Philosophy of Science*, **79**, pp. 207-232.

Bronshstein, I. N. and Semendyayev, K. A. (1985) *Handbook of Mathematics*. New York: Van Nostrand Reinhold.

Wallace, David (2016) "More Problems for Newtonian Cosmology," http://philsci-archive.pitt.edu/12036/1/newton-cosmology-phil.pdf

Field initializing equation numbering variables here:

November 7, 2013; December 4, 2016; June 25, 2018

Chapter 16

A Quantum Inductive Logic

1. Introduction[237]

The material theory of induction requires that good inductive inferences must be warranted by facts within their domain of application. In earlier chapters, we have seen many examples of individual inductive inferences warranted by specific facts. Marie Curie, for example, inferred the crystallographic system of all crystals of radium chloride from inspection of just a few specks of the substance. The inference was warranted by facts contained in crystallographic principles from the preceding century, not by some universal inductive inference schema.

In cases like this, there is little sense that the inductive inference forms part of a larger inductive logic whose overall structure could be abstracted in some measure from the specific subject matter. There are cases, however, in which this abstraction is possible. Complete abstraction is impossible; that would provide a universal logic of induction. We can find cases in which sufficient structure can be abstracted for a rich logic to appear.

The most familiar inductive logic of this type and the one that is best worked out is that of probabilistic logic. The prevalence of this probabilistic logic has given the illusion that it is the universal inductive logic and that no other inductive logic is viable. That illusion persists only because of the familiarity of the example and the lack of looking for alternatives. If we have any domain governed by some well-developed theory, then a compactly expressible inductive logic may be supported. Just which that logic will be, depends on the character of the theory. There will be cases in which the logic supported is not probabilistic.

In the preceding chapters, such cases were illustrated with simple examples: an infinite lottery machine, various forms of indeterministic systems and the non-measurable outcomes arising among infinitely many coin tosses. While we can and should demand that an inductive logic applies to these cases, one can be forgiven for finding the examples contrived or abstruse.

---

[237] I thank Rob Spekkens for helpful discussion.

They were so precisely because that enabled the systems to be simple enough for us to comprehend their physical properties fully.

Might we find an example with an immediate application to present science? This chapter presents such an example. Drawing on the work of Leifer and Spekkens (2013), we shall see that the natural mathematical structures of quantum theory afford a distinctive, non-probabilistic logic, at least for certain quantum systems, such as systems of entangled particles.

This quantum inductive logic differs from a probabilistic inductive logic in its most fundamental quantity. A probabilistic logic uses an additive probability measure to represent degrees of support or, in subjective terms, belief states. In its place, the quantum logic uses a structure that arises naturally in quantum theory, a density operator. That this is the appropriate structure derives in turn from a deeper difference. Probabilities arise naturally when all of the distinct states of a system fall under a single probability measure supplied by background facts. While there are probabilities associated with measurement outcomes in quantum theory, each measurement setting is associated with a different probability measure and, crucially, their totality does not form a single probability measure. Rather the different probability measures are both issued by and unified by a single, deeper structure, a density operator. That structure is the fundamental quantity of the quantum inductive logic.

It may seem strange at first to replace a probability measure by a density operator when probabilities can also be found in quantum theory, even if in scattered form. For, one might think, a probability measure, when it can be had, is just the right thing to use to represent partial inductive support or uncertain beliefs. This thought is driven more by familiarity and comfort than good reasons. The naturalness of a probability measure is an artifact of hundreds of years of development. It is a rather abstruse notion, as one finds when one engages in the cumbersome task of explicating precisely what it means to say that, for example, some outcome has such and such a probability. We shall see below that a density operator is no more abstruse and, since it is the central structure provided by the quantum mechanics, it functions much better as the basis of the quantum inductive logic.

Section 2 below will sketch some probabilistic inferences on the presence of a rare genetic mutation among siblings. It will serve as a foil for the quantum case introduced in Sections 3 and 4, inductive inferences over the measured spins of entangled electrons. Section 5 to 9 develop the mathematical devices needed to treat the spins of entangled electrons. Sections 10 and 11 will identify one of these devices, a density operator, as the appropriate analog in the quantum case of the probabilities of the foil. Section 12 will provide a simple geometric picture of the density operators to support this identification. Section 13 will briefly review how Leifer and Spekkens (2013) have developed the approach sketched into a fuller calculus with some

analogies to the probability calculus. Sections 14 and 15 explore analogies and disanalogies between the probabilistic and quantum inductive logics. Section 16 offers conclusions.

2. Probabilistic Inductive Inference

2.1 Rare Genetic Mutations

As a foil for the quantum case, let us consider cases in which a probabilistic logic is warranted by the prevailing facts. One arises when we have outcomes generated by physical chances. The simplest of these is a gambling casino. By careful design, a roulette wheel (with a 0 and 00) has a physical chance of 18/38 of a red outcome; and a physical chance of 18/38 of a black outcome. That fact, and others like it, warrant using the corresponding probabilities as the measure of inductive support for red and black; and employing the probability calculus as the logic of induction applicable to casino games.

Population frequencies can also provide a factual warrant for the use of probabilities in an inductive logic. Demographic data consistently shows that low educational level correlates with unemployment. People in the US without a high school diploma, for example, are the group with the most unemployment. We make the *added assumption* that some individual has been chosen randomly, where randomly just means that each individual in the population has an equal probability of being chosen. It follows that the probability that the individual selected has a certain property matches the frequency of the property in the population. We can then use these probabilities as the measures of inductive support for the propositions that the individual has various educational levels and various employment statuses. That the individual has no high school diploma increases the inductive support for the proposition that the individual is unemployed; for the probability of unemployment given no high school diploma is greater than the unconditioned probability of unemployment.

Inductive inferences concerning genetic mutations in some population combine the essential features of the last two cases. To make matters concrete, consider a human population in which a mutation of some particular gene arises, but only very rarely. To make the example more interesting, assume that the mutation can arise in n mutually exclusive variations, so we have possible alleles

$$N, m_1, \ldots, m_n$$

where N is the overwhelmingly most common case of no mutation—hence the symbol N for "No." We have a population of alleles in which the n mutations will arise with varying frequency. Physical chance process will govern the propagation of the alleles through the generations and those physical chances will determine the equilibrium distribution frequency of the various

alleles. If standard, idealized conditions are met, these frequencies will conform with the Hardy-Weinberg equilibrium.

2.2 Inductive Inference Problems

The conditions just specified are the background facts that warrant inductive inferences over the presence of the mutation in the population. Since the physical chances are probabilistic, these inferences will be within a probabilistic inductive logic.

Consider some randomly selected child. The fact of random selection means that the probability that the child carries mutation m_i matches the overall frequency r_i of mutation-i carrying individuals in the population. These facts together warrant our use of probabilities as the measure of inductive support, where those probabilities are matched with population frequencies.

Consider two sibling children in some family. The measures of inductive support that each carries the mutation m_i is given by the two probability measures:[238]

$$P(\text{child}_1 \text{ carries } m_i) = r_i << 1 \quad i=1,\ldots,n \tag{1}$$
$$P(\text{child}_2 \text{ carries } m_i) = r_i << 1 \quad i=1,\ldots,n$$

These two probabilities must also be related by the rule of total probability:

$$P(\text{child}_2 \text{ carries } m_i)$$
$$= P(\text{child}_2 \text{ carries } m_i \mid \text{child}_1 \text{ carries } m_i) \,.\, P(\text{child}_1 \text{ carries } m_i)$$
$$+ P(\text{child}_2 \text{ carries } m_i \mid \text{not-child}_1 \text{ carries } m_i) \,.\, P(\text{not-child}_1 \text{ carries } m_i) \tag{2}$$

In general, the two conditional probabilities in this last formula are quite complicated expressions of the various gene frequencies. However, for the case of extremely rare mutations, that is $r_i << 1$, they are approximated very well by

$$P(\text{child}_2 \text{ carries } m_i \mid \text{child}_1 \text{ carries } m_i) = 1/2$$
$$P(\text{child}_2 \text{ carries } m_i \mid \text{not-child}_1 \text{ carries } m_i) = \frac{r_i}{2(1-r_i)} \tag{3}$$

The first conditional probability arises from the circumstance that, if child_1 carries m_i, then it is overwhelmingly likely that just one of the children's parents carries mutation m_i. It is possible that a parent may carry two copies, or both parents may carry copies, but these cases are vastly less likely and can be neglected. If just one of the children's parents carries mutation m_i, then

---

[238] More exactly, if the allele carrying the mutation m_i arises with frequency f_i in the population and the gene distribution has arrived at the Hardy-Weinberg equilibrium, then the probability that the child carries one or both of the mutated alleles is $r_i = 2\, f_i(1- f_i) + f_i^2 \approx 2f_i$ for small $f_i<<1$.

there is a probability of 1/2 that $child_2$ inherits it. The second conditional is recovered from a short application of Bayes' theorem.[239]

We can use the conditional probabilities (3) to support an inference from the probability that $child_1$ carries m_i to the probability that $child_2$ carries m_i. Substituting (3) into (2) we find:

$$\begin{aligned} &P(child_2 \text{ carries } m_i) \\ &= 1/2 \,.\, P(child_1 \text{ carries } m_i) + \frac{r_i}{2(1-r_i)} \,.\, P(\text{not-}child_1 \text{ carries } m_i) \\ &= 1/2\, r_i \;+\; \frac{r_i}{2(1-r_i)}\,(1-r_i) = r_i = P(child_1 \text{ carries } m_i) \end{aligned} \qquad (4)$$

which agrees with (1).

This last inference is a particular case of how the rule of total probability becomes a rule of inductive inference in the probabilistic logic. Consider an outcome space that can be partitioned into mutually exclusive outcomes in two ways; that is as $\{S_0, S_1, \ldots, S_n\}$ and as $\{R_0, R_1, \ldots, R_n\}$. We start with the probability distribution $P(R_k)$ for $k=0, \ldots, n$ as representing the inductive support for the outcomes R_k. The conditional probabilities $P(S_i \mid R_k)$ for $i,k = 0, \ldots, n$, allow us to infer from the support accrued to the outcomes B_k to the support accrued to the outcomes A_i, by means of the rule of total probability:

$$P(S_i) = \sum_k P(S_i \mid R_k)\, P(R_k) \qquad (5)$$

3. From Mutations to Electrons

Quantum mechanics describes a physical realm that differs from those of more familiar systems in which probabilistic logics are appropriate. The facts that comprise quantum theory can warrant a rather different inductive logic for certain quantum systems. One of these systems, a pair of entangled particles, is analogous to the pairs of children of the mutation case above in that it is comprised of two related systems. However if we try to carry out inductive inferences analogous to those concerning mutations carried by children, we will find that we need to use a non-probabilistic inductive logic and that this logic can be read off directly from the quantum mechanical formalism.

---

[239] Writing $c_1 = child_1$ carries m_i and $c_2 = child_2$ carries m_i, we have from Bayes' theorem that

$$P(c_2 \mid \text{not-}c_1) = \frac{P(\text{not-}c_1 \mid c_2)}{P(\text{not-}c_1)}\; P(c_2) = (1/2)r_i/(1-r_i)$$

since $P(c_2) = r_i$, $P(\text{not-}c_1) = (1- r_i)$ and $P(\text{not-}c_1 \mid c_2) = 1 - P(c_1 \mid c_2) \approx 1\text{-}1/2 = 1/2$.

This is not the place to attempt a self-contained development of the standard formalism of quantum theory.[240] However my concern is that the development is accessible to those who do not work in quantum theory. I will do my best to motivate and explain the least amount needed to convey the main ideas to you, if you have less familiarity with the formalism. So do keep reading—this is written for you.

In the following, we will consider one of the simplest properties of one of the simplest, best-known particles. That is, we will consider electrons and their spins. Electrons carry angular momentum. Classically, angular momentum is a measure of the quantity of rotational motion of a body like a spinning top. It is the rotational analog of ordinary linear momentum—"mass time velocity"—and, for the spinning top, is "moment of inertia times the angular velocity." It is a vector quantity and is fully specified when we have fixed its real number magnitude and its direction in space. The magnitude is determined by the speed of rotation and the mass distribution in the top, as expressed by its moment of inertia. The direction is fixed by the axis of rotation.[241] Angular momentum acquires its importance in both classical and quantum systems since it is a conserved quantity. The total angular momentum remains constant in all closed interactions.

It is almost the same with the spin of the electron. The angular momentum of the electron has a magnitude. Unlike a classical top that can spin faster and slower and thus can carry more or less angular momentum, all electrons carry the same magnitude of spin angular momentum. It is 1/2 in units of $h/2\pi$ (where h is Planck's constant). Since it is the same for all electrons, this magnitude is unimportant for what follows. Like the top, electron spin also has a direction and this direction can take all orientations in space. That direction is the quantity that will interest us. That direction is what is measured in many foundational thought experiments in quantum mechanics.

The major disanalogy between spinning tops and electrons with spin is that there is nothing rotating or spinning inside the electron. An electron carries angular momentum in the same way that it carries electric charge, as a fundamental, irreducible property. There is no deeper story about some hidden, spinning machinery that explains how the angular momentum comes about. It is just there.

---

[240] To fill in the inevitable technical gaps, an account such as Nielsen and Chuang (2010, Ch. I) can be consulted.

[241] Which direction along the axis? Up or down? The right hand rule tells us that, if the direction of rotation follows the direction of the curled fingers of the right hand, then the hand's upright thumb indicates the direction.

4. Two Inductive Inference Problems for Electrons

The background facts fix the inductive logic appropriate to some domain. We can find situations involving electrons in which a familiar probabilistic induction is the appropriate one; and we can find situations in which it is not.

4.1 Uncertainty over Randomly Selected States

Here is an example of the first. Electrons can have spins that point in all directions. Imagine that we have prepared many (unentangled) electrons whose spin directions are uniformly distributed over all directions. One—we know not which—is selected. On the evidence of this set up, how much inductive support is accrued to each possible spin direction? This is *almost* a problem that calls for a probabilistic logic. What is missing are facts that would require this logic. They are easily supplied. We add to the set up that the selection is random, where this means that each candidate electron has an equal probability of selection. It then follows that each spin direction is equally probable for the electron selected and thus equally well supported.

A probabilistic inductive logic is warranted by the set up, for in it all possible outcomes fall under a single probability measure. No quantum peculiarity has entered. The analysis would be the same if, instead of electrons, we had prepared many classical arrows, pointing in different directions, and selected one randomly.

4.2 Uncertainty over Measurements on Electrons in Entangled States

Now consider a second problem. It is possible to entangle two electrons so that their states are highly correlated. In the simplest case of two electrons in a singlet state (to be explained below), the two electrons have spins that always point in opposite directions. If one is measured to have a spin that points North, the other will always be measured to point South; and so on for every other possible pairing of opposite directions. This singlet state can persist even when the two electrons are separated by great spatial distances. They are entangled.

If we have access to one of the electrons in this entangled state, we can perform measurements of the direction of its spin. The measurement process is foundationally quite troublesome in quantum theory, as we shall see below. However, for present purposes, all that matters is that the measurement will yield some definite direction. We do not know in advance which that will be. Quantum theory only gives us probabilities for the different possible directions. Once we know the spin direction of one of the electrons in a singlet state, then we know the spin direction of the other electron, no matter how distant that electron is from us.

The inductive inference problem starts with the evidence that we have two electrons in some state, such as a singlet state. How much support does this evidence give to the various spin direction measurement outcomes that may arise on each of the electrons? How much support does this evidence give to possible connections between the spin direction measurements on the two electrons? These questions are the analogs of those asked above about the children concerning rare mutations. Given the background facts of the distribution of the random mutation, what is the probability that the first child carries the mutation? Given that one carries it, what is the probability that the other does?

There are probabilities in the quantum inductive problem. However they prove not to be the fundamental quantities. The uncertainty is not the sort of probabilistic uncertainty that arises with random selection. For in random selection, there is a single probability measure that covers all possible outcomes. In the quantum case, there is no single probability measure covering all outcomes.

To proceed, we need to develop the elements of the quantum theory of electron spin.

5. Vector Spaces

An electron spin can point in any direction in space. It turns out that we can recover all possibilities if we start with two states, a spin that points up and a spin that points in the opposite direction, down. All other possibilities are recovered by adding together or subtracting — "superposing"—these states. Left and right pointing spin states are recovered by adding and, respectively, subtracting the up and down spin states.

This is not the way more familiar displacement vectors in space add and subtract. If we add a displacement of one foot North to a displacement of one foot South, they cancel. They do not give us a displacement to the East or the West, as would spin vectors. In this respect, spin vectors are not quite like ordinary displacement vectors. However spin vectors do share the essential property with displacement vectors that we can always add two vectors to produce another with an intermediate direction. What counts as an intermediate direction, however, will be different in the two cases.

To keep track of these different directions, we will label them in the familiar way with Cartesian coordinate axes, x, y and z and identify the "up" direction as the positive z direction. That we can add and subtract the different spin states to produce new ones, relying on the fact that they form a vector space.

Dirac's "ket" notation is a convenient and compact way to write the vectors. The vectors of unit length corresponding to the to the +z (up) and -z (down) directions are written as kets |z>

and |-z>. The x and -x pointing vectors of unit length, |x> and |-x>, are recovered by superposition as[242]

$$|x\rangle = \frac{1}{\sqrt{2}}(|z\rangle + |-z\rangle) \qquad |-x\rangle = \frac{1}{\sqrt{2}}(|z\rangle - |-z\rangle) \tag{6}$$

The summations can be pictured in the familiar vector diagram of Figure 1.

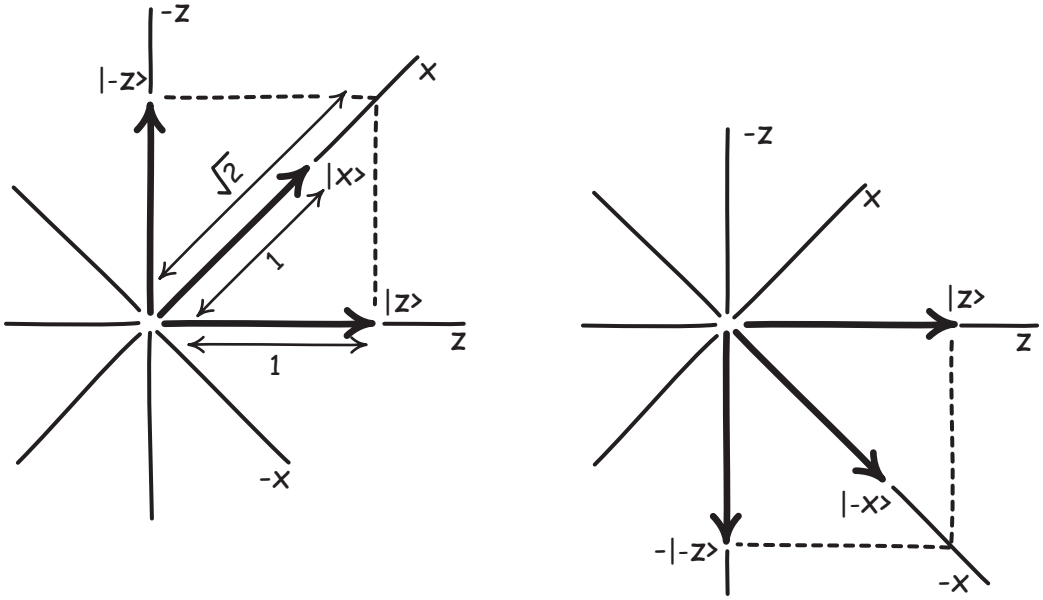


Figure 1. Superposition of Vectors

The figure also makes apparent the need for the factor of $1/\sqrt{2}$. For simply adding vectors |z> and |-z> of unit length produces a vector of length $\sqrt{2}$. It must be rescaled by this factor to recover a unit vector.

So far we have spin states pointing in the x and z directions. We can also introduce spin states in the y direction by means of superpositions that employ $i = \sqrt{-1}$.

$$|y\rangle = \frac{1}{\sqrt{2}}(|z\rangle + i\,|-z\rangle) \qquad |-y\rangle = \frac{1}{\sqrt{2}}(|z\rangle - i\,|-z\rangle)$$

In general any superposition of these vector states produces a new vector state. There is a symmetry among them all; none is more fundamental. We can start by labeling any direction as the z direction and use the above formulae to produce the complete spin space.

Figure 1 allows the vector addition to look like the familiar addition of vector displacements in space. But it is in other ways a poor representation of the spin space. It allows us to draw the vectors |-z> and (-1)|-z> = -|-z> as two separate vectors, with the second pointing

[242] The vector space is a Hilbert space, which means that there is also a notion of the length of the vectors.

in a direction opposite to the first. This gives the appearance of a difference where there is no physical difference. The distinguishing phase factor (-1) in quantum theory has no physical import so that |-z> and (-1)|-z> represent the same state. A simpler picture eradicates the duplication. It is the Bloch sphere shown in Figure 2.

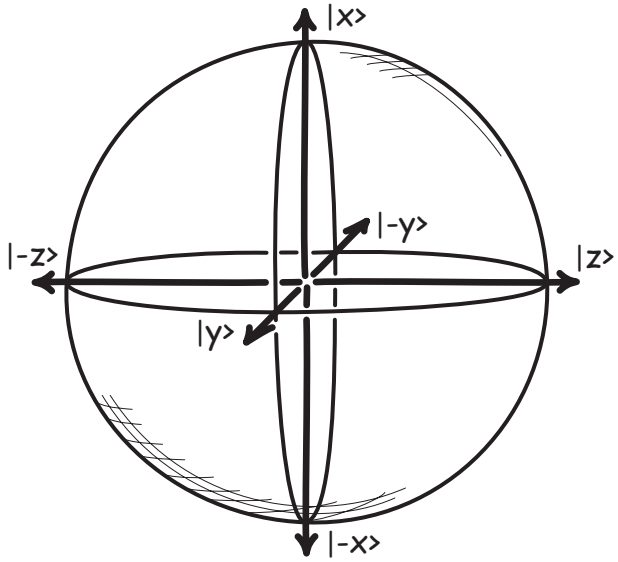


Figure 2. The Bloch Sphere

The figure looks so familiar that it is easy to misread. What is orthogonal—"perpendicular"—to what differs from Euclidean expectations. In this space, |z> and |-z> are orthogonal; as are |x> and |-x>; and |y> and |-y>. Yet |x> and |z> are not orthogonal, even though Euclidean expectations suggest otherwise. The sphere also looks like it is a three dimensional vector space that must be built from three independent basis vectors. However it is a two dimensional space, with |z> and |-z> as its basis vectors. Their linear superpositions can span the whole sphere since complex numbers can be used in forming linear superpositions; and this shift from real to complex numbers gives the added degree of freedom needed.

6. Measurement

6.1 An Oddity in Quantum Theory

In non-quantum systems, measuring the state of a system is merely a technical challenge, not a foundational problem. If we have a spinning top, we can in principle determine the direction of its axis of spin without having to destroy the top. Things are different in quantum theory.

We can learn something of the direction of the spin axis of an electron by passing it through an inhomogeneous magnetic field in a Stern-Gerlach apparatus. The magnetic dipole

moment of the electron aligns with its spin and that moment determines how the electron is deflected by the magnetic field. The direction of the deflection tells us the direction of the spin. We need not delay with further details of this measuring operation save one:

To perform the measurement, we must choose in advance some direction in space along which to align the magnetic field of the Stern-Gerlach apparatus. Our measurement will be performed along that direction. The curious and foundationally troublesome property of measurement in the quantum context is that the measurement will always return a definite result along the direction chosen, no matter what the spin state of electron.

If we measure the z-spin of an electron that has z-spin up, that is, its state is $|z\rangle$, we will measure z-spin up with certainty. If we measure the z-spin of an electron with z-spin down, that is, its state is $|-z\rangle$ we will measure z-spin down with certainty. So far, there is nothing unexpected. But if we measure the z-spin of an electron in state $|x\rangle$ with x spin up, something odd happens. Since a state of x-spin up is different from either z-spin up or z-spin down, you might expect the measurement to fail in some way. It might, perhaps, give a muddled answer of both z-spin up and z-spin down and the same time; or perhaps no result at all. That does not happen. We still get a definite z-spin measurement outcome. It will be either z-spin up or z-spin down, without any confounding. Which of the two will happen? The formalism gives us a probability of 0.5 for each.

6.2 The Born Rule

In general, a z-spin measurement always returns either a z-spin up or z-spin down outcome. The probability of each will vary according to the state measured. Standard quantum theory provides a simple rule—the "Born rule"—for computing these probabilities. Assume that we are measuring the z-spin of an electron with some general state $|\phi\rangle$. We can decompose the state vector $|\phi\rangle$ into two components in the $|z\rangle$ and $|-z\rangle$ directions.

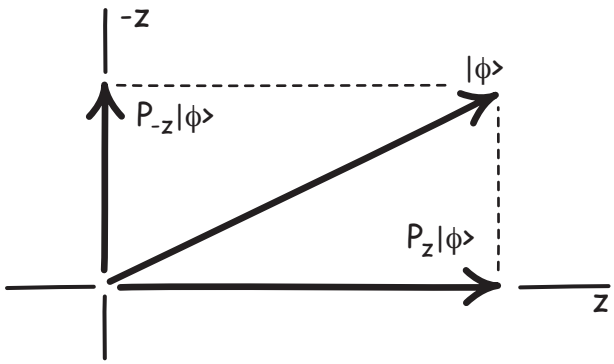


Figure 3. Components of $|\phi\rangle$

The two components are $P_z|\phi\rangle$ and $P_{-z}|\phi\rangle$, where the projection operator P_z picks out the component of $|\phi\rangle$ in the $|z\rangle$ direction; and the projection operator P_{-z} picks out the component of $|\phi\rangle$ in the $|-z\rangle$ direction. The vector $|\phi\rangle$ is the sum of these two components:

$$|\phi\rangle = P_z|\phi\rangle + P_{-z}|\phi\rangle \tag{7}$$

The Born rule tells us that the probability of measuring each outcome is given by the $(\text{length})^2$ of each of these two component vectors, where we recall that by supposition $|\phi\rangle$ has unit length.

$$\text{Probability (z-spin up)} = (\text{length } P_z|\phi\rangle)^2 \tag{8}$$

$$\text{Probability (z-spin up)} = (\text{length } P_z|\phi\rangle)^2$$

For the general case of a $|\psi\rangle$ measurement on a state $|\phi\rangle$, we have

$$\text{Probability } (|\psi\rangle \text{ on } \psi\text{-measurement of } |\phi\rangle) = (\text{length } P_\psi|\phi\rangle)^2 \tag{9}$$

For the case of $|\psi\rangle = |z\rangle$ and $|\phi\rangle = |x\rangle$, we have from (6) that

$$|x\rangle = \frac{1}{\sqrt{2}}(|z\rangle + |-z\rangle)$$

so that

$$P_z|x\rangle = \frac{1}{\sqrt{2}}|z\rangle \qquad P_{-z}|x\rangle = \frac{1}{\sqrt{2}}|-z\rangle$$

as shown in Figure 4. The probability of each outcome is just $(\text{length})^2 = \left(\frac{1}{\sqrt{2}}\right)^2 = 0.5$.

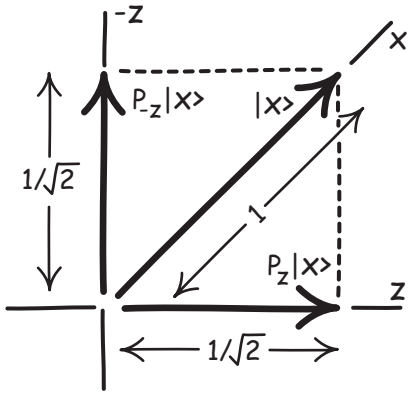


Figure 4. Projections of $|x\rangle$

6.3 The Basis of the Difference Between Probabilistic and Quantum Inductive Logics

That the Born rule gives us the correct probabilities for measurement outcomes is well established by experiment. How it can do it and what is happening during the measurement process, however, remains a troublesome issue in the foundations of quantum theory.

In the standard, text-book account, the electron state vector of the electron undergoing measurement "collapses" onto one of the two measurement states |z> or |-z>, with the probabilities given by the Born rule. That is, measurement instantly transforms a |x> state into a different one, a |z> state or a |-z> state, according to the outcome. Measurement *changes* the state. That measurement can do this is odd and puzzling. Yet it is an essential part of the standard account of quantum theory. An expansive literature has sought to find alternative accounts of measurement that avoid this oddity. None has produced a view that has been accepted widely enough to be the new standard.

Fortunately, my present purposes require no decision on how the measurement problem should be solved. I need adopt only the bare account in which the Born rule gives us the correct probabilities for measurement outcomes.

This oddity of quantum theory is decisive as far as inductive logics are concerned. For the probabilities introduced by measurement do not merely reflection an uncertainty over which prior, existing state is at hand. Measurement changes the state and then attaches probabilities to the result. As a result, the probabilities of outcomes associated with different measurement scenarios cannot be combined into a single probability measure. Rather, a different quantity synthesizes these measures and that quantity forms the basis of a quantum inductive logic.

7. Density Operators

The goal here is to find the inductive logic warranted by the quantum facts concerning electrons in entangled states. To proceed, we need to identify the structure in the quantum case that is analogous to the probability measure of probabilistic logic. That inductive structure is the density operator. It arises as follows.

For a single particle, in the most definite case, we assuredly have just one quantum state, such as |z>. It is called a "pure state." What if we are uncertain as to which of two such pure states, |z> and |-z>, is at hand? It would be nice if our uncertainty could be captured merely by taking a suitably weighted sum of the two pure state vectors. That simple option fails. We already saw that adding these two vectors just gives us another pure state vector. If we add them with equal weight, for example, we merely recover |x>, as (6) shows.

While this simple option fails, something very close to it succeeds. An alternative way of representing a pure state is by a projection operator. There is a one to one correspondence

between them, so picking one amounts to picking the other. We have already seen projection operators in the context of the Born rule of measurement above in equation (8). They pick out the component of a vector parallel to the direction of projection. For each unit vector, such as $|z\rangle$, there will be just one projection operator that finds all of $|z\rangle$ to be in the direction in which it projects. We have written that unique projection operator as P_z. More compactly, the pure state $|z\rangle$ is associated uniquely with the projection operator P_z that has the property that $P_z\,|z\rangle = |z\rangle$.

Since these projection operators are a special case of density operators, let us explore them a little more. Operators in vector spaces are the analogs of functions in ordinary algebra. A function maps numbers to numbers. The square function maps 2 to 4, 3 to 9, and so on. An operator in the vector space maps vectors to vectors. The projection operator is one of the simplest. The behavior of the projection operator P_z associated with the vector $|z\rangle$ is fully specified by the fact that it takes $|z\rangle$ back to itself and takes the vector $|-z\rangle$ to zero:

$$P_z\,|z\rangle = |z\rangle \qquad P_z\,|-z\rangle = 0 \tag{10}$$

and that it is linear, so that

$$P_z\,(A\,|z\rangle + B)\,|-z\rangle) = A\,P_z\,|z\rangle + P_z\,|-z\rangle)$$

for all complex numbers A and B. Since an arbitrary vector $|\psi\rangle$ can always be written as this sort of linear sum $|\psi\rangle = (A\,|z\rangle + B)\,|-z\rangle)$, linearity and (10) fix how the projection operator acts on any vector.

Now we return to the original problem. What if we are unsure as to which of $|z\rangle$ and $|-z\rangle$ is at hand? As long as we represent the states directly by vectors, we *cannot* just add the two vectors in a suitably weighted summation. We saw that would give us a new vector, which is just a different pure state. If we represent states with projection operators, then we *can* add them without this happening. If we weight the two states equally, then we produce the new operator for a so-called "mixed state," in contrast to the pure states with which we started:

$$\rho_{max} = \tfrac{1}{2}P_z + \tfrac{1}{2}\,P_{-z} \tag{11}$$

The subscript "max" indicates that the state is maximally mixed—that is, as far away as possible—from a pure state. (We will see how this comes about below.) This new operator is no longer a projection operator.[243] It is a density operator. We do not need to use the $\frac{1}{2}$ to $\frac{1}{2}$

[243] The quickest way to see that is to note that projection operators have the property of "idempotency." That is, after they have been applied once, nothing changes if they are applied a second or third time. That is, $P_z\,P_z = P_z$ and $P_{-z}\,P_{-z} = P_{-z}$. The operator ρ_{max} is not idempotent,

weighting. We merely need to use two positive real weights that sum to unity. The $\frac{1}{2}$ to $\frac{1}{2}$ weighting, however, is the case that will interest us most. We arrive at the most general density operator for the single electron spin by choosing arbitrary positive, real number weights w_z and w_{-z},

$$\rho = w_z P_z + w_{-z} P_{-z} \tag{12}$$

such that the weights sum to unity, $w_z + w_{-z} = 1$.

At this stage, it looks as if the density operators of (11) and (12) are behaving just like probability measures. We appear to be uncertain over which of $|z\rangle$ or $|-z\rangle$ we have with probabilities w_z and w_{-z}, respectively. That appearance is reinforced by the term "mixed state." *Something* like this is correct. But it is not quite like this. The unqualified term "mixed state" is misleading and it is in the qualifications needed that the novelty of the quantum logic will be found.

8. Tensor Product Spaces

A density operator is the appropriate structure for an inductive logic when we are inferring inductively over the properties of electrons in entangled states. These states arise as follows. Consider two electrons. Each has its own spin vector space. The first is formed by taking all linear superpositions of the states $|z\rangle_1$ and $|-z\rangle_1$ of the first electron. The second is formed by taking all linear superpositions of the states $|z\rangle_2$ and $|-z\rangle_2$ of the second particle. (The subscripts 1 and 2 just number the particles.) The two electrons together form a combined physical system with its own vector space. One state in it will be a product state such as $|z\rangle_1|z\rangle_2$. That is, the first electron state is z-spin up and the second is z-spin up also. All four of these possibilities are

$$|z\rangle_1|z\rangle_2\ ,\ |z\rangle_1|-z\rangle_2,\ |-z\rangle_1|z\rangle_2,\ |-z\rangle_1|-z\rangle_2$$

We form a new vector space, the combined space of all possible states of the two particles, by taking all linear superpositions of these four states. The space is formed in the same way as we formed the one electron vector space by taking all linear superpositions of $|z\rangle$ and $|-z\rangle$. This new space is the tensor product of the vector spaces associated with the individual particles.

---

since $\rho_{max}\,\rho_{max} = \frac{1}{4}P_z\,P_z + \frac{1}{4}\,P_{-z}\,P_{-z} + \frac{1}{4}P_z\,P_{-z} + \frac{1}{4}\,P_{-z}\,P_z = \frac{1}{4}P_z + \frac{1}{4}\,P_{-z} = \frac{1}{2}\,\rho_{max} \neq \rho_{max}$. (Note $P_z\,P_{-z} = P_{-z}\,P_z = 0$.)

This new vector space contains many new states. We will investigate one, the so called "singlet state" of total spin angular momentum of zero. It is [244]

$$|s\rangle = \frac{1}{\sqrt{2}} (|z\rangle_1|\text{-}z\rangle_2 - |\text{-}z\rangle_1|z\rangle_2) \qquad (13)$$

It is a superposition of two states: $|z\rangle_1|\text{-}z\rangle_2$ in which the first particle spin points "up" and the second "down"; and $|\text{-}z\rangle_1|z\rangle_2$ in which the first particle spin points "down" and the second "up."

9. Reduced Density Operators

Consider two entangled electrons, such as the singlet state (13). The two electrons can remain entangled in the singlet state, even when they are widely separated spatially. If we have access to just one of these electrons, we can make a measurement of the spin direction of that one electron. The entanglement means that whatever measurement outcomes we obtain on our nearby electron will be correlated with the measurement outcomes that someone else finds on the other remote electron. We read that correlation directly from the two terms in the singlet formula (13). The first term $|z\rangle_1|\text{-}z\rangle_2$ tells us that whenever the first electron produces z-spin up on measurement, the second electron produces z-spin down (and conversely). The second term $|\text{-}z\rangle_1|z\rangle_2$ tell us that whenever the first electron produces z-spin down on measurement, the second produces z-spin up (and conversely). In short, our measurement on the nearby electron will always give a spin of the opposite direction from the result of a measurement on the remote electron.

When we make our measurements on the nearby electron, we will know nothing of these remote outcomes. Let us set them aside and ask what outcomes we should expect for measurements on the one electron to which we have access. Quantum theory provides the following recipe for determining the probabilities of the various outcomes.

The first step is to eliminate explicit appearance of the second, remote electron from the description of the two electron system to arrive at a reduced description of the first, nearby electron only. We begin by replacing the vector representation of the entangled state by its corresponding projection operator, P_{12}. For example, the projection operator associated with the pure singlet state $|s\rangle$ can be written as a sum that includes projection operators associated with the individual particles that comprise it:

$$P_{12} = P_s = \tfrac{1}{2}\, P_{z,1}\, P_{\text{-}z,2} + \tfrac{1}{2}\, P_{\text{-}z,1}\, P_{z,2} + \text{further cross terms} \qquad (14)$$

[244] The factor of $1/\sqrt{2}$ ensures that the state $|s\rangle$ has unit length. Since the spins in each term point in opposite directions, the total angular momentum of the singlet state is zero.

where the ",1" and ",2" notation labels the nearby and remote electrons (respectively) to which the individual projection operators belong. The "further cross terms" contain operators that are not projection operators. While important in some applications, these further terms drop out of the calculations below.[245]

We now suppress the details of the second remote particle "2" by means of a "trace" operation "Tr." This linear operator replaces the degrees of freedom in its scope by their expectation values. The trace operator Tr_2 of the remote electron vector space suppresses the properties of the remote electron. If P_{12} is the projection operator associated with the entangled pair of electrons, we arrive an operator that represents the properties of the first electron only by means of

$$\rho_1 = Tr_2[P_{12}] \tag{15}$$

The operator ρ_1 need no longer be a projection operator, but will in general be a density operator. Since they are produced in the reducing of the two electron vector space to a one electron space, they are called reduced density operators. For the case of the singlet state when $P_{12} = P_s$, we have [246]

$$\rho_{s1} = Tr_2[P_s] = \tfrac{1}{2}\ P_{z,1} + \tfrac{1}{2}\ P_{-z,1} = \rho_{max,1} \tag{16}$$

The operator ρ_{s1} is not a projection operator.

That the reduced density operator for the nearby electron is not a projection operator captures the fact that the electron is in no definite spin state. If the entangled pair is in a singlet state, then the reduced density operator of the nearby electron (16) is the maximally mixed state (11). One might expect that the two factors of $\frac{1}{2}$ are just the probabilities of measuring z-spin up and measuring z-spin down. They are.

---

[245] For completeness, the "further cross terms" are $-\frac{1}{2}|z\rangle_1\langle -z|_1|-z\rangle_2\langle z|_2 - \frac{1}{2}|-z\rangle_1\langle z|_1|z\rangle_2\langle -z|_2$ where the linear operator $|z\rangle_1\langle -z|_1$ maps $|-z\rangle_1$ to $|z\rangle_1$ and $|z\rangle_1$ to 0; and so on for the remaining three operators.

[246] Since $Tr_2\,[P_{z,2}] = Tr_2\,[P_{-z,2}] = 1$ and the trace operator is linear, we have

$Tr_2\,[P_s] = Tr_2\,[\ \frac{1}{2}\ P_{z,1}\,P_{-z,2} + \frac{1}{2}\ P_{-z,1}\,P_{z,2}$ + further cross terms]

$= \frac{1}{2}\ P_{z,1} Tr_2\,[P_{-z,2}] + \frac{1}{2}\ P_{-z,1} Tr_2\,[P_{z,2}] = \frac{1}{2}\ P_{z,1} + \frac{1}{2}\ P_{-z,1}$, where Tr_2 [further cross terms] = 0.

This follows from the Born rule (9) for measurement outcomes for density operators. In its general form, the rule says that the probability of measuring a spin state $|\psi\rangle$ when we have an electron described by a density operator ρ is[247]

$$\text{Probability } (|\psi\rangle \text{ on } \psi\text{-measurement of } \rho) = \text{Tr}[P_\psi \rho] \tag{17}$$

The projection operator P_ψ is just the projection operator associated with the vector $|\psi\rangle$. Applying this formula to the maximally mixed state ρ_{max} we find:[248]

$$\text{Probability (z-spin up on z-spin measurement of } \rho_{max}) = \text{Tr}[P_z \rho_{max}] = \tfrac{1}{2}$$

$$\text{Probability (z-spin down on z-spin measurement of } \rho_{max}) = \text{Tr}[P_{-z} \rho_{max}] = \tfrac{1}{2} \tag{18}$$

10. Density Operators do not Represent Probabilistic Ignorance of a Unique, True State

10.1 Many Probability Measures

The density operator ρ_{max} for the maximally mixed state looks *initially* as if it just represents a familiar probabilistic uncertainty over whether the true state is $|z\rangle$ or $|-z\rangle$. The two coefficents of $\frac{1}{2}$ for the states $|z\rangle$ and $|-z\rangle$ in the expression (11) reappear as the probabilities of measuring these states according to the Born rule (18).

What makes this mixed state different from mere probabilistic uncertainty is an important fact about the density operators of mixed states: they can be written in many ways, each indicating a different sort of uncertainty with a distinct probability measure associated with it. That makes the term "mixed state" potentially quite misleading. The state is not a simple mixture

---

[247] While it is written differently, this version of the Born rule is equivalent to (9). Briefly, to go from (17) to (9), set ρ as the projection operator P_ϕ associated with the pure state $|\phi\rangle$, then $\text{Tr}[P_\psi P_\phi] = (\text{length } P_\psi|\phi\rangle)^2$. To go in the reverse direction, set the pure state $|\phi\rangle$ in (9) to be a many electron entangled state and P_ψ the projection operator associated with the $|\psi\rangle$ state of one of the entangled electrons.

[248] We have $\text{Tr}[P_{|z\rangle} \rho_{max}] = \text{Tr}[P_z (\frac{1}{2} P_z + \frac{1}{2} P_{-z})] = \text{Tr}[\frac{1}{2} P_z P_z + \frac{1}{2} P_z P_{-z}] = \text{Tr}[\frac{1}{2} P_z] = \frac{1}{2} \text{Tr}[P_z] = \frac{1}{2}$, where we have used that $\text{Tr}[P_z] = 1$, $P_z P_z = P_z$, $P_z P_{-z} = 0$ and the linearity of Tr.

that can be decomposed uniquely into its components. It is not like a mixture of sand and iron filings that can be unmixed uniquely with a magnet.

Since this is the key point for all that follows, let us be clear on how this comes about. The density operator is simply a map that takes vectors to vectors. Two density operators are the same if they map the same vectors to the same vectors. In this respect, they are no different from ordinary functions. Take $f(x) = x^2$. It is a function that maps numbers to their squares. While their expressions look different when written down, the functions $g(x) = (x+1)(x-1) +1$ and $h(x) = (x+2)(x-2) + 4$ perform exactly the same mappings. So they are the same function.

It turns out that the mapping of the maximally mixed state ρ_{max} of (11) can represented equally well by many equivalent expressions

$$\rho_{max} = \tfrac{1}{2} P_x + \tfrac{1}{2} P_{-x} = \tfrac{1}{2} P_y + \tfrac{1}{2} P_{-y} = \tfrac{1}{2} P_\psi + \tfrac{1}{2} P_{-\psi} = \tfrac{1}{2} I \qquad (19)$$

Here P_x is the projection operator associated with $|x\rangle$, P_y with $|y\rangle$, etc. and P_ψ is the projection operator associated with some arbitrarily chosen unit vector $|\psi\rangle$ in the Bloch sphere, pointing in any direction. I is the identity map that takes each vector back to itself.

Each of the expressions for ρ_{max} in (19) represent the same map on the vector space, which is written most simply as the last expression on the list, $\frac{1}{2} I$. That is, ρ_{max} is the map that merely takes each vector in the space back to a half-sized version of itself. To see that they are equivalent, we need only recall from (7) that an arbitrary vector $|\phi\rangle$ is the sum of its two components, when decomposed in the $+\psi$ and $-\psi$ directions:

$$|\phi\rangle = P_\psi |\phi\rangle + P_{-\psi} |\phi\rangle = (P_\psi + P_{-\psi}) |\phi\rangle$$

It follows that $(P_\psi + P_{-\psi})$ is just the identity operator I—that is the operator that merely maps a vector back to itself. Thus $\frac{1}{2} (P_\psi + P_{-\psi}) = \frac{1}{2} P_\psi + \frac{1}{2} P_{-\psi} = \frac{1}{2} I$. This is true no matter which unit vector $|\psi\rangle$ is used to define it. Thus the maximally mixed state density operator ρ_{max} is defined equally well by any of the formulae in (19).

These equivalent representations of the maximally mixed state ρ_{max} provide further probabilities for measurement outcomes analogous to (18):

$$\text{Probability(x-spin up on x-spin measurement of } \rho_{max}) = 1/2 \qquad (20)$$

$$\text{Probability(x-spin down on x-spin measurement of } \rho_{max}) = 1/2$$

$$\text{Probability(y-spin up on y-spin measurement of } \rho_{max}) = 1/2$$

$$\text{Probability(y-spin down on y-spin measurement of } \rho_{max}) = 1/2$$

Probability(ψ -spin up on ψ -spin measurement of ρ_{max}) = 1/2

etc.

10.2 No Single Probability Measure Unifies Them

The combined measurement outcomes of (18) and (20) are incompatible with the ordinary notion of probabilistic uncertainty as mere ignorance of some definite but unknown state. That sort of ignorance can be captured by a single probability measure, whereas there can be no single probability measure covering all the results of (20). For each of the states returned by measurement are incompatible with all the others. An x-spin up state is different from either a y-spin up and a z-spin up state. An effort to treat these probabilities as generated by ignorance over some true but unknown state fails and does so rapidly.

Take the probabilities of (18). If we interpret them as this sort of ignorance, then we have with probability one that the true state of the system is |z> or |-z>. For the two states are mutually exclusive so that

P(state is truly |z> or |-z>) = P(state is truly |z>) + P(state is truly |-z>) = 1/2 + 1/2 = 1

It now follows that the probabilities of all the other states must be zero, which contradicts the probabilities reported in (20).

Might we solve the problem with a simple expedient? Take a large outcome space whose primitive events are of the form:

We measure x-spin and recover x-spin up.

We measure x-spin and recover x-spin down.

We measure y-spin and recover y-spin up.

We measure y-spin and recover y-spin down.

etc.

We can form a single probability measure over this larger outcome space, such that the probabilities of (20) can be recovered as conditional probabilities. For example:

Probability(x-spin up on x-spin measurement)

= Probability (we measure x-spin and recover x-spin up | we measure x-spin)

The difficulty with this proposal is that our space now includes probabilities over our freely chosen actions, such as:[249]

Probability (we measure x-spin)

[249] Since Probability (we measure x-spin) = Probability(x-spin up on x-spin measurement) + Probability(x-spin down on x-spin measurement).

The probabilities of (20) are provided directly by quantum theory itself. These new probabilities over our actions bring nothing but trouble. What grounds these new probabilities? To secure a grounding in physical chances, we might employ some physical randomizer to instruct us in which measurement to make. Then our inductive logic has been restricted to this special case. Or if we wish to leave the setting as open as possible, then that very openness means that there are no specific facts that warrant the introduction of the probabilities. In the worst case, they are arbitrarily chosen subjective probabilities and we corrupt the objectivity of our inductive logic by mingling them with the objective probabilities of (20). Setting aside this extreme case, we have still compromised the quantum inductive logic by interweaving inductive support from two distinct arenas: the inductive support for various quantum measurement outcomes as guided by quantum theory; and the inductive support for certain of our choices as guided the vagaries of the human circumstances surrounding our choices.

These are serious difficulties and best avoided. Inductive support for quantum outcomes ought to be independent of human affairs. There is no need for us to face these difficulties. For nothing compels us to combine the probability measures of (20) into a single huge measure. We can arrive at an inductive logic that does not need them, as long as we are willing to give up the idea that an inductive logic must be probabilistic.

10.3 Density Operators as the Fundamental Inductive Structures

The maximally mixed state ρ_{max} already represents some sort of uncertainty over the electron state. It is not the same as the probabilistic uncertainty familiar from cases of ignorance arising through random sampling, for such uncertainty cannot issue in the measurement probabilities (20). The direct way to understand the sort of uncertainty represented by ρ_{max} is that it is the inductive structure that does manifest as the infinite list of the measurement probabilities (20). It is a compact representation of them all.

For many, the predisposition to favor probabilities is strong. They might be inclined to say that means that the logic is still probabilistic--here finally we have probabilities. However these probabilities are not the central quantities. They are intermediates that mediate between the density operator and the measurement outcomes. To capture the inductive situation fully, we need the entire infinite set. It is insufficient merely to report a subset associated with fewer than all directions of measurement. One cannot use the rules of the probability calculus to infer from the measurement probabilities for x-spin measurements, for example, to those for y-spin measurements.

The density operator is the natural and compact representation of the capacity of the electron to deliver different measurement results. When we form the new inductive logic adapted

to this quantum case, the density operator is the central quantity that replaces the probability measure of the more familiar probabilistic inductive logics. It is the quantity that figures centrally in the physics of entangled electrons, in the same way as physical chances figure centrally in the physics of roulette wheels. It is the quantity around which we should build an inductive logic for entangled electrons, just as we build an inductive logic for roulette wheel outcomes around physical chances.

Note for experts in quantum foundations: My goal here is not to contribute to the literature in the foundations of quantum theory. Rather it is to find a context in which a non-probabilistic inductive logic is warranted. Such a context arises, I argue here, with the bare version of quantum theory that merely employs the Born rule to determine measurement outcomes but does not probe what happens in the measurement process. If we deviate from this bare formulation, matters may change. If, for example, we adopt a Bohmian approach, then we augment our ontology to include hidden electron position properties, possessed always by electrons and revealed on measurement. Our uncertainties may then revert to the sort of probabilistic uncertainties that arise with random sampling. Exploring that possibility is not my project here.

11. Is the Density Operator Really an Inductive Structure?

Is it really admissible to treat density operators as inductive structures that can serve in an inductive logic? They seem to be a poor choice for it is hard to say precisely what sort of uncertainty they represent. They do not represent the familiar sort of uncertainty captured by probabilities. Why should we erect an inductive logic for quantum theory around density operators when, perhaps with some effort, we might find a way to replace them with probability measures?

The short answer is that we should use these density operators since they are the appropriate structures delivered by the applicable physics. The uncertainty they represent is more opaque to us than that represented by a probability measure merely because the latter are familiar and their problems largely tamed. We should not mistake the resulting transparency of probability measures for their necessity in inductive logics. Indeed the sorts of analyses that make probability measures interpretationally transparent can be applied equally successfully to density operators.

To see this, note that probability measures initially require considerable interpretive work before their meaning becomes clear or clear enough. If we are unprepared, we encounter severe difficulties when we try to give an explicit definition of probability talk. The challenge is to complete the formula:

"An outcome has probability 0.65."	means	[some text *here* that does not already contain "probability"].

The difficulty is that "probability" always seems to creep into the text requested. We cannot complete the formula by saying that the frequency of success in repeated trials approaches 0.65 in the limit of arbitrarily many trials. We must say that this limit is approached *with probability one*.

While these are serious difficulties, they do not mean that probability talk is meaningless. Indeed, we can constrain the meaning of probability talk quite effectively with a few familiar devices that amount to partial, implicit definitions.[250] First, we require that a probability measure conforms to the standard axioms of probability theory. Second we give interpretations of certainty to the probabilistic extremes: probability one is certainty of occurring; probability zero is certainty of not occurring. We can use these components to provide interpretations for cases of intermediate probability. The trick is to embed the probability talk into a larger discourse in which the already interpreted cases of unit or zero probability arise.

For example, take the proposition that an outcome has probability 0.65. It is a theorem of the probability calculus that, with probability one, on repeated independent trials, the outcome will arise with a frequency that approaches 65%. Most people find that gives them enough to grasp the difference between the two propositions:

An outcome has probability 0.65.

An outcome has probability 0.05.

Loosely speaking, the first outcome happens 13 times as often in repeated trials.

If this sort of interpretive apparatus is sufficient to dispel the clouds around probability talk, then the clouds surrounding the density operator as an inductive structure can also be dispelled. For a quite analogous interpretive apparatus can be employed for them.

---

[250] I set aside other approaches that interpret probabilities operationally in terms of the behavior supposedly manifested by people who harbor those probabilities as belief states. For example, to believe that the probability of an outcome is 1/2 is to be equally ready to accept either side of an equal stakes bet on the outcome. In so far as these operational definitions are constitutive of the probability of an inductive logic, they must be resisted. They entangle probabilities with human utilities and that is a mortal threat to the objectivity of the bearing of evidence in a probabilistic logic. For our preference for $100 over $10 ought to have no bearing on whether observation of the 3K cosmic background radiation increases the probability of the big bang.

First, density operators obey a quite definite axiom set and thereby accrue meaning implicitly, just as do probabilities.[251] Second we can identify extreme cases. The most definite is the density operator corresponding to a pure state such as |z>, a projection operator such as P_z. This projection operator expresses certainty that we *do* have the state |z>; and certainty that we *do not* have the state |-z>. It is analogous to an outcome of probability one if |z> is the true state at hand. The maximally mixed density operator of ρ_{max} is the least definite. It favors all spin directions equally, for under it every possible spin direction has the same probability upon measurement. It is the analog of a uniform probability measure that assigns the same probability to all simple outcomes.

These most and least definite density operators are the extreme cases. For all the intermediate cases, we will be able to give a list analogous to (20) of the probabilities of all possible measurement results. That list gives us the same sort of interpretive purchase on the associated density operator as does saying something like "probability 0.65 means that the outcome happens roughly 65% of the time." Analogously we can say that having some particular density operator entails that we have such and such probabilities of outcomes on this or that measurement, where the list includes all possible measurement and outcomes. That is, we know probabilistically what it is to have some density operator as an inductive structure in terms or all possible measurement experiences in the world. If we are confident in our understanding of probabilities as inductive structures, then we should be confident in our understanding of these density operators as inductive structures.

12. A Geometric Picture of an Electron Spin Density Operator

Part of our comfort with probability measures is that there are simple physical or geometric models for them. For example, distributing probabilities over different outcomes is akin to dividing a unit mass into parts and locating different parts on the different outcomes. The weight of evidence appears directly in the analogy as a mass. Additivity of the probability measure is captured by the fact that we can only increase the mass on one outcome by reducing the mass on others by exactly the same amount.[252] If we have a probability density over some

---

[251] The details do not matter but are stated here: A density operator is linear operator in the vector space that is positive and of unit trace.

[252] This makes it natural for us to think that increasing belief or inductive support in one outcome *must* come from diminishing it for other outcomes. There is no necessity for this

continuous space of outcomes, we can picture the space as an area and the probability density at each point as the altitude of some mountainous surface spread over it.

A fertile picture of all possible probability distributions over n+1 mutually exclusive outcomes is provided by an n-simplex. For three mutually exclusive outcomes, A, B and C, the n-simplex is a triangle, as shown in Figure 5. The three vertices represent A, B and C and each point in the triangle represents a distinct probability measure. The probabilities of each of A, B or C increase with the proximity of the point to the corresponding vertices. The figure shows contours of constant probability for P(A), P(B) and P(C).

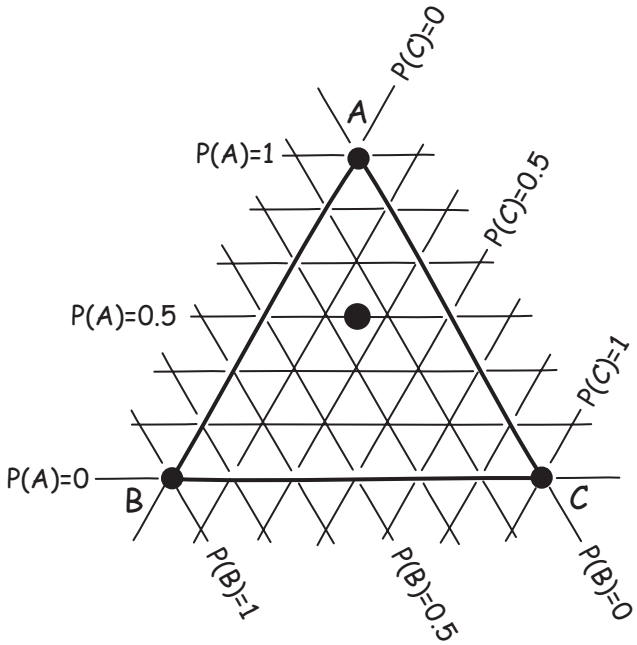


Figure 5. Probability measures for mutually exclusive outcomes A, B and C.

The interior point shown in the figure represents a probability measure for which P(A) = 0.5 and P(B) = P(C) = 0.25. For this measure, P(A) is greater than P(B) or P(C) since the representative point is closer to the A vertex than to the B or C vertex.

In general, there are no correspondingly simple geometric pictures for density operators. The exception is the special case of the spin space of an electron. All possible density operators can be represented elegantly in a three dimensional sphere, as shown in figure 6.[253] Each density operator is represented by a single point in or on the sphere.

---

compensation. It is or it should be a reflection of the fact that out system happens to be one for which additive measures are warranted as the appropriate inductive structures.

[253] This beautiful picture is elaborated in Penrose (2004), §29.4 and Fig. 29.3.

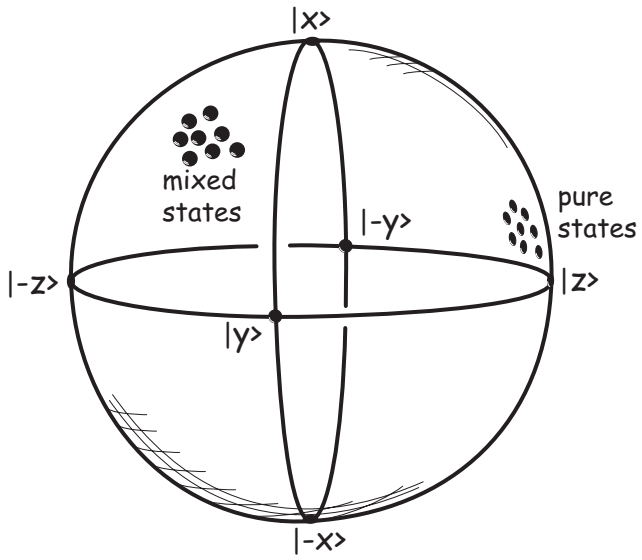


Figure 6. Pure and Mixed States

The pure states, corresponding to projection operators, occupy the surface of the sphere. (This surface by itself is the Bloch sphere we saw in Figure 2 above.) These surface points correspond to the most definite cases of a single pure state. The points inside the sphere represent density operators that are not also projection operators. They represent mixed states. The ones closest to the surface are least mixed and closest in their properties to pure states. The deeper one proceeds inside the sphere the more mixed the states become. The central point is the maximally mixed state ρ_{max}.

The sphere representation also affords a simple picture of which pure states are mixed to yield each density operator. The maximally mixed density operator ρ_{max} lies at the center of the sphere. Any diameter through the center connects two opposite points on the surface of the sphere, as shown in Figure 7. The points connected are two pure states that form ρ_{max}.

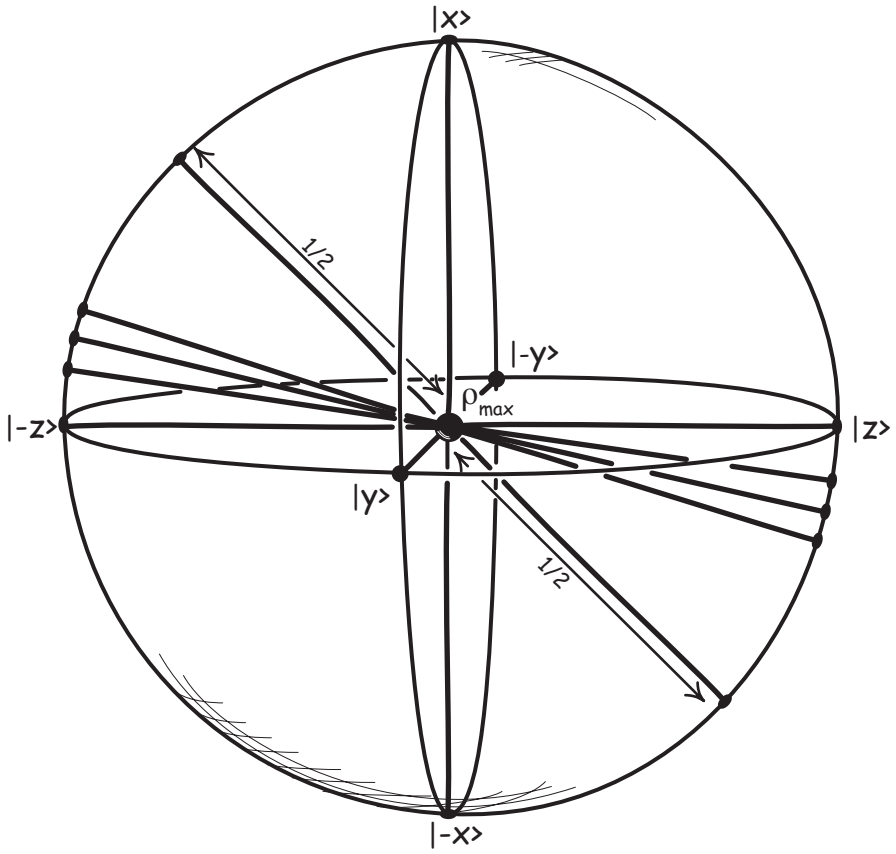


Figure 7. The Maximally Mixed State

We read directly from the figure that ρ_{max} can be formed by equal mixtures of pure states |x> and |-x>; or |y> and |-y>; and so on, as summarized in (19). The multiplicity of possible decompositions of the mixture is represented by the multiplicity of possible diameters through the center.

There is a corresponding representation for the remaining density operators. Consider another density operator ρ that is *not* the maximally mixed ρ_{max}. Any chord through it will intersect the surface of the sphere at two points, as shown in Figure 8.

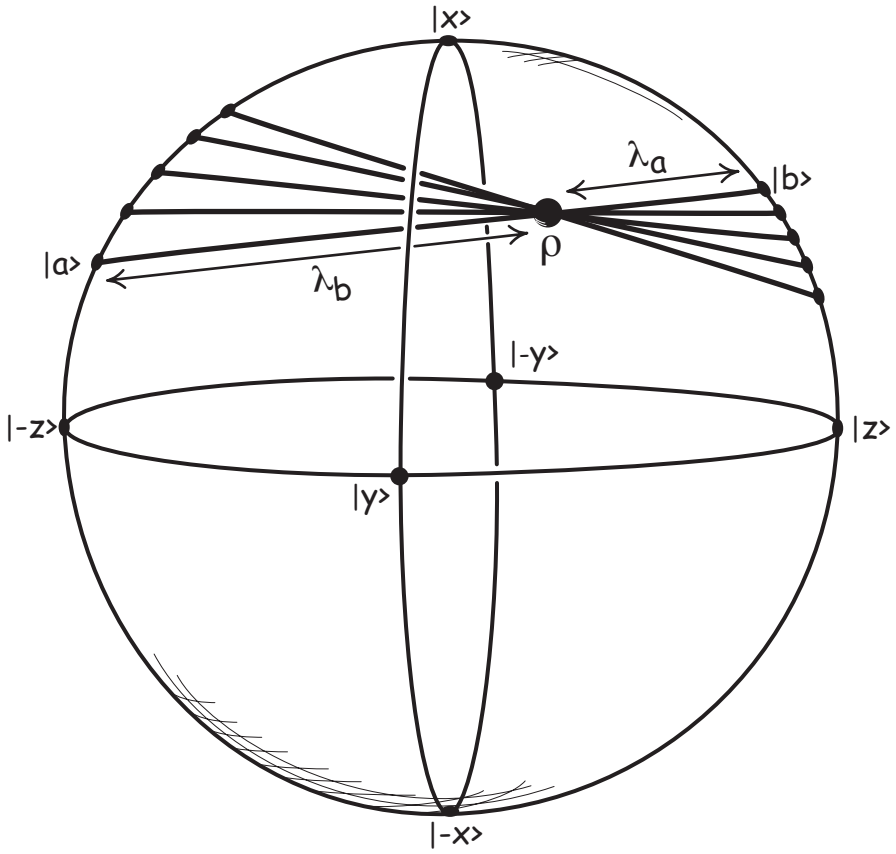


Figure 8. A Mixed State

The two pure states at either end of the chord, |a> and |b>, are the two that are mixed to form ρ. Since there are infinitely many chords through an arbitrary point inside a sphere, a given density operator ρ can be constituted from infinitely many distinct pairs of pure states.

For each case, the geometric construction provides the weighting. The point representing ρ on the chord divides it into two lengths, λ_a and λ_b, where the lengths have been scaled so that $\lambda_a + \lambda_b = 1$. (Note that λ_a is the length of the chord segment between ρ and |b>; and similarly for λ_b.) These are the two weights used to form ρ. That is, if P_a and P_b are the projection operators associated with pure states |a> and |b>, then[254]

[254] (For experts) To see this, note that density operators ρ are mapped onto the unit sphere by $\rho(\mathbf{r}) = (I + \boldsymbol{\sigma}.\mathbf{r})/2$, where $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are the three Pauli matrices and $\mathbf{r} = (x, y, z)$ are the Cartesian coordinates of the unit sphere $\mathbf{r}^2 \leq 1$. (Nielsen and Chuang, p. 105.) A point $\mathbf{r} = \lambda_a \mathbf{r}_a + \lambda_b \mathbf{r}_b$, where $\lambda_a > 0$, $\lambda_b > 0$ and $\lambda_a + \lambda_b = 1$, lies on the straight line connecting $\mathbf{r}_a$ and $\mathbf{r}_b$. Since the

$$\rho = \lambda_a P_a + \lambda_b P_b \qquad (21)$$

The general density operator of (21) can no longer by recovered by tracing away the degrees of freedom of a remote particle in a singlet state (13). We need to replace the entangled singlet state by another. Many choices are possible. A simple one is:

$$|\psi\rangle_{12} = \sqrt{\lambda_a}\ |a\rangle_1|z\rangle_2 + \sqrt{\lambda_b}\ |b\rangle_1|\text{-}z\rangle_2$$

When we trace away the degrees of freedom of the second particle, this operator reduces to the density operator (21).

The maximally mixed ρ_{max} divides each unit diameter into two equal parts of length 1/2 and these weighting factors correspond to the probability of measurement outcomes coinciding with the pure states at either end of the diameter. Something similar holds for the general case of (21), in which the density operator lies on the chord connecting pure states P_a and P_b. We have[255]

$$\text{Probability}(|a\rangle \text{ on an a-measurement of } \rho) = \lambda_a + \lambda_b P(a|b) \qquad (22)$$

where

$$P(a|b) = \text{Probability}(|a\rangle \text{ on an a-measurement of } |b\rangle)$$

That is, the probability of an a-outcome on an a-measurement is given by the weighting factor λ_a, with the addition of a correction factor in P(a|b). This correction factor arises only when the two states mixed, $|a\rangle$ and $|b\rangle$, are not orthogonal, that is, not mutually exclusive. It does not appear in the case of the maximally mixed ρ_{max}, since ρ_{max} arises from mixing orthogonal states such as $|z\rangle$ and $|\text{-}z\rangle$.

Combining all these considerations, we recover a quite serviceable representation of the sort of uncertainty represented by density operators in this simple case. A density operator P_a on the sphere's surface is a projection operator associated with a pure state $|a\rangle$. It is the most definite case. For an a-measurement, it will assuredly give us an a-outcome. A density operator close to P_a will give an a-outcome on a-measurement with high probability. For it will most

---

map is linear, the density operator $\rho(\mathbf{r})$ at $\mathbf{r}$ satisfies $\rho(\mathbf{r}) = \rho(\lambda_a \mathbf{r}_a + \lambda_b \mathbf{r}_b) = \lambda_a \rho(\mathbf{r}_a) + \lambda_b \rho(\mathbf{r}_b)$ and is the λ-weighted sum of the two density operators $\rho(\mathbf{r}_a)$ and $\rho(\mathbf{r}_b)$ at the endpoints $\mathbf{r}_a$ and $\mathbf{r}_b$.

[255] The probability of an a-measurement on $\rho = \lambda_a P_a + \lambda_b P_b$ yielding $|a\rangle$ is

$$\text{Tr}[P_a \rho] = \text{Tr}[P_a (\lambda_a P_a + \lambda_b P_b)] = \lambda_a \text{Tr}[P_a P_a] + \lambda_b \text{Tr}[P_a P_b] = \lambda_a + \lambda_b P(a|b)$$

commonly be associated with a value of λ_a close to one.[256] As the location of the density operator approaches the midpoint, the probability of an a-outcome on a-measurement will approach 0.5, which is the probability associated with the maximally mixed density operator at the center of the sphere. This maximally mixed density operator treats all pure states alike: the probability of an a-outcome on a-measurement is 0.5, no matter what |a> is. That it must do this is immediately clear from the fact that the sphere has a rotational symmetry about the center of the sphere. From that central point, no pure state is closer than any other. It must treat all alike.

13. Leifer and Spekkens' System of Quantum Inference

So far, we have seen only a part of the inductive logic appropriate to entangled electrons. We have identified the reduced density operator in each single electron's vector space as the structure corresponding to the probability measure in a probabilistic logic. We need to do only a little more to specify the full logic. That is, we need a full specification of which density operators arise in which circumstances. As it happens, no further theorizing is needed to arrive at this specification. It is given to us by the standard formalism of quantum theory. When the theory lays out the physics of how the reduced density operators of entangled electrons relate, it is also giving us the inductive logic.

One may wonder, however, if what results really is an inductive logic. If one is used to and is expecting a probabilistic logic, it will be unfamiliar, just as density operators are not quite like probability measures. But that is no reason to dismiss it. Lack of familiarity is not the same as failure.

Leifer and Spekkens (2013) have shown, however, that the inductive logic based on density operators is not so unfamiliar after all. Once we adopt the density operator as the basic inductive structure, they have shown how we can rewrite basic results in quantum theory so that they are structurally analogous to formulae in a probabilistic logic. Their system is elaborate and distinguishes connections between variables according to whether they are causally or acausally related. To give a quite preliminary sense of the system, I will describe how it treats the case of acausally related systems, such as the two particles in a singlet state.

The following Table 1, based on Leifer and Spekkens (2013, p. 7), summarizes the correspondences:

---

[256] If the density operator is close to P_a but λ_a is not close to one, it is because the density operator lies on a chord whose other endpoint, P_b, is also close to P_a. Then the correction term $\lambda_b P(a|b)$ will ensure that the probability of an a-outcome on a-measurement remains high.

Probabilistic logic	*Quantum inductive logic*
Classical variables R, S, … over an outcome space.	Systems A, B, … supporting (Hilbert) vector spaces $H_A, H_B,$
Probability measures P(R), P(S), …	Density operators $\rho_A, \rho_B, \ldots$
Joint probability distribution P(S&R) over Cartesian product space.	Density operator ρ_{AB} over the tensor product Hilbert space $H_{AB} = H_A \otimes H_B$
Conditional probability measure P(S|R) defined through P(S&R) = P(S|R) P(R) P(S|R) = P(S&R) / P(R)	Conditional density operator defined through $\rho_{AB} = \rho_{B\|A} \bigstar \rho_A$ $\rho_{B\|A} = \rho_{AB} \bigstar \rho_A^{-1}$
Normalization $\Sigma_S P(S\|R) = 1$	Normalization $Tr_B (\rho_{B\|A}) = I_A$ where I_A is the identity operator in H_A.
Total probability[257] $P(S) = \Sigma_R P(S\&R) = \Sigma_R P(S\|R) P(R)$	$\rho_B = Tr_A (\rho_{AB}) = Tr_A (\rho_{B\|A} \bigstar \rho_A)$

Table 1. Correspondences between Probabilistic and Quantum Logics

These correspondences are fairly straightforward. In the mutation example, the classical variables R, S, … are the genetic make-ups of each child. When R, S, etc. take specific values, then the genetic makeup of the child is specified as a particular mutation, $m_1, m_2, \ldots$ and their totality forms the outcome space. In the case of entangled electrons, systems A, B, … correspond to $electron_1, electron_2, \ldots$ and the vector spaces $H_A, H_B, \ldots$ are the vector spaces of electron states described above.

The remaining formulae have been written in a way that emphasizes the parallels between the two cases. The classical summation operation "$\Sigma_S \ldots$" sums away the variable S. Correspondingly, the trace operator Tr_B (...) averages away the degrees of freedom associated with B. The star operation $\bigstar$ is a particular multiplication operation designed to keep the parallel in the formulae as close as possible. The goal is to find the quantum analog of P(S&R) = P(S|R) x P(R), where the "x" is just ordinary arithmetic multiplication, since the probabilities P(.) are

[257] This the same rule at (5) above, but here written in the notation used by Leifer and Spekkens.

real numbers. One might write $\rho_{AB} = \rho_{B|A} \cdot \rho_A$, as a direct analog of the probabilistic formula. But caution is needed, since there are important disanalogies. The operation joining $\rho_{B|A}$ and ρ_A is not simple multiplication, but the sequential application of operators, since $\rho_{B|A}$ and ρ_A are operators that act on vectors. This produces two problems.

The first is that the two operators act on different vector spaces. $\rho_{B|A}$ acts on vectors in $H_A \otimes H_B$. ρ_A acts on vectors in H_A. If they are to be combined, they must act on the same vector space. The simple remedy is to expand ρ_A to $\rho_A \otimes I_B$, where the addition of $\ldots \otimes I_B$ makes a new operator that acts as ρ_A on H_A and as the identity ("do nothing") on H_B.

The second is that the order in which we combine the operators will matter, whereas it does not matter when we multiply real numbers. The formula $\rho_{B|A} \cdot \rho_A$ says first act with ρ_A and then with $\rho_{B|A}$. The formula $\rho_A \cdot \rho_{B|A}$ says first act with $\rho_{B|A}$ and then with ρ_A. There is no assurance that the two will yield the same result; and in general they will not. Which is the correct order? It turns out that neither is correct if the resulting product is to be a new density operator. To make sure it will be a density operator, we split the operator ρ_A into a product of its square root, so that $\rho_A = \rho_A^{1/2} \cdot \rho_A^{1/2}$. Instead of multiplying $\rho_{B|A}$ by ρ_A, we multiply it from either side by $\rho_A^{1/2}$. The formula that results from both changes is the *definition* of the star operator:

$$\rho_{AB} = (\rho_A^{1/2} \otimes I_B)\, \rho_{B|A}\, (\rho_A^{1/2} \otimes I_B) = \rho_{B|A} \bigstar \rho_A \tag{23}$$

An inversion gives an explicit expression for $\rho_{B|A}$

$$\rho_{B|A} = (\rho_A^{-1/2} \otimes I_B)\, \rho_{AB}\, (\rho_A^{-1/2} \otimes I_B) = \rho_{AB} \bigstar \rho_A^{-1} \tag{24}$$

14. Analogous Inferences: Mutations and Electrons

In the case of mutations among children, we used the rule of total probability (5) in the series of computation (2), (3) and (4), to infer from the probabilities of various mutations in one child in the family to the corresponding probabilities for a second child. We can use these new quantum formulae to display the corresponding inference for pairs of electrons in the singlet state.

Take two electrons in the singlet state (13)/ (14) with projection operator P_{12}. Using (16) and (19), the reduced density operator representing each of the electrons individually is

$$\rho_1 = I_1/2 \qquad \rho_2 = I_2/2 \tag{25}$$

These are the quantum analogs of the probabilistic equation (1) of the mutation case:

$$P(\text{child}_1 \text{ carries } m_i) = r_i << 1 \quad i=1,\ldots,n$$

$$P(\text{child}_2 \text{ carries } m_i) = r_i << 1 \quad i=1,\ldots,n$$

A short calculation shows that[258]

$$\rho_{1|2} = 2\, P_{12} \tag{26}$$

The analog of the rule of total probability in Table 1 is

$$\rho_1 = Tr_2\,(P_{12}) = Tr_2\,(P_{1|2} \bigstar \rho_2)$$

Substituting for $P_{1|2}$ and ρ_2, we use this rule to infer from state ρ_2 of the second electron to that of the first ρ_1. We find

$$\rho_1 = Tr_2\,(P_{12}) = Tr_A\,(2P_{12} \bigstar I_2/2) = I_1/2 \tag{27}$$

in agreement with (25). This last computation (27) is the quantum analog of the application of the classical rule of total probability in (2), (3) and (4).

15. Disanalogies

These last comparisons underscore the analogies between a probabilistic inductive logic and the quantum inductive logic induced by the laws of quantum theory onto electrons in entangled states. That these analogies are present shows that the quantum logic is of comparable richness to the probabilistic logic. The key point for our purposes, however, is that the analogies are incomplete. The quantum inductive logic is a distinct inductive logic.

That the analogies are incomplete is already established by the investigation of the properties of density operators in Section 10. When the formal properties of the quantum inductive logic are explored, further disanalogies emerge. They derive from the fact that probabilities are numbers, whose products are insensitive to the order of multiplication, whereas density operators are sensitive to the order of multiplication. Switch that order and one may get a different result.

One consequence of the lack of commutativity of operators is the following disanalogy discussed in Leifer and Spekkens (2013, p. 33). The probability P(S&R) can be expanded as a simple product

$$P(S\&R) = P(S|R)\, P(R)$$

The rule is robust and holds if all the probabilities are themselves further conditionalized on another variable T:

---

[258] This follows directly from (24) once we note that $\rho_1^{1/2} = I_1/\sqrt{2}$ so that $\rho_1^{-1/2} = \sqrt{2}\, I_1$.

$$P(S\&R|T) = P(S|R\&T)\ P(R\&T)$$

This is extremely useful in probabilistic analysis since it means that we can collect all background information into some huge proposition T and then treat all probabilities conditionalized on T, P(.|T), as if they were unconditional probabilities P(.).

The first of these two formulae has a quantum analog

$$\rho_{AB} = \rho_{B|A} \bigstar \rho_A$$

However the second does not. That is, we do not in general have

$$\rho_{AB|C} \stackrel{?}{=} \rho_{B|AC} \bigstar \rho_{A|C}$$

This means that the rule for forming conditional states will differ according to whether or not we begin with a state that is itself already conditional.

16. Conclusion

The material theory of induction requires that the inductive logic applicable in some domain be dictated by the facts that prevail in that domain. In many domains, facts do warrant a probabilistic inductive logic. The prevalence of such domains has helped foster the misimpression that a probabilistic logic is the universal logic of induction.

The burden of this chapter has been to illustrate how a formally rich, alternative inductive logic can be warranted. The domain is that of entangled quantum mechanical particles. The inductive logic appropriate to them employs density operators where a probabilistic inductive logic employs probability measures. This new logic looks very different, initially, from a probabilistic logic. There is no single real valued measure of support that tells us which state is more or less well supported. Differences of support are expressed by density operators. In the most definite case of narrowest support, the density operator is a projection operator. It identifies a unique state as the true state. At the opposite extreme of the most distributed support, the maximally mixed density operator accords equal support to all states, just as does a uniform probability measure in the probabilistic case. The intermediate cases are captured by density operators between these extremes. For the case of a single electron, the range of cases and their properties are represented in readily interpretable form by the spheres of Section 12.

There are further structural analogies between the quantum inductive logic and a probabilistic logic, as described in Leifer and Spekkens (2013). That assures us that we do have a logic of comparable richness. Eventually, the analogies break down, for the two logics are different.

References

M. S. Leifer and Robert W. Spekkens (2013), "Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference," *Physical Review* A 88, 052130 (2013); preprint: "Formulating Quantum Theory as a Causally Neutral Theory of Bayesian Inference," arXiv:1107.5849 [quant-ph]

Nielsen, Michael A. and Chuang, Isaac L. (2010) *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press.

Penrose, Roger (2004) *The Road to Reality.* London: Jonathan Cape.

Epilog

The many chapters of this book all aim to sustain a single conclusion. Inductive inferences are not warranted by formal schemas or rules. They are warranted by background facts. Over the last few years, I have had the opportunity of presenting this thesis and arguments for it in various philosophical forums. The reactions to it have been varied. Some find the idea illuminating and even obvious, once it is made explicit. They are supportive and I am grateful for it. Others are more neutral, reacting with various forms of indifference or incomprehension. Some set aside the question of whether they are or are not convinced by the main claim; or whether there is some way that they could help the speaker advance the project. Rather they hold to the lamentable idea that, no matter what, the job of an audience in a philosophy talk is to try to trip up the speaker with some artful sophistry. Still others are, perhaps, not quite sure of precisely what I am proposing and arguing. But they are nonetheless sure that it is a Very Bad Thing that must be opposed and stopped.

For audiences in these last two categories, a common strategy is to pursue this line:

> "If every inductive inference is warranted by contingent facts,
> how do we know those warranting facts?"
> "By more inductive inferences, warranted by further warranting facts?"
> "Doesn't that mean that there's a regress problem?"
> "Aha—Gotcha!"

My response to them then and to you now is the same. Yes—they are right. There's something like a regress lurking about. It is something I should address. It is, very roughly, the analog of Hume's problem of induction, but now played out in the material theory.[259]

Hume's problem of induction is the classic exemplar of an intractable philosophical problem. While many solutions for it are offered in the philosophy literature, I do not think that there is any one solution that commands universal assent. To have a theory of inductive inference that does not also solve Hume's problem would put me in good company with all the other accounts of inductive inference. If failing to solve Hume's problem is sufficient to damn the material theory, then we must also damn all other accounts.

For the purposes of this book, I wish to stop with that last conclusion. My hope is that

[259] Hume's problem can be set up as a circularity or an infinite regress. Something like this second regress form is the one that threatens in the material theory.

readers will think about the issues I do raise and the arguments I do offer in this book. There is ample material here for readers to ponder, endorse and dispute. I hope that they will not let themselves be distracted by an easy critique afforded by Hume's problem. It is one that can be applied to all accounts of inductive inference and fails to connect with what is distinctive about the material approach.

It is precisely because I wish to avoid this distraction that I have not raised the issue of Hume's problem so far in this book. For I find it entirely adequate to say that, if the material theory fails to halt the regress of justifications of Hume's problem, then it fares no worse than all the other accounts. However, in closing, I alert readers that I do believe that the material theory is not derailed by a regress akin to Hume's problem. My reasons have already been summarized in a paper (Norton, 2014) and I have elaborations in preparation.

In short, I argue that Hume's problem is an artifact of the formal approach to inductive inference. There we warrant an inductive inference by an appeal to a rule; and we justify that rule by inferring inductively over its past usage using another rule; and so on indefinitely. We thereby trigger a fanciful regress of inferences rules applied to inference rules applied to inference rules… It is fanciful since it is nothing like what we see in real science. Attempts to implement even the first few steps of the regress lead us far from contexts in which reasonable judgments can be made. How do we apply rules of severe testing to vindicate the use of inferences to the best explanation when they are used to justify instances of enumerative induction?

In the material theory, we have something similar. An inductive inference is warranted by a fact; and we support that fact by an inductive inference warranted by another fact; and so on. As we trace out these connections, we find ourselves mapping out an increasingly tangled network of inferential pathways that can quite quickly span across much science. However this regress is not fanciful. Rather it is a mundane exploration of the connections among the facts that support our science. Curie's inference on the crystallographic form of radium chloride is justified by Haüy's principle that in turn is justified by inferences that draw on much of the physics and chemistry of the nineteenth century. It is complicated, but not fanciful.

So far all is well. Yet one may still wonder: must not all the pathways of this network terminate in something like the singular facts of brute experience? The totality of those singular facts cannot warrant any universal generalization. For, when all we have are singular facts, we can call up no warranting facts of general scope to support inductions from singular facts to generalities. Or so it might appear.

Here appearances are deceptive. This last failure requires as a tacit assumption that relations of inductive support are hierarchical, something like the courses of stones used to build a tower. Each course is supported solely by the course below it. Analogously, the propositions of science reside in layers, with lower layers closer to the singular facts of experience. An inductive

inference that starts with facts in one layer can only call upon warranting facts in that same layer or those below it.

This hierarchical assumption fails for science. Its relations of inductive support are not hierarchical, like the relations of structural support among courses of stones in a tower. They are interconnected in very many complex ways. Relations of inductive support are closer to the relations of structural support in complicated systems of arches and vaulted ceilings. Each stone in such a system is supported structurally both by those below it and those above it.

How can these structures come about? An arch cannot be built simply by piling up stones, layer by layer. Rather we must temporarily support stones higher up in the arch by scaffolding. As further stones are put in place, support for these higher stones shifts to the permanent security of other stones and the scaffolding can be removed.

It is the same in science. To get our inductive inferences started, we make various general hypotheses. These hypotheses are used to warrant inferences, even though they are themselves inductively unsupported at this initial stage. They are the analog of stones in arches supported by scaffolding. We must recall which these hypotheses are, for their use places an obligation on us. As our investigations proceed, we must return to them and give them proper support. When we do this fully, what results is an inductively self-supporting structure. Its simplest propositions will be singular but nonetheless they are able to support inductively other propositions of universal scope. When this process is complete, every proposition is well-supported inductively.

Here I have sketched my account so that readers see that my impudent boast of having evaded Hume's problem has a real basis. However I hope that readers can keep their interest and focus on the material in the many chapters preceding this epilog. There will, I promise, be ample opportunity elsewhere to dispute my solution of the regress problem in the material theory of induction.

References

Norton, John D. (2014) "A Material Dissolution of the Problem of Induction," *Synthese*. **191**, pp. 671-690.