**INSTRUCTIONAL NOTE**

It is my personal view that deduction and induction, i.e. deductive and inductive models and systems, are just two ways of modeling different aspects of a single reality: proper human reasoning. Inductive logic cannot be reduced to deductive logic. However, inductive logic can be formalized to contain deductive logic within it. If so, then deductive systems model the limits of proper reasoning (probability values of 0 and 1), the extremes at which thinkers are certain, while inductive systems model the areas between the limits (probability values between 0 and 1). To teach students how to properly think requires teaching them how to make use of the entire spectrum of human reasoning, not only to play at the limits.

However, most of the important developments in theories of induction and confirmation are presented by those who double as both philosophers and mathematicians, and most students are simply not capable of following the foundational research without devoting considerable time to it. This short primer is intended to remedy that by serving as an introductory supplement for Critical Thinking and Logic courses, without relying too heavily on mathematical formulations of the probability calculus.

This primer is the first part (Vol. I) of a larger project, hopefully to be completed in the near future. However, Vol. I and its anthology section are a complete unit on its own, and, when combined with a deductive logic text or primary critical thinking text, can take up an entire semester.

Here is how I piloted the early versions of this text in my Logic and Critical Thinking courses: students first learned the basics of *linear* reasoning--how to structure linear arguments (premise-conclusion form), and the basics of formal, symbolic reasoning (i.e. propositional logic). Once students became proficient in interpreting and structuring arguments, translating into symbols, and carrying out proofs by natural deduction, they were made aware of the limitations of what they have now mastered. (I also included truth tables, but some disagree with me about their utility for undergraduates.) This filled up a little over half a semester (8 weeks). Only the most important informal fallacies were covered.

It was then made clear to students that formal, deductive logic cannot serve as a model for the vast majority of human inquiry, inference and deliberation. We cannot understand human inquiry in purely deductive terms. (Instructors may do well here to introduce Hume's problem(s) of induction, and his examples of unsound predictive arguments that appeal to principles of regularity, as an illustration of the failure of deductive logic to justify future predictions.) Confirmation Theory was then introduced as the field working on this set of problems. After this, the primer was taught.

In short, I have used this primer as a way to teach students *non-linear* reasoning and argumentation, after having taught them *linear* reasoning and argumentation skills. But, the modular nature of this text allows for considerable flexibility--instructors may use it as they wish.

Vol. I covers, in plain language, the basics of a neo-Peircean model of inquiry. It first covers the nature of "theorization," defining the term "theory" for students. This leads to a discussion of the differences between *explanations* and *mere predictions*. Then, the notions of *weighing the evidence* and *inference to the best explanation* are surveyed. Finally, a summary section synthesizes these various elements into a coherent whole: the process of inquiry. Students are given a rough model of the process of human inquiry, a replacement for the overly simplistic models of the "scientific method" most students come across in their lower-level science curricula.[1]

[1] There are so many instances of poor science curricula that it is hard to cite anything in particular. One biology text I learned from explicitly identified hypothetico-deductive methodology as not only *essential* to science, but conflated the scientific method with the hypothetico-deductive method, as if they were synonymous. (Hickman, *Principles of Integrated Zoology,* Ch. 1.3.) This is not the only text to do so. This is not even an oversimplification,

It should be noted, however, that this book is *not* a "theoretically neutral" text. There is little consensus on the nature of philosophical probability, and many of the ideas covered here are extraordinarily controversial in the literature. In other words, this book offers a broad-brush strokes model of inquiry and human reasoning, and needs to be critically evaluated. There is simply no "standardized model" of inquiry or inductive inference, no "authoritative" system to offer students. It is an attempt to advance research in Philosophy of Science and Confirmation Theory by providing a synthesis of various schools. Still, I think this modified Peircean model, at the very least, is powerful, a useful way of dividing up the various elements of inquiry and synthesizing them into a whole. It provides a solid starting point from which students and researchers alike can investigate the particular aspects of inquiry they are most interested in. And yet I hope that, within my lifetime, someone will modify this model and make something much better.

Early versions of this book were tested in my Logic and Critical Thinking courses at Oklahoma State University during my time as an adjunct there (2020-2022). I was very pleased with the results. Students seemed highly engaged, and some of them even changed their majors or adopted a philosophy minor after reading it. Most importantly, when paired with simple examples from the history of science, this text helped demystify the nature of scientific inquiry for many students. Various students gave me input on the method of delivery, and have helped simplify the text greatly, although I am sure there is much more work to be done.

-Pierce Alexander Marks

---

but an outright falsity. Science does not proceed merely through deductive testing of hypotheses--deductive testing is not even *essential*. On the most charitable reading of this text, the authors simply understand "hypothetico-deductive" to mean "creating hypotheses and looking to test them." But even this would reveal their disconnect from the wider philosophical and historical literature on scientific methodology. The danger in this is that it may leave students under the impression that "real science" must be positivistic, able to be confirmed or disconfirmed by finding empirical data that flat out contradict or imply their hypotheses. But we are far past the age of positivism in science, and science has certainly been done for millenia in much more creative ways. The hypothetico-deductive method is, and always has been, just one tool in the scientist's toolkit.

# Table of Contents

How IBE and Bayesian Probability are Shot Through with Pragmatic Decisions

## Vol. II: Confirmation Theory and Inductive Logic

I. The Various Senses of 'Probability'
II. Some Rational Principles of Confirmation/Disconfirmation
III. A Simple Probability Calculus
IV. The Modeling Problem: What is the Probability Calculus a Model Of?
V. Some Paradoxes of Confirmation
VI. Assigning Prior Probabilities: The Major Issue

# Part 1: The Basics of Theoretical Reasoning

**Introduction**

Confirmation theory--a field that studies the rational principles for forming beliefs and adopting theories on the basis of evidence--is one of the most important subjects philosophers study. For most of our important beliefs are derived on the basis of evidence, not known by some faculty of intuition or miraculous experience. If we want to know the truth with confidence, or ensure that the greatest portion of our beliefs are successful, warranted and rational, we will thus need to understand the relationship between the acceptance of some proposition as true (belief) and the evidence that supports that acceptance.

Theories of explanation and prediction provide the wider context for confirmation theory. Before we ask, "how do I know that some theory is rational/successful/probable/correct?", I must first understand what a theory is. Theories of explanation and prediction are attempts to analyze the nature of theoretical reasoning--they are theories *about* theory in general. Theory includes both explanations and predictions. Explanations are attempts to answer questions like, "why did E happen?" "What makes E the way it is?", etc. Predictions are attempts to make rational,successful, reliable or probable guesses, when the truth is unknown. A prediction about what the past was like is sometimes called a 'retrodiction.'

Theories of explanation and prediction situate theories of confirmation within a wider topic. Likewise, theories of inquiry embed theories of explanation and prediction within an even broader framework. Inquiry is an activity, it is the activity of asking questions and coming up with successful answers. Very often, the goal or success of inquiry is thought to be *truth*, or something like or approximate to truth: a pursuit of the truth by asking questions, and seeking answers, with those answers in the form of explanations and predictions. But there may be other ways of conceiving of the aims and success of inquiry--we may wish to find the most useful answer, the most consistent answer, the most empirically successful answer, etc.

Inquiry is not done in an instant, but is a process that unfolds over time. Theories of inquiry thereby study the process of question-asking, the formation of theories, and the confirmation or adoption of theories on the basis of evidence. Our task here is to begin an inquiry into inquiry itself: inquiry into inquiry, theories about theory, explanations for explanation. We might call these "meta" theories.

We thus have three questions we want answers to: "what is inquiry, and what are its aims?" "what are theories?" and "how should we decide which theories to adopt, especially on the basis of evidence?" An answer to the first question will not only require, but shape, answers to the second and third. Theories of Inquiry contain Theories of Explanation and Prediction, which in turn contains Theories of Confirmation. (Additionally, Theories of Confirmation will contain an even narrower topic: how does statistics relate to rational belief formation? That is, how can statistical probabilities and ratios serve as evidence? Accounts of these we might call Theories of Probability, or Probability Theory.)

These meta-theories come in two flavors: descriptive and normative. A normative account of inquiry, for example, will specify the *proper*, *good*, *reliable* or 'right' way of inquiring. On the other hand, a descriptive account will merely seek to describe, in a value-neutral way, our actual practices and methods of inquiry. Descriptive accounts are typically the starting point, since we assume that we do inquire well, or at least somewhat successfully, before pursuing a meta-theory (if we didn't presume this, we couldn't even start inquiring into inquiry; we would be paralyzed!). The successes of our current methods will inform, but not necessarily determine, our *normative* theory--the processes and kinds of reasoning we *should* use.

Keep in mind, however, that inquiry can have different aims. Some philosophers have attempted to recast the inquiry, not as the pursuit of truth, but as the pursuit of a maximally useful answer, a maximally consistent answer, the most unified answer, etc.[2] Given each aim, the process may change--for the goal of any endeavor will determine the steps necessary to achieve it (aim determines method and the methodological process). Thus, radically different aims of inquiry may result in different models, or *paradigms* of inquiry, theory, explanation, prediction and success. The best we can do here is delineate one paradigm which appeals to some intuitive notion of truth or 'rational expectation,' of probability as a useful guide to truth or rational expectation, and the methods normally supposed to aid us in arriving at our goals.. Note that different paradigms or models of inquiry are not necessarily incompatible. They may have overlap, and one thinker may easily have different aims at different times, requiring different practices.

Theory construction and theory adoption lie at the foundations of every human discipline. All inquiry relies on theory at some point. The principles of evidence and acceptance do not just apply to ordinary beliefs, but also to theories constructed and put forward as "the truth" by scholars. Evidence and belief--including evidence and the adoption of any scientific or academic theory--are governed by general rational principles, and these principles should constrain both scholars and laymen equally. And so here Philosophers are in a unique position: our findings will undoubtedly affect how we understand ourselves as intellectual, believing creatures. Not only that, but Theories of Inquiry, Explanation, Prediction and Confirmation will shape our understanding of every field of study, from Physics to Literary Criticism.

Despite the importance of inquiry for human life and understanding, books introducing the topic have mostly emerged from the anglo-analytic philosophical tradition, and have been unnecessarily technical, boring, and symbolic.[3] Even seasoned philosophers have significant trouble picking up Popper's *The Logic of Scientific Discovery,* due to what is, in my view, its overreliance on mathematical formulas. Worse, once a student struggles through the overly technical literature, they often find that the key insights could have been communicated in less technical jargon. This has been a major drawback of the impressive and rigorous analytic tradition: many of the leading figures in early 20th century analytic philosophy have been extraordinarily intelligent, but have been far too embedded in mathematical topics. These geniuses often seem to lose sight of the fact that human beings are not primarily computational machines. As a result, most students, even students of philosophy, are never introduced to theories of inquiry, explanation, prediction, or confirmation. Confirmation Theory in particular, because it involves the development of symbolic systems of probability, has been locked away from the public.

---

[2] Pierre Duhem, in *To Save the Phenomena,* argues that, while Aristotle (384-322BCE) and the Aristotelian tradition attempted to give *true*, *accurate* theories (models), the philosophical/scientific traditions has, as far back as we can tell, allowed for a different aim. Taking astronomical theory as our prime example, Aristotle's aim for inquiry was truth, or at least approximate truth, while Ptolemy (born c. 100AD) conceived of the aim of inquiry as merely successful modeling--theories and models that could accurately predict the apparent motions of the heavenly bodies. This division between Scientific Realists and Scientific Non-Realists (which include a variety of thinkers with very different views) goes back to the early days of Philosophy and the natural sciences. (Duhem, Pierre. *To Save the Phenomena*.)

[3] "It is easy to get lost in the complexities and controversies in the philosophy of science about how this process works. Scientific discovery is a fascinating and timely topic for research, to be sure, but much light could be thrown on abduction by also analyzing some less technical cases familiar from everyday reasoning. One problem is that the philosophy of science, although very important in its own right, tends to favor examples that are so specific and technically controversial that they do not really serve well to pose basic questions that tend to be overlooked. The extensive scientific examples, although deeply interesting, do not illustrate abductive reasoning in such a compelling way that the reader can say, "Aha, now I know what it is." Thus some reconsideration of other examples could be useful." (Walton, Douglas. *Abductive Reasoning*, 12.)

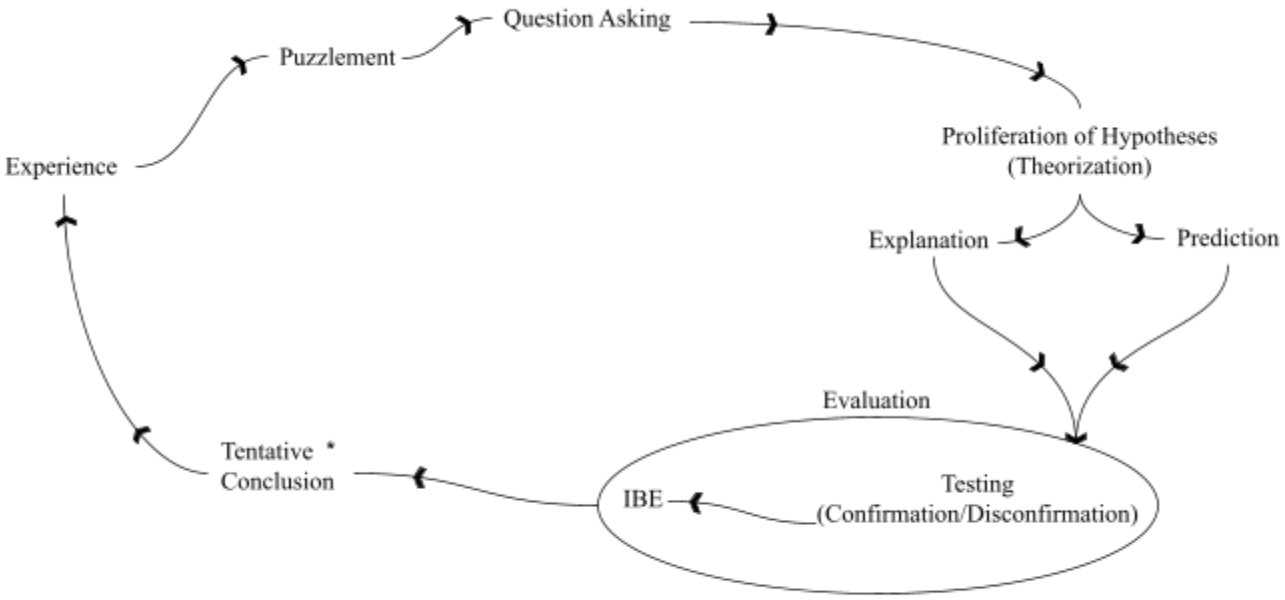
Here, I am attempting to bring confirmation theory to not only the masses, but to philosophers, like myself, who have not spent their formative years engaged in mathematical subjects. I take as my models C.S Peirce (1839-1914) and John Dewey (1859-1952). Peirce is, in many ways, a model of a well-rounded scholar. Despite being a polymath capable of rigorous symbolic and mathematical reasoning, Peirce's writings are accessible to students of the humanities, and even interested laymen. Despite this accessibility, Peirce is the father of modern confirmation theory: his writings eventually led to the development of key concepts of confirmation, namely Inference to the Best Explanation (IBE). Carl Hempel (1905-1997), who wrote perhaps the most influential works on confirmation and explanation in the 20th century, notes that a rudimentary form of IBE is clearly presented in Dewey's manual for thinking, *How We Think* (published 1910, revised in 1933)*;* Hempel, in constructing his models of explanation, begins with Dewey's observations about our reasoning processes.[4] Dewey himself was educated in logic and the nature of inquiry by Peirce at Johns Hopkins University, although he was put off by what he took to be an overly-mathematical approach to logic.[5] Dewey wrote *How We Think* as a manual for teachers covering the basics of inquiry without being overly mathematical. Peirce's work on inquiry, trickling down through Dewey and now accessible through his unpublished writings, is so central to confirmation that his notion of "abduction" is often, though wrongly, used as synonymous with IBE.[6]

**Summative Preview**

This part of our book is intended to describe the process of inquiry by providing a conceptual model of the process as a whole. In addition, we provide models and analyses for some of the components of this process: explanation, prediction and Inference to the Best Explanation. The model of inquiry provided here can be summed up in the following diagram.

[4] Hempel, Carl. "Explanation in Science and History."
[5] Dykhuizen, George. *The Life and Mind of John Dewey,* Ch. 2
[6] (Frankfurt, Harry. "Peirce's Notion of Abduction.") (Mcauliffe, William H. B. "How did Abduction Get Confused with Inference to the Best Explanation?")

**The Process of Inquiry**




** Hypothesis H is the best explanation/prediction out of its current competitors, given our current body of evidence.*

A brief overview is necessary before approaching this subject matter from the 'bottom-up,' as we do below. First, all inquiry or *investigation* begins with experience. Second, some experiences grab our attention and puzzle us. Third, we begin to ask questions, like 'what?' 'why?' 'how?' and 'what is it?' about the puzzling experience (or the objects we encounter in experience). Fourth, we begin to formulate answers to those questions: theories, also called hypotheses.

These hypotheses can take the form of explanations, which are stories or sets of posited, conjectural facts that 'shed light on' the puzzling event by describing potential 'reasons why' the puzzling event obtained. The entities, events, conditions and supposed facts posited in an explanation are called 'posits.' Hypotheses can also take the form of predictions, which are simply attempts to establish some conclusion as more probable than not.

Fifth, after we formulate various hypotheses, we move to evaluate those hypotheses, and find the 'best' (i.e. 'true,' 'correct' or 'most successful') of them. We gather relevant data and evaluate evidence by testing the implications and predictions of those hypotheses. Then, we weigh all the 'pros and cons' of each hypothesis against each other, and tentatively conclude in favor of one hypothesis over another. However, our job is not done: because our body of data (evidence) and the number of hypotheses are constantly changing as time goes on, we must constantly reevaluate our conclusion--that is why the conclusion is *tentative.* And thus, the process begins again.

This, we suggest, is the process of inquiry universal to all human investigations, across every field. However, it is an idealized model, and real-life investigation will be much messier than presented here. Still, all the steps will be present in any successful inquiry, and will bear the same logical, if not temporal, order. Let us now move towards understanding the steps of this process in more detail.

**A Note on 'Truth'**

In what follows, I will sometimes make use of the word "truth." I don't mean to presuppose any particular theory of truth. "Truth," in this primer, is to be taken merely as "success." There are many ways to conceive of success, and so it may be best to replace the word 'truth' with something like 'rationally correct,' 'rationally proper,' 'most reasonable,' or 'empirically adequate,' etc. "Truth," here, is merely a convenient shorthand.

Our goal is to introduce, in fairly broad brushstrokes, some rules of probability, rational investigation, and acceptance of theories. Statements about noetic probabilities can be understood as statements regarding what is rational to believe, accept, or affirm, and with what degree of confidence. We might even define noetic probabilities as quantified measurements of our normative reasons to act, with the action in particular being something like affirmation, acceptance, belief, expectation, or 'act as if it will happen.' Does this notion require that of 'truth?' If so, truth in what sense? There are a variety of theories of truth, after all. Many will say that probability and truth are bound up together, such that "H is probable" just means "H is probably true." I do not know the answer to these questions. Yet, it seems perfectly clear to me that the ways of reasoning laid out in this book are correct, rational principles of reasoning, explaining, theorizing, and investigating. We have to start somewhere.

In fact, how could we develop a *theory* of truth, without first being able to understand how to *theorize*? So I don't think the question, "what is truth, and how does it relate to probability?" should bother us here. It is very interesting, and important. But it cannot undermine our project, since answering that question will involve accepting some method of inquiry like what I lay out below.[7] A blinded horse might not understand who is driving her stagecoach, nor where they are headed, and yet still reliably arrive at her destination.

[7] Goodman?

## Section I: Theory, Explanation, and Prediction

The primary stumbling block in front of students of any discipline is confusion over method. How do, say, philosophers come up with answers to their questions? How do they develop theories, and confirm that they are true (or verify that they are false)? How do scientists? What about historians, anthropologists, and the like? Typically, students are not given answers to these sorts of questions, but learn through experience: they take introductory courses, and "get a taste" for how scholars reason. Some catch on, some are turned off, others remain puzzled. Students become researchers, researchers become professors: the danger is that the lack of methodological training results in sloppy scholarship and further propagates lack of understanding to the next generation of students. The result is that students and professionals alike are working without really understanding what they are doing.

So let us begin by making clear two foundational notions of inquiry--explanation and prediction--and by introducing the various topics and problems related to them, in the hopes that giving students a clearer picture of how philosophical inquiry works will save them from the evils of sloppy thinking.

To be clear, explanation and prediction only make up part of the human reasoning toolkit. There are straightforward arguments or inferences, of the form:

Premise 1
Premise 2
...
Premise *n*
---
∴ *Conclusion*

Each premise is evidence, or support, for the conclusion, and, when taken together, are meant to warrant the conclusion. The line between premises and conclusion indicates an 'inference,' saying, roughly, "we have inferred what is below the line from what is above the line." We make arguments like this all the time. When the conclusion is intended to follow from the premises with logical necessity, so that if the premises were true, then the conclusion would *have* to be true, the argument is *deductive*. To understand the nature of straight-forward, *deductive* argumentation, students should take a course in Symbolic Logic. When the conclusion is intended to follow from the premises with mere probability, so that if the premises were true, then the conclusion would be more probable than the alternative, the argument is *inductive.* Much of what we cover in Part III of this book involves developing a logical system for evaluating and symbolizing inductive arguments. Some universities, especially well-funded ones, will offer a course in inductive logic which deals with these sorts of arguments.

Theorization, however, cannot be reduced to mere straightforward argumentation, and the issue of understanding theory is much more complex, and there are not many books that concisely introduce the notion of theory. This is part of our goal below.

**On Theory**:

Theory, or theorization, is something like an attempt to discover or describe realities that aren't directly observable or verifiable. The past is not directly observable, neither is the future. Some things in the present aren't directly observable either--like ethical truths, the existence of God, the nature of the soul, logical principles, or even material things so small we can't see them.

Can we know anything about reality if it's not directly, empirically observable? Most Philosophers have thought we can, and have used theoretical reasoning, including *explanatory* and *predictive* reasoning, to do so. Theoretical reasoning has been around as long as inquisitive people have asked questions, and, overall, its methods have largely remained steady.[8]

Before delving into the details of explanation and prediction, let me describe them roughly. "Theories," also called "hypotheses," can come in two kinds: explanatory theories and predictive theories. An explanatory theory tries to *explain*, and a predictive theory tries to *predict.* A strangely technical weatherman might say, "I have a hypothesis: tomorrow, it will rain." This is a predictive theory. On the other hand, a detective investigating a murder might say something like, "my theory is that Jeeves, the butler, was responsible for killing Mr. Gasbottom." This is an explanatory theory. A predictive theory is merely trying to make a correct, probable guess, while an explanatory theory is trying to show who, or what, was responsible for some event. Still, very often explanatory theories also predict, although in a loose sense of 'predict.' For the detective is, in a sense, trying to predict *what really happened*, who really killed Mr. Gasbottom. Explanations, or explanatory theories, are attempts to answer questions like "why did this happen?" "What made this happen?", "What even is this thing?," etc. So, prediction and explanation overlap.

**Explanation:**

Humans use *explanatory reasoning* to reconstruct the past and present, and *theoretical prediction* to attempt to give accurate predictions of the future. Both of these utilize highly similar sorts of reasoning structures, and differ primarily in their aims: explanatory reasoning constructs explanations, while predictions attempt to establish some set of propositions as probable (without necessarily trying to *explain* anything).

An explanation is a story that makes sense of something puzzling. It is an attempt to construct a story, a model, or a series of hypothetical events and conditions that, if true, would answer questions like "How did E happen? Why did E happen? What caused E to happen? What *even is* E? What makes E what it is, and not something else?" Here is an example of explanatory reasoning that many of us might come across in day to day life:

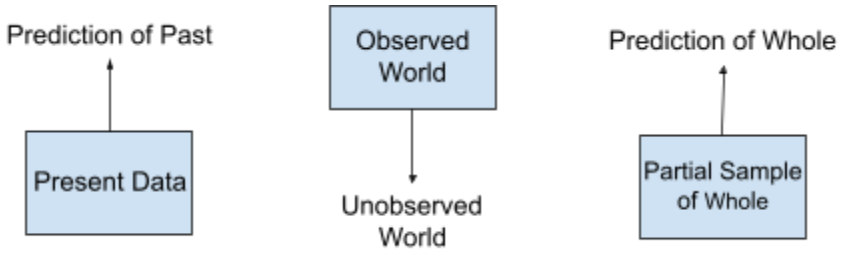
> *We turn the key in our car's ignition, and nothing happens. The car doesn't start, and doesn't make any noise. We ask ourselves, "why isn't my car starting?" Immediately several possibilities come to mind. First, we wonder if the car's battery is dead. If the car's battery was dead, then that would make sense of why the car won't start. Second, we wonder if the car's battery has been disconnected. If the car's battery was disconnected, then we know the car wouldn't have enough electricity to start, and that would make sense of why the car won't start. Third, we wonder if rats have gotten into the car and chewed through the wires connecting the ignition switch to the ignition cylinder. If rats had chewed through these wires, we know the car would not be able to start. With these three possibilities in mind, we pop the hood, and begin trying to figure out which possible story is correct.*

Each of these three possible stories is a different *explanation, theory* or *hypothesis* about some puzzling data: our car won't start, and doesn't make any noise when we turn the key. In ordinary language, each of these different stories would *explain* the fact that our car is not starting. Other ordinary ways of talking

[8] For a historical survey of theoretical reasoning through time, see: (Duhem, Pierre. *To Save the Phenomena.)*

about explanation might be to say that each of these stories would *make sense of, shed light on,* or *lead us to expect* the fact that our car does not start.

Explanations like these are stories with the power to answer "why?" or "how?" or even "what?" questions. They are most often attempts to reconstruct the past to make sense of the present, as in the example above. That is, each explanation is an attempt to reconstruct what might have happened in the past that caused, or brought about, the puzzling event we just observed (this is called an *etiology*, which means "causal history"). In a sense, we are trying to predict *how the past was, and how it led up to the present event,* even though we cannot directly observe it, nor its link to the present. We can also predict the future on the basis of the past, or predict the unobserved from the observed. Here are some common kinds of explanation diagrammed below;the box represents our vantage/starting point, and the arrow represents the act of predicting or explaining.




In each case, we infer from some starting data to some theory, hypothesis, explanation, prediction, conclusion, etc.

In the case of the car not starting, we can say that we have three theories, hypotheses or explanations, and label them H1-H3 (the "H" standing for "hypothesis"):

> *H1: The car battery is dead. And if a car's battery is dead, then the car will not have enough electricity to strongly turn over the starter. Most times (say 7/10 times) that a car does not have enough electricity to strongly turn over the starter, turning the key in the ignition does not result in any noise at all.*
>
> *H2: The car battery is disconnected. And if a car's battery is disconnected, then the car will not have enough electricity to turn over the starter at all. If a car does not have enough electricity to turn over the starter at all then turning the key would certainly not result in any noise.*
>
> *H3: Rats have chewed through the cables connecting the ignition switch (the part where you insert the key) to the ignition cylinder (the part that makes the starter attempt to turn over). And if a car's ignition cables were chewed through, then the electricity from the car battery would not be able to flow to the ignition cylinder. And if electricity from a car's battery could not flow to the ignition cylinder, then the car would not start, and turning the key would not result in any noise.*

Each of these stories have something in common: if they were true, they would all imply or make it likely that the car won't start. This is at the core of explanatory reasoning: explanations have logical or probabilistic relationships to the things we want to explain. Explanatory stories are stories that, if true, would make the thing we want to explain true, likely true, or more likely true than they are by themselves. Think of the logical relationship this way: if we were in the past, before we turned the key in the ignition,

and we knew that the story in H1 was true, that would lead us to *expect* (predict, or not be surprised by) the fact that, when we put the key in the car and turn it, nothing happens. If we were in the past, and knew that our battery was dead, we would be able to predict, or expect, that turning the key would not work. If we were in the past, and knew that rats had chewed through the wires, we would not be surprised to find that turning the key over does nothing. If only we had known that our battery was disconnected, we would not have been surprised that the car wouldn't start!

**Explanatory Power**

Let's label our puzzling data/event we want to explain "E," for "event."

E: Car won't start; turning the key doesn't produce noise.

We can diagram the relationship between the different hypotheses (H1-H3) and the puzzling event (E) like this...

$$H1 \Rightarrow E$$
$$H2 \Rightarrow E$$
$$H3 \Rightarrow E$$

...where the arrow represents "explains." This double-lined arrow represents the relationship between some story (some set of hypothetical, posited facts) and the puzzling event. This relationship is the "explanatory" relationship. It is a complex question exactly what that arrow means and what that relationships consists in--what it means for "H to explain E"--but we can, by example, grasp it intuitively, and build up from there.

In older textbooks, you'll often find the various parts of the explanatory story called "explanans" and the thing to be explained the "explanandum." And these books will often diagram the most vague and general form of an explanation like this:

$$\text{Explanans} \Rightarrow \text{Explanandum}$$

Again, the idea is that we can invent stories or point to facts we know in order to answer questions about what makes something happen, or why things are the way they are, etc. This is what it means to *explain* something. We can break apart each hypothesis (story) into individual details, and give it a structural form. These individual details--parts of the story--are called *explanans.* And the events, conditions, entities and 'facts' in them, especially when they are purely hypothetical guesses, are called *posits.* We *posit* certain conditions, facts, entities as part of these stories, and these *posits* and the conditions they find themselves in are meant to explain the explanandum. The rats in H3, or the battery being disconnected in H2, are both *posits.* Let's give some structure to H1 and H2:*

H1

Explanans1. The car's battery is dead.

Explanans2. If a car's battery dies, then the car would not have enough electricity to strongly turn over the starter.

Explanans3. Most times (say 7/10 times) that a car does not have enough electricity to strongly turn over the starter, turning the key in the ignition does not result in any noise.

⇓

*Explanandum: The car does not start, and turning the key doesn't result in any noise.*

H2

Explanans1. The car battery is dead.

Explanans2. If a car's battery dies, then the car will not have enough electricity to turn the starter over at all.

Explanans3. If a car does not have enough electricity to turn the starter over at all, then turning the key in the ignition will certainly not result in any noise.

⇓

*Explanandum: The car does not start, and turning the key doesn't result in any noise.*

Look at H1. Do you see how, if all the explanans were true, then the explanandum--the puzzling event--would be probable, or likely? That is, if we know that all the explanans of H1 were true *before* turning the key, we would have been able to predict that, probably, the car will not start, and turning the key will probably not make any noise. We wouldn't be surprised when it ended up happening, because all the explanans *make the explanandum more probable than it was before,* and in this case (but not always), *probable overall.* There are no guarantees with H1, but H1 would give us very strong reasons to expect the explanandum.

Now look at H2. If all the explanans were true, then the explanandum would be not just probable, but *guaranteed*. If we knew for certain that the explanans of H2 were true *before* turning the key, we would have been able to predict, with certainty, that the car would not start, and that the turning of the key would not result in any noise.

Part of the relationship between explanans and explanandum--between theory and the puzzling thing we want a theory about--is called *explanatory power.* This concept, also called 'explanatoriness,'[9] is normally spoken of vaguely. On one usage of 'explanatory power,' an explanation has some explanatory power just in case that explanation, if the story it tells was true, would have made the puzzling event *somewhat more probable.* This does not necessarily mean that the hypothesis makes the explanandum *probable overall.* It just means that H, if it was true, would increase the probability of E, even just slightly. In another sense, a potential explanation is powerful in that it specifies some of the reasons, or kinds of reasons, that contributed to the puzzling event occurring. As we will see below, these two senses of 'explanatory power' are the two kinds of *explanatory relevance*. Again, philosophers do not speak univocally about explanatory power, but freely use the term to refer to either statistical or reasons relevance, which we will discuss shortly.

Explanatory power in the first sense comes in degrees. Look at H1 again. H1 does not guarantee, with 100% confidence, that, when we turn the key, no noise will happen--in 3/10 cases, there has still been some noise from the starter. Let's say that we have a 70% probability that turning the key won't result in noise (because 7/10 cars with a dead battery have not made any noise in the past). This is weaker than, or *has less explanatory power* than H2, because H2 would guarantee, with 100% probability, that the car will not start and turning the key will not result in any noise. Think closely about H2--if all the

9

explanans in H2 were true, there would be no way that the explanandum could be false. So, explanatory power comes in degrees: the degree to which an explanation is powerful is the degree to which the explanans (the explanatory story) would make the conclusion probable.

Note, however, that philosophers often use terminology loosely, and, in different areas at different times, philosophers might use the same word to evoke multiple concepts. 'Explanatory power' is one such word. Keep in mind that some philosophers use 'explanatory power' in different ways than I am. Some mean by 'explanatory power' that a hypothesis increases the probability of the data, *and* that it specifies the *causes* (or reasons why) that data obtains. Others use it as I am here, to merely refer to probability relations. And finally, I have encountered others who use it to mean only that the hypothesis specifies, or could potentially specify, the causes of the data it predicts. We must allow ourselves to have some flexibility in vocabulary.

**On Probability**:

The exact nature of probability is even more complex than the nature of explanation, and less amenable to understanding via example. The study of the nature of probability would have to include the following: (a) an analysis of the meaning of "probability" (a theory of probability); (b) the discovery of a set of normative principles governing the use of probabilities and an analysis of how to apply these principles for adopting theories and beliefs as true (confirmation theory); and (c) a symbolic system to represent all these things and to facilitate precise applications of probabilistic thinking (probability calculus). It is thus far beyond us, at this stage, to address anything more than the following three topics, and in a very common sense, pre-theoretical way.

First, **there are many kinds, or senses, of 'probability.'** Very often, we use the word 'probability' to indicate something like degrees of confidence: degrees of how confidently we *should* or *do* hold some belief, or make some prediction. Other times, we use the word 'probability' to talk about ratios of past observations, or kinds of events. And still, at other times, we use the word 'probability' to talk about the physical world, or an object's propensity to behave in certain ways.

The first kind mentioned above are *noetic probabilities*, also called *epistemic probabilities,* and have to do with how confidently we should believe some proposition P, or expect some event E to occur.[10] *Statistical probabilities* have to do merely with ratios derived from past or present observations, or ratios projected into the future. *Physical probabilities* have to do with features of the physical world--we might say that some object, in certain conditions, has a 70% propensity of behaving in a certain way, or causing a certain event: if it is put in a certain situation 10 times, it will behave in this way 7 times.

It is an interesting and complex question as to how exactly these probabilities relate to one another, and far beyond us at this stage to theorize about the precise nature(s) of probability.

Second, **probability is often spoken of as coming in degrees.** Regardless of our theory of probability, probability has a lower limit, 0% (or 0), and an upper limit, 100% (or 1).

Talking naively about noetic probability: if some proposition or event E is 0% probable (maximally improbable), that means it is *certainly* false or will *certainly* not obtain. If E is 100% probable (maximally probable), that means it is *certainly* true, or will *certainly* obtain. If E is 50% probable, that means it is just as probable as it is improbable, and so have no reason to believe that E will happen or not happen. We often talk about things we are totally unsure of as being "50/50." When we say, "I have no

[10] *Noetic* just means "relevant to the mind," and I am using it to mean "relevant to beliefs." Epistemic means "regarding knowledge or justification," which also pertains to beliefs.

idea, I could go either way, I think it's 50/50," we are saying that the (noetic) probability of whatever we are talking about is 50%--we are truly agnostic, and feel it rational to withhold judgement, from it.

With statistical probabilities, 0% means that we have observed E 0 out of some number *n* times. 100% would mean that we have observed E every time we have looked, so 10/10, 100/100, or simply *n*/*n* times. So, our ordinary, common sense way of talking about probability is to conceive of it as ranging from 0%-100% (from 0-1).

These two kinds of probabilities--noetic and statistical--must be related. If, on tossing a coin, there is a 50% statistical probability of it landing tails, then it seems to be the case that it is rational to think that the coin toss could go either way--to assign an epistemic probability of 50% to getting tails. So, clearly, these two kinds of probability are related. But how this is so is too complex to deal with here, much less the relation between statistical, noetic and physical probability.[11]

Third, **probabilities are spoken of as conditional.** We say things like, "given the fact that the battery is dead, the car probably won't start." The idea of conditional probabilities is just this: certain propositions or events can affect the probability of other events. If my car battery is dead, that *raises* the probability that the car won't start. If I drink spoiled milk, that *lowers* the probability that I will be able to go dancing later tonight. So, there can be *conditional probability relationships* between events and propositions. The phrase "the probability of H given E is *n%*" means that E makes the probability of H *n*%--it is a way of talking about conditional probability relationships. Sometimes, conditional probabilities are called *extrinsic* probabilities--the probability of H given *external* factors, factors outside of H. Some philosophers also call probability of this sort *posterior* probability--probability of H *after* (posterior to) considering some evidence.

Fourth, **probabilities are often spoken of as intrinsic and "prior."** We often talk of probability as *intrinsic*, or regardless of evidence. And what a surprise!... this is also a complex topic in its own right, and so my comments here should be taken as extraordinarily rough. This sort of probability is typically brought up when comparing two hypotheses. We often say that "the simplest explanation is probably the right one," or things to that extent. One idea is that, when we're comparing two hypotheses, the one with fewer convoluted assumptions is more probable than the other. But the number of assumptions in some hypothesis has nothing to do with evidence, or other events. The simplicity of a theory has to do with the theory itself, its internal or intrinsic features. These intrinsic features like simplicity can increase or decrease the probability of some hypothesis, at least when compared to others. Hence, *intrinsic probabilities,* also called *prior* probabilities: the probability of H *before* (prior to) considering some evidence.

**Explanatory Relevance:**

Some philosophers think that an explanation must have a great deal of explanatory power--it must make the explanandum *more probable than not* (greater than 50% likely).[12] However, Wesley Salmon has shown very persuasively that explanations do not absolutely require that they make the conclusion >50% probable.[13] Rather, for Salmon, all an explanation must do is be *relevant* to the explanandum. Consider the following example. Imagine we are playing heads or tails with a crooked penny. The bend in the penny

[11] For a very concise overview of the issues in probability theory regarding the kinds of probability there are, see: (Swinburne, Richard. "Introduction," in *Bayes Theorem*.)

[12] (Hempel, Carl G. and Paul Oppenheim, "Studies in the Logic of Explanation.") But this goes back at least to Peirce in 1903; his principle of abduction seems to limit explanations to hypotheses which deductively entail the conclusion. (Peirce, C.S. "Pragmatism as the Logic of Abduction.")

[13] Salmon, Wesley. *Statistical Explanation and Statistical Relevance*, Ch. 2.

results in the penny having a severe bias towards heads. However, there is still a small possibility, due to the shape of the penny, that it lands tails. We throw the penny and to our surprise, it lands on tails. We then construct the following explanation:

Explanans1 *This* penny has a bias towards heads--90% of the time it has been thrown in the past, it landed heads, and 10% of the time it landed tails.
Explanans2 *This* penny is thrown.

⇓

*Explanandum: The penny landed tails.*

This explanation seems a perfectly legitimate explanation. Yet notice how the explanans (the parts of the hypothesis) have not made the explanandum probable overall. In fact, the hypothesis actually shows that the explanandum was *very improbable*--only a 10% chance! Yet, it is still a perfectly good explanation, because it specifies some of the *reasons why* the penny landed tails--its shape was such that, when thrown, it *could* land tails. Further, the penny would not have landed tails without being thrown--the *intrinsic probability* of the penny landing tails is 0% *all by itself*. The intrinsic probability of some event is how likely that event is in its own right (without considering any other data). In this case, pennies do not "land on tails" without being thrown. So, the hypothesis does increase the probability of the penny landing on tails--it increases it from 0% to 10%. Call this ability to increase the probability of the explanandum "statistical relevance" or "explanatory power."[14] And call the ability of the hypothesis to specify reasons why the explanandum occurs "reasons relevance." (Note that I am using the term "statistical relevance" broadly--it could refer to either noetic or statistical probabilities.)

Further, if a theory does not have these kinds of relevance, then it couldn't be a correct explanation of the explanandum. For if a theory does not have any explanatory power--if it does not to any degree increase the probability that the explanandum will obtain over its intrinsic likelihood--then that theory is ruled out *as an explanation for the explanandum.* That is, a theory with no statistical relevance cannot be the correct explanation of the explanandum. Though the claims of that theory may be true, they are irrelevant to the explanandum, and so cannot be the proper explanation of why the explanandum occurred. The entire point of an explanation H is to point out the reason why E happened--a necessary ingredient for an explanation *being* a real explanation is that it has explanatory power/statistical relevance. If a friend asked us the reason why our car won't start, and we told them that it was because 5,000 years ago a pebble was swallowed by some animal, they would look at us like we're crazy.

Why does this strange fact not count as an explanation? Again, it seems to be because the strange fact about the pebble is *irrelevant* to the explanandum, and so does not count as an explanation of it. It neither increases the probability that our car won't start (statistical relevance), nor does it specify any reasons why our car won't start (reasons relevance).Wesley Salmon, writing in 1971, gives an extraordinarily clear expression of the need for explanatory relevance.[15] Consider the hypothesis:

> *John Jones avoided becoming pregnant during the past year, for he has taken his wife's birth control pills regularly, and every man who regularly takes birth control pills avoids pregnancy.*

[14] McGrew, T., 2003. "Confirmation, Heuristics, and Explanatory Reasoning," 558.
[15] (Salmon, Wesley. "Statistical Explanation," pg. 34-35.)

From the perspective of anyone who knows the basics of human reproduction, this explanation is hardly an explanation at all. Why? For two reasons: first, the birth control pills are *causally irrelevant* to the fact that John Jones did not get pregnant. Birth control pills have no effect on the ability of a biological male to become pregnant. It plays no causal role. Second, the birth control pills are *statistically irrelevant* to the fact that John Jones did not get pregnant. His taking the pills does not, in reality, make any difference to the probability that he would become pregnant (because it was already a certainty, without the pills).

A hypothesis must describe the *reasons why* the explanandum obtains (happens/is true). In regards to physical events, reasons why are typically *causal reasons*--an explanation of physical events will specify the causal mechanisms, processes and physical laws that led up to the event. But there may be other sorts of reasons, and so other ways in which explanations can be relevant, besides causal reasons. For instance, in metaethics, we seek explanations of moral truths. Why is murder wrong, and care good? It seems implausible that a relevant explanation for these ethical truths will involve causal reasons. Similarly with mathematical and logical truths--we look for explanations of logical truths, but these do not seem to be causal explanations, and causality seems an irrelevant category. Further, consider what it means for a material compound entity to exist as itself: it is matter in a certain formal configuration. The existence of the compound entity can in some sense be explained by pointing out what constitutes the being of the compound: matter + form. We could call the form a *constitutive* reason for why the entity exists as itself. All of these are potential *reasons why* E obtains. But just what are "reasons why?" An analysis of "reasons" is one of the most troubling and fundamental problems of philosophy, one that lays at the foundations of all human reasoning, in any discipline or subdiscipline. Given the complexity involved in this topic, I can do no more than to gesture at the problem, and move on.

For now, let us say only that an explanation must have both *statistical* and *reasons* relevance.[16] And let us also note that statistical and reasons relevance are bound up together. For if something has reasons relevance, especially when those reasons are causal reasons, it will also likely have statistical relevance. For instance, if car batteries being disconnected are true *causal reasons* why cars do not start, then a car battery being disconnected will raise the probability that the car will not start (reasons relevance sometimes implies statistical relevance). Further, statistical relevance can itself indicate or aid in establishing a causal relationship. Consider a hygienist who is trying to discover the cause of some sickness in a hospital.[17] The hygienist has a guess that the sanitary practices of the hospital staff are causally relevant--the staff doesn't wash their hands before examining patients! To test her hypothesis, the hygienist forces all staff to regularly disinfect their hands before exams. And, what do you know, suddenly the mortality rate of patients at this hospital plummets! That is, the hygiene practices seem to have probabilistic, statistical relevance to the mortality rate. And this is ordinarily taken to indicate causal, or reasons, relevance.[18]

All this leads us to our first principle of inquiry:

> **Principle 1:** A hypothesis H, in order to count as an actual, correct/successful explanation of explanandum E at all, must have some degree of explanatory power (statistical relevance) to E, *and* must additionally have reasons-relevance to E. That is, H must, to some degree, increase the probability of E above its intrinsic probability; and, H must, in some sense, specify a genuine

[16] Note that, if there are not other kinds of reasons why besides causal reasons, that this would destroy the ability of philosophers to explain ethical truths. The entire branch of metaethics would be crippled.

[17]

[18] Again, see the case study discussed in: (Lipton, Peter. *Inference to the Best Explanation,* 74).

> *reason why* E obtained.[19] The correct explanation must be both statistically-relevant and reasons-relevant.

Given the concept of reasons relevance, we can also differentiate between two broad classes of explanation: a hypothesis H counts as a *full explanation* of E just in case it is an explanation that specifies *all the reasons why* E occurred. While H is merely a *partial explanation* of E just in case it is an explanation that specifies just *some* of the reasons why E occurred. Most of our explanations are partial because of our finite, limited perspective. If we wanted to explain why some event in the material world occurs, a full explanation may require following a causal chain back to the beginning of the universe… a task that is impossible for us.[20]

**The Structure(s) of Explanation:**

Our intuitive, common-sense introduction to the nature of explanation has proceeded mostly by example. From these examples of hypotheses, theories and explanations, we can extrapolate a useful way of writing out the structure of explanation that is loosely inspired by Carl Hempel's work.

According to Hempel, whose work was highly influential throughout the 20th century, theories can come in three forms, or can be modeled in three ways. Each of these ways is differentiated by their structure and degree of explanatory power.[21]

Our first two models are deductive models, in that the hypothesis H leads us to predict the explanandum E with complete certainty, due to H logically entailing E. The explanans of H entail the explanandum E by being paired with a general "law"[22] or a relevant generalization. When H makes use of a general law, it is called a "Deductive Nomological" explanation (D-N) ("nomological" just means "pertaining to law"). When H makes use of a mere generalization, it is called a "Deductive-Statistical" explanation (D-S). Both of these kinds of explanation have the same form, or logical structure, and differ only in their contents: one posits a law, another merely posits a generalization. So here, I lump them together under one header.

In the forms of explanation below, you will see a line, called the "line of inference." This line indicates that anything below it has been derived (evidenced, supported) from the propositions above it. You will also see this symbol: ∴ . This symbol simply marks the conclusion of an argument. I make use of these because, in each of Hempel's models, explanans are taken as giving hypothetical evidence for the explanandum, so that we are led to expect the explanandum as true. However, as I will show shortly, this is too strict of a requirement for explanation.

---

[19] A similar, though weaker, principle was noted much earlier (in 1903) by C.S. Peirce. Peirce asserts that a hypothesis, in order to even count as a hypothesis, must be abductively successful, where "abductively successful" means that the hypothesis deductively entails the explanandum. I do not agree with Peirce that H must entail E, but the point remains: H, in order to count as an explanation at all, must have power to predict E (even if we disagree about what sort of power that is--probabilistic or deductive). (Peirce, C.S. "Pragmatism as the Logic of Abduction," pg. 234). Wesley Salmon gives the stronger version of this principle that I write here (Salmon, Wesley. *Statistical Explanation and Statistical Relevance,* Ch. 3, pg. 36).

[20] Peter Lipton discusses as a case study the work of Ignaz Semmelweis, who was attempting to explain differences in maternity mortality rates between two hospital divisions. Though Lipton does not use the same terms I or Salmon do (relevance), he does note that Semmelweis only had success in explaining the disparity in mortality rates once he had discovered some difference between the two divisions that (a) seemed to impact the probability of mortality (statistical relevance), and (b) was a likely causal contributor to mortality (reasons relevance). (Lipton, Peter. *Inference to the Best Explanation,* 74).

[21] Hempel, Carl. "Explanation in Science and History."

[22] Hempel, Carl G. and Paul Oppenheim, "Studies in the Logic of Explanation"

**Deductive Models (D-N or D-S)**

Explanans1. (Particular Conditions) Conditions of kind C1...Cn obtain.

Explanans2. (Law or Generalization) Conditions of kind C1...Cn obtaining entail that an event of kind E obtains.

---

∴ *Explanandum*: *An event of kind E obtains.*

As you can see, H2 is an instance of D-S explanation, because E follows from (is predicted by) explanans 1-3 with logical necessity. In H2, explanans 2 and 3 are generalizations regarding the conditions under which the car will be able to start.

General laws are what we normally call "laws of nature." Laws are broader, and have less exceptions than a mere generalization about how things have worked in the past. There is a sense that, when describing a law, we are describing something more fundamental and universal than mere in generalization. A mere generalization might look something like "all the apples in my basket are red" (which is equivalent to, "if an apple is in my basket, then it is red"). Whereas a general law might look like this: "matter cannot be created, nor destroyed." What exactly separates law from mere generalization is complex, and beyond the scope of this primer. Hempel and Oppenheim offer extended remarks on it in their paper, "Studies in the Logic of Explanation,"[23] and students are invited to pick the issue up.

Second, the Inductive Statistical (I-S) model, which is differentiated from the deductive models in virtue of its explanatory power, and the nature of the generalization it makes. In I-S explanations, H does not logically entail the explanandum, but, instead, makes it more probable than not. H1 is an example of I-S explanation in action. Like D-N, I-S explanations also include a generalization. However, the generalization can merely be a statistical generalization, about how things normally behave, and not offer any guarantees. The general form of I-S explanation is as follows:

**Inductive Statistical Model (I-S)**

Explanans1. (Particular Conditions) Conditions of kind C1...Cn obtain.

Explanans2. (Statistical Generalization) Conditions of kind C1...Cn obtaining make the probability of events of kind E obtaining >50%.

---

∴ *Explanandum*: *An event of kind E obtains.*

However, I do not intend to endorse Hempel's view of explanation whole heartedly--it is, given the relevance considerations discussed above, untenable. Hempel's model of explanation suggests that explanations are kinds of *reverse arguments* for the explanandum. They are reverse in the sense that we begin with the argument's conclusion and work backwards to construct premises that would support it.

For Hempel, explanations/hypotheses must make the explanandum probable overall (>50% likely). Following Salmon, this is false: all a hypothesis must do is have statistical and reasons relevance--H must only show some of the reasons why E happens, and increase the probability of E

[23] Hempel, Carl G. and Paul Oppenheim, "Studies in the Logic of Explanation," pg.152

*somewhat* above its intrinsic probability. Thus, hypotheses aren't arguments for the probability or certainty of the explanandum.

I propose that we can take the following as the paradigm of explanation. This paradigm does not cast high degree of explanatory power as necessary for explanation, although it allows for explanations to do so.

**Paradigm of Explanation**

Explanans. C1...Cn.[24]

⇓

*Explanandum*: *E.*

Here, the hypothesis is the collection of explanans--the conditions, supposed facts, and conjectures about how things were/are that we point to as potentially explaining E. The double arrow should be taken to indicate the double relevance relations mentioned above: one line for explanatory power (statistical relevance), the other for reasons relevance. That is, the double lined arrow means:

> *C1...Cn obtaining raises the probability of E obtaining above its intrinsic probability, and C1...Cn are reasons why E obtains.*

As described above, I believe these conditions are what the explanatory relation consists of: H correctly (though partially) explains E in virtue of it somewhat raising the probability of E, and specifying (some) reasons why E occurred. The arrow represents the supposed explanatory relationship, which just is relevance, and so we do not need to state this relationship separately, or within the explanans itself. Consider again H1:

Explanans1. The car's battery is dead.
Explanans2. If a car's battery dies, then the car would not have enough electricity to strongly turn over the starter.
Explanans3. In the past, most times (say 7/10 times) that a car does not have enough electricity to strongly turn over the starter, turning the key in the ignition does not result in any noise.

Hypothesis 1

⇓

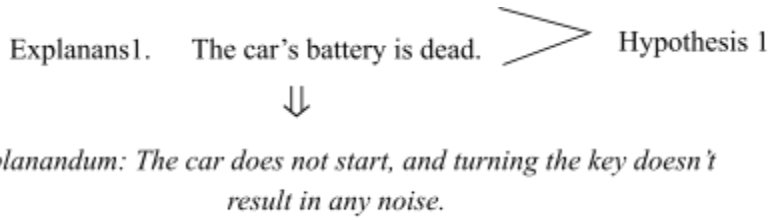
*Explanandum: The car does not start, and turning the key doesn't result in any noise.*

Though it may seem like it at first, these explanans do not necessarily contain any claim to statistical or reasons relevance *to the particular event that happened.* The explanandum is often a particular event, it is the car not starting *this time, today.* Explanans 2-3 support the claim that Explanans 1 is relevant to the general category of "times cars won't start and don't make noise." But the claim that Explanandum 1-3

[24] Keep in mind that conditions could be events, true propositions, membership of some object in a class, etc. C1 could stand for a proposition like "the car's battery won't start," or for "x is a member of class C." The same is true of the explanandum.

are relevant to *this particular time,* that it is the reason why the car won't start *this time,* seems to me to be a separate claim than that, *in the past*, or *generally,* similar conditions had relevance to a similar explanandum. This separate claim, that the explanans are relevant to *this* explanandum, is represented by the double-lined arrow.[25]

But we can also choose, given the peculiarities of our subject matter and our particular theories, to divide up the explanans in different ways. Perhaps we could model H1 like this, instead:




Here, the hypothesis H1 is just the posited conditions that we feel would be relevant, if they were actually the case. Our hypothesis is just that "the battery is dead." Our other explanans, which describe known regularities and past observations about how batteries have *previously* been relevant to cars not starting, are now taken as support for thinking that Explanans1 is relevant (really does have the link symbolized by the arrow) to the Explanandum. In sum, C1...Cn--the conditions we posit to explain E--can include events, propositions, and statistical facts, and our claim is that *these conditions* when taken together are statistically and reasons relevant to E. Precisely how we choose to divide up our explanans may matter for developing a symbolic probability calculus, but for now, we need not worry about it any further.

This structural model would include both deductive and inductive explanations, since C obtaining could raise the probability of E to 100% (deductive entailment), or below that (including both strong statistical relevance, >50%, or weak statistical relevance, <51%). The model is thus compatible with Hempel's explanations! D-N, D-S, I-S would all be instances of this general paradigm of explanation. However, this model does not restrict explanation to D-N, D-S, or I-S. And again, C1...Cn and E could stand for particular propositions or events, or classes/kinds of events, so that this paradigm of explanation models explanations that involve generalizations, as well as explanations only in terms of particular events or propositions.

Finally, note that this is just one way of modeling the structure of explanation. There may be other, equally good, or even better ways of modeling explanation. Models are useful in different ways, in different contexts. Some models are inconsistent, but others are equivalent, or at the very least compatible. For an overview of the various models of explanation, see Peter Lipton's wonderfully clear *Inference to the Best Explanation*, Ch. 2-3. The model I have given above is similar to what Lipton calls the "causal model" of explanation, and resembles Hempel's . However, I have synthesized multiple models of explanations (reasons, deductive nomological, and causal) into a single model.[26] With Lipton, I

[25] When we actually put forward a hypothesis to explain E, we normally don't know that the explanation is correct (we don't know if it is truly relevant). Instead, we put it forward as a *potential* explanation, not as the *actual* explanation.# And so we put forward some supposed facts and considerations (conditions) with the hope that these, taken together, really do meet the relevance conditions. It is then a separate step to establish that the explanans really do have this relationship to E.

[26] Further, my paradigm of explanation is compatible with Lipton's *contrastive* model of explanation. For Lipton, our explanandum question is often of the form, "why did F1 occur, rather than F2?" All one has to do is make the explanandum E represent "F1 occurs, but F2 does not occur." We can even plug in for E: "F1 occurs, but F2 does not, and both were possible." (Lipton, 33-35) For these contrastive explanations, Lipton is welcome to put more

have not restricted the model to casual reasons; instead, I have left open the possibility of other kinds of reasons to figure in genuine, correct explanations (moral reasons, constitutive reasons, etc).[27]

> *NOTE: Again, modeling explanatory theories in the way described above may give the impression that they are reverse arguments, where the explanans stand in for premises and the conclusion is the explanandum. It is "reverse" in the sense that we begin with the conclusion, and work backwards to try to reconstruct the reasons why it happened. However, there are good reasons to think that this is just one way of modeling theories, of giving them structure, and is not to be taken as normative. That is, this structure is not part of the "essence" of theories, but is a clear way of representing our theories on paper. For many every-day explanations, this structure works quite well. However, when one begins to study the complex theories of the hard sciences, other modeling techniques will be more effective. It is often practically impossible to list out all the explanans of a detailed scientific theory, at least not without taking up many pages of paper. For more on this, see the anthology section, particularly the entry on the Semantic Approach to Theories.

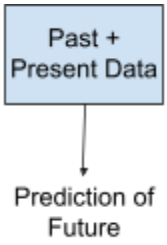
**Theoretical Prediction:**

So far, we've only focused on theoretical reasoning that aims at, roughly, reconstructing the unobservable to explain the observable. But theoretical reasoning can also be used to predict the unknown on the basis of the observable/known/already believed. Remember that predictions are not just predictions of future events, but predictions about what might already be the case, or what was the case in the past, etc. When a prediction is about the past, not the future, it is sometimes called a 'retrodiction.' I am using "prediction" very loosely. This is the sort of reasoning that physicists use when trying to guide a rocket to the moon--they consult well-established theories about how matter behaves in general and about the orbit of the planets, combine those with information about their present conditions, and infer a prediction of where the rocket will land, if they launch it in a certain way, at a certain time. Giving a similar diagram as before, we could represent theoretical prediction of the future as:

---

restrictions on the model, so long as he understands these restrictions as defining a certain *kind* of explanation that the paradigm above *includes*. For instance he might say that the explanans need to specify some relevant difference between F1 and F2 that is a reason why F1 occurs but not F2 (39, 42).

[27] (Lipton, 32)




Again, keep in mind that predictions are not necessarily predictions *of the future,* but of the *unknown* or *currently unobservable.* But I'm no physicist, and, probably, neither are you. So here is a more straightforward--and *definitely* relatable--example of theoretical prediction.

H4

Data1. Most times I have gone to Taco Bell, it has been cause of why my stomach hurts the same night.

Data2. I am going Taco Bell today.

---

∴ *Prediction: My stomach will end up hurting tonight.*

The explanatory theory here is Data1--we have come to believe that the correct explanation for why my stomach has hurt in the past is that eating Taco Bell causes my stomach to hurt. We already have a theory or explanation for why I get a stomach ache, and we already believe it; now, however, we are applying it to make a prediction.

Imagine that you have already eaten taco bell earlier in the day, and now have a stomach ache. You might try to explain why you have a stomach ache like so:

H5

Explanans1. Most times I have gone to Taco Bell, it has been the cause of why my stomach hurts.

Explanans2. I went to Taco Bell today.

⇓

*Explanandum: My stomach currently hurts.*

Do you see how the explanation and prediction have similar structures, but differ in terms of our temporal starting point, and the use we have for them? One difference is in *our* relationship to the event in question. In H5, we are starting from the conclusion (explanandum), and working backwards to predict the unobservable past--we are trying to understand what is under the line. In H4, we are starting from the accepted theory and data, and working towards the conclusion (prediction)--we are trying to understand what the implications of a theory we already believe are. Whether we are predicting or explaining just depends on our location in time and goal, but they are structurally very similar. Are we starting from data we want to explain, and trying to reconstruct the past? That's an explanation. Are we starting from a theory that we already accept, and using it to make predictions about the future? That's a theoretical prediction.

Note, however, that I have retained the line of inference for predictions, instead of the doubled-lined arrow. This is because predictions *are* argumentative: they are inferences from hypotheses and data to some conclusion that an event will occur (or did occur, or is currently occurring, unbeknownst to us). Predictive theories are inductive arguments. This means that a prediction will require its premises to have a higher degree of explanatory power/statistical relevance: the premises paired with the hypothesis must make the prediction greater than 50% probable. I give the general form of prediction like this:

1. Data D1...Dn obtain.
2. Data D1...Dn obtaining makes the probability of E obtaining >50%.

---

∴ *Prediction:* E will obtain.

Or, like this:

1. Data D1...Dn obtain.
2. Data D1...Dn obtaining makes the probability of E obtaining >50%.

---

∴ *Prediction:* We should have >50% confidence that E will obtain.

An important difference between theoretical prediction and explanation is that a prediction need not always involve reasons relevance: we can make reliable predictions by looking at informative correlations, even if we don't have any idea how these correlations provide insight into the *reasons* why the predicted event will/would/might occur. Mere correlations can be good indications of what might happen in the future, even if we don't understand the reasons why those correlations occur. For instance, consider this cliche example of a theoretical prediction:

Data1. The sun has risen every day in human experience.
Data2. We have no good reason to think that the sun will not rise tomorrow.

---

∴ *Prediction:* The sun will rise tomorrow (and we should have very high confidence that it will do so).

This is a reasonable prediction! The data cited seems to make the conclusion >50% probable. Yet, the data does not cite any information on *why* the sun will rise again tomorrow. Yet, it is a perfectly good prediction. Thus, theoretical prediction need not cite reasons, as explanation does. All that prediction needs is *strong statistical relevance*--the data cited must allow us to predict P with greater than 50% confidence. Of course, citing reasons why P will/did/does obtain can greatly help our confidence. Predictions can be made stronger by citing reasons why the prediction will come true, because reasons why P might occur will certainly affect the probability that it will occur. But we can, and often do, make theoretical predictions without being able to cite reasons why our predictions will come true.

We can think of the relationship between explanation and prediction like this: Explanation is an attempt to answer "why?" "how?" "what?" questions. Answers to these questions necessarily involve both statistical and reasons relevance. Finding the most plausible explanation of E involves establishing that

the posited conditions probably obtain, AND that these conditions are most probably the reasons why E happened.

Theoretical prediction is an attempt to show that some conclusion is probable overall, with >50% probability. This conclusion could be a hypothesis about what the future will be like, what the past was like, or what the present, unbeknownst to us, is like. With theoretical prediction we are only trying to show that our guesses about these things are probable, or even more probable than other guesses. This surely involves statistical relevance, and probability relationships between data, previously adopted hypotheses, and the predicted conclusion. However, this does not necessarily require reasons relevance. Of course, predictions can be made stronger by citing reasons why the prediction will come true, because reasons why E might occur will certainly affect the probability that it will occur. But we can, and often do, make theoretical predictions without being able to cite reasons why our predictions will come true.

Explanation is thus a much more ambitious goal than prediction, because it is trying to reconstruct *why* things are the way they are, not just show *that* they are. But both are vital. In fact, explanations that have a high degree of explanatory power *include* prediction. For if H makes E >50% probable, then H also *predicts* that E will be the case. So, explanations can include, or be, predictions, and predictions can be explanations, but they can also be separated from one another.[28]

At this point, let me introduce two symbolic ways of saying that "H explains E" and that "D makes the probability of E greater than 50%." If you don't have any use for symbolic thinking, feel free to ignore this. These symbols will be useful later on, when we develop a logical calculus of probability.

$$\mathbf{H \Rightarrow E}$$

***H explains E*** which means ***H is statistically and reasons relevant to E***

$$\mathbf{P(E|D) > 50\%}$$

***D predicts E*** which means ***the probability of E given D is greater than 50%***

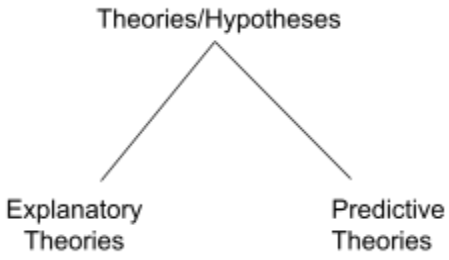
**Summary:**

What should jump out at us right away is that the general method, form or structure of theoretical reasoning in science is the same as in the humanities and everyday life. Scientists and scholars do not have a special way of reasoning, hidden from the public. The research disciplines instead apply human reasoning very carefully, and processes of peer-review and publication are meant to filter out poor applications of reasoning. But the reasoning used is the same the average person already makes use of. This was one of the greatest insights of Charles Sanders Peirce, an American 19th century philosopher. In his work, "On the Logic of Drawing History from Ancient Documents," he demonstrated, quite convincingly, that the patterns of reasoning in the sciences (he would consider history a science) are just technical, complex applications of the patterns of reasoning common to all mankind, across all cultures and time periods.[29]

Overall, the picture of theoretical reasoning that has emerged allows us to understand what we normally mean when we talk about "theories" or "hypotheses." A theory or hypothesis could be either an explanation or a prediction. We often talk about our predictions as "theories." "I have a theory: tomorrow

[28] Hempel's D-S, D-N, I-S explanations are actually, on this model, cases of theoretical prediction!
[29] Peirce, C.S. "On the Logic of Drawing History from Ancient Documents."

the world will end." But we also talk about explanations as theories: "my hypothesis is that the car won't start *because* the battery is dead." This could be diagrammed like so:




Below, you will find a summative glossary of the key terms covered in Section I.

- **Theoretical Reasoning:** A kind of reasoning from known to unknown, observable to unobservable.
- **Explanatory Power/Statistical Relevance:** The relationship between a hypothesis H and event E, such that H increases the probability of E above its intrinsic probability. Also called "explanatory power." Comes in degrees--H can increase the probability of E above 50% overall, which means that H makes E probable. But H can also merely increase the probability of E above its intrinsic probability, which might not increase it above 50% overall (see the penny hypothesis above).
- **Reasons Relevance:** The relationship between a hypothesis H and event E, such that H specifies some of the reasons why E obtained. These could be causal reasons, but there may be other kinds of "reasons why."
- **Explanatory Reasoning:** A kind of theoretical reasoning, in which we construct a hypothesis H which sheds light on why some event or proposition, the explanandum E, occurred or is true. Explanatory reasoning is constrained by relevance conditions--an explanation must have statistical and reasons relevance in order to be an explanation at all.
- **An Explanation/Explanatory Hypothesis/Explanatory Theory:** An attempt to answer "why?" "how?" "what?" questions. A story, or set of posited facts and general principles, which "sheds light on" some puzzling event. By "sheds light on," I mean that the explanation has both statistical and reasons relevance: it indicates why the puzzling event happened. In our paradigm of explanation, the hypothesis is the collection of all the explanans together.
- **Explanandum:** A puzzling event, or proposition about a puzzling event, which we want to explain, or understand "why" it happened.
- **Explanans:** The parts of an explanatory story/hypothesis/theory which contain information that is statistically and reasons relevant to the explanandum. In our paradigm of explanation, the hypothesis is the collection of all the explanans together.
- **Posits:** Entities, events, or states of affairs 'posited' by a theory in order to explain some puzzling piece of data or make a prediction. Posits can sometimes be known to exist or to have occurred, other times posits are theoretical, unobserved, unconfirmed subjects that we predict exist, in order to explain why some puzzling event occurred. A totally unconfirmed, speculative posit is often called a 'black-box' entity. These black box posits are not understood, or even known to exist. Rather, we posit an entity and define it in terms of how it performs a certain function, like 'entity x causes E to occur.' We might confirm that some entity which causes our puzzling event exists, without actually understanding that entity, as if it was sealed away in a dark box, unable to be

seen or studied. However, we can study the object indirectly, but looking at its effects, and trying to establish that it exists, even if we cannot yet directly observe it.[30] This is common not only in science, but in the philosophy of religion, in that God is, in the mystical theological traditions, thought to be knowable only by His effects, and not in Himself.[31]

- **Full vs Partial Explanations:** H counts as a *full explanation* of E just in case it is an explanation that specifies *all the reasons why* E occurred. While H is merely a *partial explanation* of E just in case it is an explanation that specifies just *some* of the reasons why E occurred. Most of our explanations are partial because of our finite, limited perspective.
- **Theoretical Prediction:** An attempt to show that some unknown conclusion E is probable overall, with >50% probability.
- **Noetic-Confidence Probability:** A kind of probability that quantitatively represents what noetic attitudes we should have, or would be rational to have.
- **Statistical Probability:** Any kind of probability that quantifies past or current observations/data-points, and creates a ratio of target obtaining to sample set. For example, I look at the sides of a 6-sided die. I want to know the simple statistical probability of rolling a 6. Two kinds of statistical probabilities might be worth considering. First, I count the number of sides, and treat these as my sample set. There are six sides total. Out of those six, only one side is "6." So, the naive frequency probability is 1/6 (16.6%). Second, I roll the dice ten times. These ten rolls are my sample set. I find that, out of ten rolls, only one of them ended up being a 6. Som my statistical-frequency probability in this sense is 1/10 (10%). Statistical probability does not directly express what we believe or should believe. Confirmation theory will have to show how statistical probability should inform our noetic attitudes (how statistical probability relates to noetic probability).

[redacted]

[30] Black box paper
[31] Aristotle East and West

## Section II: The Absolute Basics of Evidential Confirmation and Disconfirmation

Alright, we now know what explanation and prediction are--we understand what it is to make a theory. That theory, once accepted, can then sometimes be used to make further predictions, as we saw in H4 (when the hypothesis paired with data make the predicted event more likely than not, >50%). We can 'draw out' further predictions from our theories: if *this* hypothesis was correct, then we would also expect to find some other bit of data. And this drawing out of further predictions is a crucial ingredient in the *testing*, or confirmation and disconfirmation, of our theories.

How do we know that a theory is correct, probable, true? How do we know that this set of supposed facts that we've posited are the *true* facts, the events that *actually happened,* the conditions and events that *actually* contributed to the puzzling event's occurrence? We have only talked about constructing theories, but haven't talked at all about how we pick which theory is true out from all the different, competing theories we can invent. This topic is, regrettably, complex. Philosophers have not definitely worked out exactly how we should pick which theory among many is true. The best I can do is to make you aware of the basic findings, and introduce you to the various topics and problems philosophers are still trying to work out.

**Evidence, Testing and Degrees of Probability**

Contemporary, post-enlightenment 'common sense' has it that we should accept or reject a theory based on *evidence*. But what exactly is evidence? Evidence is not a kind/type of data (proposition, fact, event). Rather, evidence is a role a piece of data might play in relation to some hypothesis. Data has an *evidentiary role,* or evidential use. When we look for evidence, we are looking for data that is able to have some impact on the probability of our hypothesis. If some data confirms (increases the probability of) a hypothesis, we call it *positive evidence*; and if some data disconfirms (reduces the probability of) a hypothesis, we call it *negative evidence*.

Not just any data will be able to count as positive or negative evidence, but only data with a certain kind of relationship to our hypothesis. What is, then, this evidentiary relationship? Here are the bare-minimum, absolute basics of evaluating a theory by looking for evidence.

Here is a quick run down of the *testing* step of inquiry: A piece of data D counts as evidence for some hypothesis H whenever H makes some prediction regarding D. We draw out implications or predictions from H--data that would probably obtain given that H was true, data we would expect to find if H was actually correct--and then check if that data obtained. If H predicts D, and D obtains, H successfully predicted D, and this counts in favor of H. That is, we ordinarily take it that a *successful prediction* increases the probability of H. On the other hand, if H predicts D, and D does not obtain (~D obtains instead), then H has made a *failed prediction*. This counts against H: a failed prediction decreases the probability of H.

We formulate our theory and our possible hypotheses proliferate. We ask ourselves, 'if this hypothesis was true, then what else would we expect to be true?' 'What can be predicted on the basis of our hypothesis?' This is to conjecture about the kinds of data that would count as positive or negative evidence for our hypothesis. Finally, we *test* the hypothesis, or perform *experiments*: we check if the data predicted by our hypothesis obtains. If it fails to obtain, then we count that as an indicator of the incorrectness of our hypothesis (disconfirmation), and if it does obtain, then we count that as an indicator of the correctness of our hypothesis (confirmation). Successful predictions are positive evidence, failed predictions are negative evidence. We gather as much data as we can, we look around the world, perform experiments, collect observations, in the hopes of verifying whether the predictions of a hypothesis

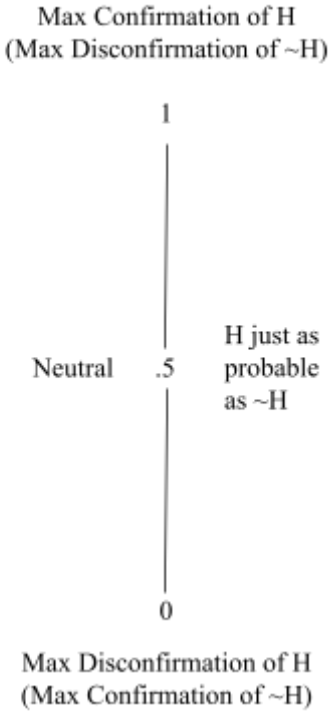
obtained, and how many. We take all these bits of data together, all the successful and failed predictions, and use these as a *body of evidence*.

We then evaluate the overall plausibility of our hypothesis given that entire body of data. One way, I think the most common way, we do this is by asking, 'given that it made *this many* successful predictions, and *this many* failed predictions, how plausible is the hypothesis?' Some degree of probability will be given to the hypothesis given our body of evidence, whether that evidence be a single piece of data, or a collection of data. Again, this is the *conditional probability* of a hypothesis H: the probability of H given some evidence (data).

Evidence for a hypothesis does not have to be conclusive, knock down evidence in order to confirm. Evidence confirms a theory just in case it increases, even just a little bit, its probability. Disconfirmation is the opposite: evidence disconfirms a theory just in case it decreases, even just a little bit, its probability. For most hypotheses important to human life, human beings cannot 'prove' or 'disprove' them with complete certainty: there is only probability.

Our common sense way of talking about probability presumes that probability comes in degrees: 100% probability (or simply 1/1) is something like total certainty in regards to H, or a maximal proof of H. 0% (or 0/1) is total certainty against H, or a maximal disproof of H; another way to put it: if H is 0% probable, then 'H is false' (~H) is 100% probable. 50% probability indicates that H is just as probable as it is improbable (we have just as much reason to suspect H as we do its opposite, ~H). If H is greater than 50% probable, then it is more probable than not, and we can, to some degree, expect H to be true, obtain, etc, though any value under 100% is without complete certainty. When H is less than 50% probable, that means it is more likely than not false, or not to be expected. We have, then, a scale of degrees of probability that can be usefully represented like this:

***Degrees of Probability***



Imagine asking someone if they believe in God. Some ardent believers have no doubts about God; they feel that, for whatever reason, they have absolute certainty that God exists. These people would say that "God exists" is 100% probable--totally, maximally certain or proved (at least from their point of view). A dogmatic atheist might say the opposite: they might have no doubts about the non-existence of God: "God exists," they may say, is 0% probable: "God does not exist" is 100% probable (again, from their point of view). However, a true agnostic will estimate their confidence in "God exists" as the same as their confidence in "God does not exist." That is, a true, idealized agnostic will describe themselves as totally uncertain, totally neutral about the existence of God: God's existence is just as likely as God's non-existence. They may describe their confidence as 50%, or say that they think the existence of God is "50-50." Most of us fall somewhere between a dogmatic atheist and dogmatic theist, without being truly agnostic. We will say things like, "I believe in God, but have some doubts," or "I don't think God exists, but I may be wrong." The greater degree of confidence, belief, certainty, or rational expectation, the higher or lower the probability.

The point is this: testing, or evaluation of H on some piece or body of evidence, works most often by drawing out predictions from H, checking if those predictions succeed or fail, then estimating how much those successful or failed predictions confirm or disconfirm (increase or decrease the probability of) H.

**Maximal Degrees of Confirmation/Disconfirmation (Estimating Conditional Probabilities)**

Sound simple? It's not! We test hypotheses like this all the time, though not with great precision, due to the complex nature of estimating exactly how much, and under what conditions, a successful or failed prediction affects the probability of H. To understand precisely how we should test will involve solving easily one of the most difficult problems in contemporary philosophy: the problem of confirmation. What exactly is it for some bit of data to confirm (increase the probability of) a theory? And how are we supposed to tell how *much* a piece of evidence increases or decreases some theory? That is, how are we supposed to estimate and quantify how much some successful prediction increases the probability of H? How are we supposed to estimate and quantify how much some failed prediction decreases the probability of H? Are there rules for this sort of thing?

There are three issues regarding confirmation and testing that philosophers have devoted serious energy towards, the first being descriptive, the second being normative, the third being theoretical. First, what are the rules or rational principles we use to estimate, or gauge, the probability H given some data D? Second, *should* we use these rules or principles, are they reliable or justified? Third, just what is probability? What does it mean for D to "make H *n*% probable?" These issues are, regrettably, very complex, and I can only hope to survey the first question here by laying out some of the clearest principles of confirmation, and providing copious references (in the footnotes) to works regarding the second and third problems.[32]

[32] David Hume, though not totally inventing a new question, drew significant attention to the first and second problem in *An Enquiry into Human Understanding,* IV. Hume's argument, normally called "Hume's problem of induction," is multifaceted, as he was, in my view, intuiting various problems with confirmation and probability without being able to divide them up cleanly. Hume's overall point is that, at the time of writing in 1748, no one (to Hume's knowledge) had provided explanations for how we can reliably, justifiably predict some unknown H from some body of past data D. Contemporary work on the problems discussed here were motivated in great part by Hume's problem. For a survey of various attempts at justifying inductive inferences, see: (ed. Swinburne, Richard. *The Justification of Induction.*)

We can start, thankfully, with the extremes, or maximal limits, and work inward on the scale. Given our understanding of deductive logic, we know how to gauge when a hypothesis H is totally proved or totally disproved.

Say that H necessitates D, so that, necessarily if H is true, then D must also be true. In symbols, (H → D).[33] This means that, if H was true, then D would *have* to be true, it could not be false. Thus, we could also put this as "H would make D 100% probable." *Though we do not need to master any symbolic system of deductive or probabilistic logic to understand this section, this relation may be symbolized as either (H → D) or P(D|H) = 100%, the latter of which reads "the probability of D given H is 100%."* The flip side of this relation is that, if H is true, then D must happen. That is, D is necessary, or necessarily required for, H.

So, say that H makes D 100% probable: if H is true, then D must also be true. Now say that D is false. Well, *if* H was true, then D would be true, but D is false. Therefore, H must be false as well! For students of deductive logic, this principle, *modus tollens,* will be familiar. If (H → D), but D is false (~D), then H is guaranteed, with necessity, to be false (~H). H requires D, so whenever D is false, H is false. This leads us to our first principle of confirmation/disconfirmation.

> **Principle of Maximal Disconfirmation:** If H necessitates D (predicts D with 100% probability), and D does not obtain, then H is 0% probable.
>
> ↪ If P(D|H) = 100%, then P(H|~D) = 0%.[34]

This principle describes how we can completely *rule out hypotheses*. Here is a (fairly) clear, though still trivial, example:

> *Imagine that we are tracking down an unknown animal in our neighborhood--it keeps killing birds, and we want to catch it. I have a theory: this animal is a cat. Now my theory predicts, with 100% certainty, that this animal will be a mammal--because all cats are mammals (being a mammal is required for being a cat).*

Let us say that I somehow find out that the animal we're tracking is not a mammal (~M). I would know, with total certainty, that my theory was wrong: it's not a cat (~C), precisely because the cat theory would *guarantee* that the animal is a mammal (C → M). Thus, this example illustrates that, if a theory predicts some data with 100% certainty (necessarily (C → M)), and the prediction does not come true (~M), the theory has been totally refuted, or maximally disconfirmed (the probability of ~C = 100%; or the probability of C = 0%).

Now, say that necessarily (D → H). That is, if D was true, it would necessarily guarantee H: the only way D can be true is if H is true. We could again symbolize this as either (D → H) or P(H|D) = 100%. Say that D obtains. Well, since D requires the truth of H, then D obtaining guarantees the truth of H.

---

[33] Technically, we need modal operators to symbolize these principles. For (H → D) is too weak for our uses, we must say that, necessarily, H guarantees D. Which would be symbolized as □(H → D).

[34] I have included, below each principle written in English, equivalent symbolizations for advanced students. P(D) is the intrinsic probability of D. P(H|D) is the conditional probability of H assuming D. P(H|~D) is the conditional probability of H assuming that D is false. P(H|D) > P(H) means that the probability of H given D is greater than the probability of H alone, meaning that D increases the probability of H. P(H|D) < P(H) means that the probability of H given D is lower than the probability of H alone, meaning that D decreases the probability of H.

> **Principle of Maximal Confirmation$_1$:** If D necessitates H (D makes H 100% probable), and D obtains, then H is 100% probable.
>
> ↪ If P(H|D) = 100%, then P(H|D) = 100%.

However, as students can quickly see, this is an absurdly trivial, almost unhelpful principle. It reads: a hypothesis H is maximally confirmed by a prediction which maximally confirms it… A more helpful way of writing the principle, though still trivial and fully equivalent to the first principle, might be:

> **Principle of Maximal Confirmation$_2$:** If D necessarily requires H (the probability of D without H (given ~H) is 0%), and D obtains, then H is 100% probable.
>
> ↪ If P(D|~H) = 0%, then P(H|D) = 100%.

Here is another trivial example:

> *My theory is that the unknown animal killing the birds in my neighborhood is a cat. Now I know that all cats are felines, because that's just what it is to be a cat: all cats are felines, all felines are cats, necessarily. So, my theory predicts with 100% probability that the animal is a feline. At the same time, the relationship goes both ways: since being a feline is just what it is to be a cat, then being a feline requires being a cat.*

If the animal ends up being a feline, then this data itself makes the theory "it is a cat" 100% probable. Why? Because being a feline requires being a cat, and vice versa. The probability that some animal is a feline without it being a cat is 0%! Without the animal being a cat (~C), the animal could not be a feline (~F). So, let's say that I find out the animal is a feline, F. Since I know the probability that F would be true while C is false (~C) is 0%, I can conclude that C *must* be true. So, when a hypothesis predicts successfully, and the prediction without the hypothesis could not be true, then the hypothesis is maximally confirmed, or proved with maximal probability: total certainty.

**Two Kinds of Confirmation/Disconfirmation Principles: Confirmation/Disconfirmation Conditions and Confirmation/Disconfirmation Estimation**

However, these are only the maximal limits, the extremes of how data, serving as evidence, can affect the probability of a hypothesis. The vast majority of hypotheses and predictions in real research are not so cut and dry. There are many predictions we make from theories that neither rule out nor totally prove a theory. Let's say that H doesn't guarantee a prediction D, but only makes it somewhat probable that D will occur. If D obtains, what then? How much does D increase the probability of H? Or, say that D *could* have possibly happened without H, but probably wouldn't. How, and how much, would D obtaining effect the probability of H, in this situation?

We thus are in need of two kinds of principles. First, principles that tell us when, under what conditions, some data D affects the probability of a hypothesis H, and whether it increases or decreases that probability. Second, principles that tells us how much, under these conditions, D increases or decreases the probability of H. The second kind of principle regards laying out rules for how to estimate the *amount* or degree of confirmation of disconfirmation, while the first kind of principle only lays our rules for *when* some data confirms or disconfirms. At the maximal limits, we can easily discover principles which fulfill both roles: I have given them above as the Principles of Maximal Disconfirmation and Confirmation. When we are dealing with all-or-nothing extremes, of decisively ruling out or decisively establishing, the principles are simple. However, the issue becomes much more complex when

we try to codify principles of the second sort that tell us how to work within the extreme limits. To avoid confusion, let us, at this stage, work primarily with the first kind of principle. I will also describe the second kind of principle shortly, but will keep the discussion minimal: such principles are complex, and full development of them is best left for a later time, during which students more interested in advanced topics will be shown how to develop a precise logic of probability, i.e. a probability calculus.

**Principles of Confirmation/Disconfirmation Conditions**

The most straightforward ways in which data can confirm a hypothesis regard, again, successful or failed predictions. Say that H predicts something--it increases the probability of the prediction above what it would have been otherwise (the prediction's intrinsic probability), and it does so strongly enough to make it >50% probable. That is, given H, the prediction will probably happen, and H actually contributes to the probability that it will happen. Calling back our notion of *statistical relevance*, we are saying that H is statistically relevant to the prediction, *and* makes it likely overall. If H was really true, if H really obtained, then we would expect the prediction to come true.

Now consider what the effect on H would be if the prediction failed. It seems to me clear that, if H leads us to expect something in such a way, then its failure to obtain would disconfirm, or lower the probability of, H. Leading to my first principle...

> **Principle of Disconfirmation via Failed Prediction:** If H predicts that D will probably not obtain, then D disconfirms H to some degree. Equivalently, if H predicts that D will probably obtain, then ~D disconfirms H to some degree.
>
> ↪ If P(D|H) < 50%, and P(D|H) < P(D), then P(H|D) < P(H)
>
> ↪ If P(D|H) > 50%, and P(D|H) >P(D), then P(H|~D) < P(H)

Consider an example:

> *A detective is trying to solve a murder. Mr. Pimplewise was killed by someone, that much is clear. But how did he die? Mr. Pimplewise's body is brought in for autopsy. The medical examiner, together with the lead detective, find that, in Mr. Pimplewise's chest, there is a deep, round, roughly circular hole with smooth edges. This hole is very near his heart. It looks either to be a stab wound, or a small-caliber gunshot. If it was a stab wound, the medical examiner is certain that it was with a pointed, smooth, rounded object--something akin to an ice-pick. Accordingly, our detective formulates two hypotheses:*
>
> > *H1: The hole in Mr Pimplewise's chest was caused by being shot in the chest with a small-caliber firearm.*
> > *H2: The hole in Mr Pimplewise's chest was caused by being stabbed in the chest with a weapon similar to an ice-pick.*
>
> *On closer examination, there is no wound on Mr. Pimplewise's back, nor his sides. Whatever produced this wound did not pierce through his back or sides.*
>
> *Our detective draws out some predictions on the basis of H1: if this wound was produced by a small-caliber gunshot, then the bullet would very likely still be in the body, since there was no exit wound. If a full autopsy is performed, then, if H1 was true, the medical examiner would find the bullet. It is possible, of course, that the killer may have extracted the bullet himself, to cover his tracks. It is also possible that the bullet, being such a small caliber, shattered into very*

> *small pieces, and that, even given an autopsy, the medical examiner will not be able to find them. Still, it is plausible that, if H1 was true, then we will find the remains of a bullet in the body during an autopsy.*

Poor Mr. Pimplewise! The autopsy is performed, and the medical examiner cannot find a bullet, nor can she find any fragments of a bullet. How does this effect H1? It seems to me that the failed prediction lowers the probability of (disconfirms) H1. It does not, of course, rule H1 out--maybe the killer already extracted the bullet, or maybe the medical examiner missed some incredibly small fragment. It simply decreases the probability of H1, to some degree, though I cannot precisely quantify by how much (that would, after all, take codifying the second sort of confirmation principle discussed above). Turn now to consider how a successful prediction effects the probability of H:

> *Our detective now considers H2, and draws out some predictions: if the wound was caused by being stabbed with a weapon similar to an ice-pick, then we would not expect to find a bullet, or remains of a bullet, upon autopsy. For, obviously, a stab wound will not result in a bullet being placed in the body. So, if H2 was true, then we would expect not to find any bullet or bullet fragments in the body. However, there is no guarantee, if H2 was true, that no bullet would be found. It is technically possible that Mr. Pimplewise had been shot long ago in an unrelated incident, and the bullet never fully removed, so that bullet shards may be found upon autopsy. So, H2 does not guarantee that there will be no bullet. Further, it is also possible that Mr. Pimplewise was first shot, then stabbed through the bullet hole--even if we find a bullet, it does not disprove that Mr. Pimplewise was stabbed. So, H2 doesn't guarantee that a bullet will not be found, it merely makes it probable. And a bullet being found does not guarantee that H2 is false. All we can say, is that, if H2 was correct, then, probably, there would be no bullet found.*

The medical examiner digs in once again to the pitiful Mr. Pimplewise and, still, finds no bullet, nor bullet shards. It seems very clear to me that this raises the probability of H2--it indicates that H2 is the correct explanation, despite not giving any guarantees. Simply because H2 raises the probability that no bullet will be found, and because it makes it overall more probable than not that no bullet will be found, finding no bullet counts in favor of H2. Leading now to my second principle:

> **Principle of Confirmation via Successful Prediction:** If H predicts that D will probably obtain, then D confirms H to some degree. Equivalently, if H predicts that D will probably not obtain, then ~D confirms H to some degree.
>
> ↪ If $P(D|H) > 50\%$ and $P(D|H) > P(D)$, then $P(H|D) > P(H)$.
>
> ↪ If $P(D|H) < 50\%$ and $P(D|H) < P(D)$, then $P(H|\sim D) > P(H)$

Finally, consider one more case study as an illustration of these confirmation and disconfirmation principles:[35]

> *In Israel, archeologists have found many wax/clay seals ("bullae") pressed by signet rings dating back thousands of years. These seals were used to bind documents together while marking the documents with a signature, to ensure their authenticity or authority. One such seal, the*

[35] Veen, Peter van der, Robert Deutsch and Gabriel Barkay. "Reconsidering the Authenticity of the Berekhyahu Bullae: A Rejoinder."

> *BEREKHYAHU BULLAE, reads "Belonging to Berekhyahu, Son of Neriyahu, the Scribe." Scholars tell us that the names "Berekhyahu" and "Neriyahu" are the ancient Hebrew spellings of "Baruch" and "Neriah," mentioned in Jeremiah 36:4. Baruch was Jeremiah's scribe, and so would have his own signet ring to bind together documents for dispersal around the Kingdom of Judah. It seems, then, that this bullae could belong to the scribe of Jeremiah the prophet!*
>
> *Some people had doubts, however, because of the shady origins of the bullae--it was bought from a black-market antiquities dealer. Many of these dealers sell forgeries, fakes, to naive and unsuspecting historians. Historians and archeologists were split: is it real? Is it fake? One thing was clear, however: if this bullae was truly a fake, then it wouldn't require detailed historical and paleographical knowledge to make. Think about it: if the bullae was a fake, then it was faked by someone who is probably not a scholar, and so we would not expect to find writing methods that would require detailed historical knowledge of, say, paleography (the study of ancient writing systems). So, if the bullae was a forgery, we would predict that the bullae would not feature any ancient writing systems unknown to the vast majority of the common people. Keep in mind that these were discovered in the 1960s-70s, and forgers did not have access to the internet!*

We could divide this theoretical prediction more clearly, and model is like this:

Data1. The Baruch Bulla is a forgery, made by a contemporary antiquities forger. (Our previously adopted hypothesis)
Data2. The vast majority of antiquities forgers are not historians nor paleographers, and so do not have technical knowledge of ancient writing systems.
Data3. If one does not have technical knowledge of ancient writing systems, then one will likely not be able to create a fake bulla with accurate paleographical inscriptions.

---

> *Prediction 1: The Baruch Bulla very probably has inaccurate paleographical inscriptions.*
> *Prediction 2: The Baruch Bulla very probably lacks any traces of accurate writing techniques that would require either technical, paleographical expertise or for the maker to have lived contemporaneously with Jeremiah.*

Do you see how the forgery theory leads us to predict, with very high probability, that there must be some inaccuracies on the bulla? Some scholars, looking at the bulla, pointed to some features that do not match other writing techniques common in the 7th-6th century BCE (the time period of Jeremiah and Baruch). The forgery theory, after careful inspection of the bulla, made a successful prediction! A theory making a correct prediction is intuitively thought to offer the theory *a degree of (positive) confirmation.* However, the story has a twist:

> *On closer inspection, some scholars have found traces of accurate, 7th-6th century BCE paleography on the bulla. It's a complicated, technical issue, but there does seem to be some reason to think that certain writing techniques on the bulla would qualify as ancient.*

That is, Prediction 2 failed! What then, for the forgery theory (H6)? If these scholars are right--if there are really marks indicating knowledge of ancient paleography on the bulla--the forgery theory has made a failed prediction, and this offers a degree of disconfirmation against it. Similar to how a failed prophecy

counts against the claim of the prophet to be inspired by God, a failed or inconsistent prediction counts against the theory we draw the prediction from.

**Principles of Confirmation/Disconfirmation Estimation**

Now, very briefly, and in as minimal fashion as possible, I wish to lay out some rough principles for estimating the *degree* to which some data D confirms a hypothesis H. We, so far, know the conditions under which D confirms/disconfirms H, as well as the upper and lower limits of degree of confirmation/disconfirmation. Now, I offer two principles for estimating degree within those limits. However, these principles will not allow us to precisely quantify our estimations.

We will, as always, proceed by using thought experiments and drawing out intuitions about how likely some failed or successful prediction would make our hypotheses. Imagine again that…

> *A detective is trying to solve a murder. The murder took place this morning, in a rural, American town, at the front of the Roadside Motel. Reports from eyewitnesses are conflicting. Supposed eyewitnesses offer two stories: first, some say that the suspect shot the victim from a distance; second, others say that the suspect ran up and stabbed the victim, making physical contact with the victim. Accordingly, our detective formulates two hypotheses:*
>
> > Ha: *The victim was shot to death from a distance, and did not make physical contact with the killer.*
> > Hb: *The victim was stabbed to death, and made physical contact with the killer.*

Now, if Ha was true, then we would almost certainly not expect to find fresh stab wounds on the victim (although it is possible that the victim was stabbed by someone else before or after being shot, so there is no guarantee). Call this prediction D1. Also, if Ha was true, then we would expect to find an empty bullet cartridge around the scene of the crime. For most guns, when fired, eject the empty shell, and it falls to the ground. Of course, this prediction could easily fail, because the killer could have quickly picked up the casing, or could have fired with a gun that does not automatically eject the empty casing. Call this prediction D2. As such, Ha predicts D2 with much less confidence than it predicted D1.

We have here two predictions from Ha, D1 and D2. What sets these predictions apart is how strongly Ha predicts them: Ha makes D1 *extremely likely*, while Ha only makes D2 *somewhat likely*.

Now, assume that both predictions fail. Certainly, if it was found that the victim *had* fresh stab wounds (~D1), then that would severely reduce the probability that they were shot from a distance (Ha). And, obviously, if crime scene investigators could not find an empty cartridge around the crime scene (~D2), then the probability of Ha would also be reduced.

But now ask yourself, "which failed prediction is *worse* for Ha?" It seems very intuitive to suggest that the presence of fresh stab wounds (~D1) lowers the probability of Ha more than the failure to find the empty bullet casing (~D2). But why? It seems to me that the relevant difference between them lies in how strongly Ha predicted D1 compared to how strongly Ha predicted D2. The more strongly the hypothesis predicted the data, the more a failed prediction harms (disconfirms) the hypothesis! This leads to a rough principle:

> **Principle of Degree of Disconfirmation via Failed Prediction**: The more strongly H predicts D, the more the failure of D (~D) disconfirms H.

> ↪ That is, the more H increases the probability of D, the more ~D reduces the probability of H.

Now imagine a slightly changed scenario. Ha also predicts D3 with great confidence: on the body, we will find at least one fresh gunshot wound. Upon examining the body, there are multiple gunshot wounds. And, at the crime scene, we find a spent bullet casing (D2). Ask yourself, which successful prediction is more beneficial for the hypothesis Ha, D2 or D3? Again, it seems to me that D3 increases the probability of Ha much more than D2 does. And, I think, it is in part because of how strongly Ha predicted D2 compared to how strongly it predicted D3.

But, I think, there is also another reason why finding gunshot wounds (D3) increases the probability of Ha more than finding the bullet casing (D2): finding an empty bullet casing near the crime scene was already fairly probable, given the location. Anyone who has lived in a rural, American city and who has looked closely enough will have seen many spent bullet casings lying around on highways, on street corners, in alleyways, etc. There are many ways empty bullet casings can end up as litter in America: people (illegally) messing around with firearms, police or criminal shootings, accidental discharges, and the loss of souvenirs.[36] So, even though Ha predicts that there will be an empty cartridge at the crime scene, simply finding an empty cartridge is not a very powerful find. And this seems to me to be because the probability that one would find an empty cartridge *even if the victim was not shot* (~Ha) is fairly good. However, finding a fresh gunshot wound on the victim (D3) would be *extremely unlikely* if the victim was not shot, as Ha describes. So, without Ha, it is not very probable that the victim would have a fresh gunshot wound, much less *multiple* fresh gunshot wounds.

So, it seems that the degree to which a piece of confirms a hypothesis is a function of two things: the strength of the prediction, and the likelihood of the prediction without the hypothesis being true:

> **Principle of Degree of Confirmation via Successful Prediction:** The more strongly H predicts D, and the less likely D would be without H, the more a successful prediction of D confirms H.
>
> > ↪ That is, the more H increases the probability of D, and the less probable D is without H, the more D increases the probability of H.

Still, these principles do not enable us to precisely *quantify* or precisely *estimate* degrees of confirmation and disconfirmation. However, they do allow us to make useful comparisons. Say that you have two hypotheses, Ha and Hb, and you find some data D. Now, you know that Ha makes D very unlikely, but Hb only makes D slightly unlikely. Given that you found D, you now know that Ha is more strongly disconfirmed than Hb, and so, on this piece of evidence alone Ha is less probable than Hb.

Likewise, we find some data D that Ha makes very likely, and Hb makes only slightly likely. Further, D would probably not happen if Ha was false, while D could very likely still happen even if Hb was false. Even these rough principles would then enable us to compare Ha and Hb given this piece of evidence: on this piece of evidence alone, Ha is more likely than Hb. Of course, when making these kinds of comparisons, we are making comparisons based on a single piece of evidence. Real inquiry requires us to take multiple pieces of evidence into consideration, as well as other factors, like the simplicity of the hypotheses (as we will see shortly). Still, we now have some idea, in principle, how to do such things.

**The Relevance Restriction(s)**

[36] (many people, after going to a gun range, will bring home one or two empty casings to remember the fun they had).

These principles, oversimplified for practical purposes, require qualification and nuance in order to be fully accurate. As I already mentioned above, I am assuming that H has statistical relevance to D, in that H raises the probability of D. In these principles, when I say that "H predicts D," I mean that H increases the probability of D somewhere above 50%. I do not mean merely that D is probably true given that we assume H. I mean, instead, that H *actually contributes* to the likelihood of D obtaining: if H occured, then D would be more probable than it was on its own, or without H: H makes a difference to D. Thus, D only confirms or disconfirms H if H truly makes a difference to the probability of D, by raising it above its prior or intrinsic probability.[37]

If we did not keep this restriction in mind, then we would run into absurdities. Imagine that H makes no difference to whether D happens or not. Yet, D might still have a very high probability for other reasons. It would then be technically true that, if H obtained, then D would be very probable (because D is probable anyways, and H has no effect on D). But it would not be because of H! If D happens, it had nothing to do with H, H did not make it any more likely or unlikely. In this case, I do not see how D could count as positive evidence for H. Think back, again, to the example given by Salmon, slightly tweaked to illustrate the point:[38]

> *John is a biological male. As such, the probability that John gets pregnant is 0%--John will certainly, with 100% probability, not get pregnant (~D). To explain this fact, John's wife comes up with a hypothesis:*
>
> > *H: John has been taking my birth control pills.*
>
> *However, clearly, the probability of 'John will not get pregnant' (~D) has nothing to do with his taking birth control pills! The probability of 'John will not get pregnant' stays the same either way. Still, John's wife is absolutely correct in making this prediction: if H was true, if John has been taking the pills, then he will not get pregnant.*

Now, say that John does not get pregnant (~D). Does this confirm H? For H did predict ~D with very high probability, and ~D did obtain… Clearly not, because H is *irrelevant* to ~D: H makes no contribution to the probability of ~D. If we did not keep this restriction in mind, we would have to admit that John not getting pregnant is evidence of his taking birth control pills. Which, intuitively, seems ridiculous. ~D cannot confirm H if H makes no difference to ~D!

All that to say: in order for some data D to confirm or disconfirm H, H has to actually affect the probability of D, it has to be *statistically relevant* to D by raising or lowering it's probability. I have already built these restrictions into the Principles of Confirmation/Disconfirmation above. We can write the relevance restriction as such:

> **Relevance Restriction(s):** For a hypothesis to be confirmed or disconfirmed by data, the hypothesis must actually make a difference to its probability.
>
> ↪ $P(H|D) > P(H)$ only if $P(D|H) > P(D)$
>
> ↪ $P(H|D) < P(H)$ only if $P(D|H) < P(D)$

[37] Richard Swinburne gives similar restrictions in *Epistemic Justification,* although I have made the restrictions a bit weaker.

[38]

**Some Principles of Confirmation and Disconfirmation**

Taking all our (useful) principles covered so far together, and ordering them so as to match our probability scale (from max confirmation to max disconfirmation), we have:

**Principle of Maximal Confirmation$_2$:** If D necessarily requires H (the probability of D without H (given ~H) is 0%), and D obtains, then H is 100% probable.

↪ If P(D|~H) = 0%, then P(H|D) = 100%.

**Principle of Confirmation via Successful Prediction:** If H predicts that D will probably obtain, then D confirms H to some degree. Equivalently, if H predicts that D will probably not obtain, then ~D confirms H to some degree.

↪ If P(D|H) > 50% and P(D|H) > P(D), then P(H|D) > P(H).

↪ If P(D|H) < 50% and P(D|H) < P(D), then P(H|~D) > P(H)

**Principle of Degree of Confirmation via Successful Prediction:** The more strongly H predicts D, and the less likely D would be without H, the more a successful prediction of D confirms H.

↪ That is, the more H increases the probability of D, and the less probable D is without H, the more D increases the probability of H.

**Principle of Disconfirmation via Failed Prediction:** If H predicts that D will probably not obtain, then D disconfirms H to some degree. Equivalently, if H predicts that D will probably obtain, then ~D disconfirms H to some degree.

↪ If P(D|H) < 50%, and P(D|H) < P(D), then P(H|D) < P(H)

↪ If P(D|H) > 50%, and P(D|H) >P(D), then P(H|~D) < P(H)

**Principle of Degree of Disconfirmation via Failed Prediction**: The more strongly H predicts D, the more the failure of D (~D) disconfirms H.

↪ That is, the more H increases the probability of D, the more ~D reduces the probability of H.

**Principle of Maximal Disconfirmation:** If H necessitates D (predicts D with 100% probability), and D does not obtain, then H is 0% probable.

↪ If P(D|H) = 100%, then P(H|~D) = 0%.

**Relevance Restriction(s):** For a hypothesis to be confirmed or disconfirmed by data, the hypothesis must actually make a difference to its probability.

↪ P(H|D) > P(H) only if P(D|H) > P(D)

↪ P(H|D) < P(H) only if P(D|H) < P(D)

These are not necessarily the only ways in which data D can serve as evidence for or against a hypothesis H. That is, these principles are almost certainly not exhaustive. There may be other conditions under which D increases or decreases the probability of H. However, these, in my view, are *intuitive, justified*, and *powerful.* Just by grasping these principles, we are in a better place to inquire, to test theories, to evaluate them on the basis of evidence. The hard work of establishing rules for quantification, or principles for estimating degrees of confirmation/disconfirmation below the limits, is still yet to be done.

[redacted]

## Section III: The Problem of the Proliferation of Hypotheses[39]

We've seen what explanatory theories are, and two very intuitive ways we can try to support them with evidence. We also saw, though we did not dwell on it for very long, that a fact, observation, or data point can be explained by multiple hypotheses. The mysterious, bird-killing animal mentioned in Section II could have been a cat, a dog, a raccoon, etc. Each of these represent multiple hypotheses. Further, positive evidence might confirm all of them at the same time. What if we find out that the mysterious-animal has four legs? Each theory--it's a cat, it's a dog, it's a racoon--makes the same prediction: the animal probably has four legs. By our principles of confirmation, finding out that the animal has four legs confirms all three hypotheses. When this happens, and there's not enough positive or negative evidence (successful or failed predictions) to decide which of the competing theories are true, how do we decide which theory is true (or the most likely to be true)? Only one of them can be, so which one is it? Which theory is the *best*?

The problem only gets worse. Technically, for any given data, there are a potentially *infinite* number of competing explanations for each theory, all of them with equal explanatory power. The human mind is so clever that we can slightly tweak or modify a theory in small ways, so as to make a new theory. We can always add one more detail, one more piece of nuance to our hypothesis, and thereby produce a new hypothesis. Maybe the animal killing all the birds in my neighborhood is a dog. Maybe it's not just a dog, but a stray dog with no vocal chords (I might say this to account for why it doesn't bark when I approach it). Maybe there are actually two animals, a dog and a cat. Maybe the dog kills birds on Monday, Wednesday and Friday, but the Cat kills birds on Tuesday, Thursday, Saturday and Sunday. Or maybe it's a cat, a dog, and a fox, and they alternate schedules in even more complicated ways… Technically, each of these theories has equal explanatory power--each of these ridiculous inventions of the human mind equally well lead us to predict the data we want to explain: something has been killing birds in our neighborhood.

Maybe we want to explain what caused the Big Bang. Maybe God caused the Big Bang. Maybe God created another being, and together they caused the Big Bang. Maybe God created another being, and that being alone caused the Big Bang. Maybe the Big Bang was uncaused, and random--something came from nothing. Maybe the Big Bang was caused by some naturalistic chain of events before it, without God. Maybe the universe has always existed and moves in cycles: Big Bang, unfolding events, a collapse, then a Big Bang again, and the cycle continues on forever. Each of these would explain why the Big Bang happened, and why the world exists as a result of the Big Bang. But which one is true? Here is one final example, to drive the point home:

> *Imagine you just started a job as a banker, and you are asked to help audit the bank account of a client. This client has only $100 in their account. Their previous balance as of yesterday was $1,000. Strangely, all records of previous transactions have been wiped from bank records. Your boss, Mr. Moneyson, has asked you to figure out exactly how this client has ended up with only $100. The bank is worried that there may be fraudulent withdrawals from the bank, but want to make sure.*

---

[39] For more in depth studies of this problem, see:
(Swinburne, Richard. *Simplicity as Evidence for Truth.*)

For an interesting application of the problem of proliferation, see: (Griffiths, P. E. "The Historical Turn in the Study of Adaptation."). Griffiths contends that significant changes in how modern evolutionary biology conceives of Darwinian theory have been driven by the need to reduce the total number of competing hypotheses.

Mr. Moneyson, the greedy capitalist that he is, has given you a totally impossible task. Why? Consider the various hypotheses that would all equally well explain the fact that the client has only $100 in their account:

**H1.** Balance yesterday was $1,000, withdrew $900.
**H2.** Balance yesterday was $1,000, deposited $100, then withdrew $1,000.
**H3.** Balance yesterday was $1,000, deposited $1, then withdrew $901.
**H4.** Balance yesterday was $1,000, deposited $1,000,000, then withdrew $999,900.

Each hypothesis here would result in today's balance being just $100. We could invent a literally *infinite* number of hypotheses to explain why there was yesterday $1,000, and today only $100. Each has the same explanatory power, each makes the same prediction. How are you supposed to pick which hypothesis is the best, most likely true, etc?

This is an instance of the Problem of the Proliferation of Hypotheses (PPH):

> ***PPH:*** *for any given explanandum, there will always be at least two hypotheses with equal explanatory power; that is, there will always be at least two hypotheses that make all the same predictions with equal confidence. So, even if we had all the relevant data and evidence in front of us, we could not rule one out or decrease the probability of one in favor of the other (by finding failed predictions). We also could not find positive evidence that would increase the probability of one over the other (by finding successful predictions). For each hypothesis makes the same predictions--they will have all the same successes and failures.*

So, we cannot use confirmation and disconfirmation via evidence to make the final decision between competing hypotheses. We can use it to narrow down or elevate some hypotheses above others initially, but there will always be at least two hypothesis left. That's a major problem! We need some other method to decide between competing explanations, or else we will be left with a set of equally plausible guesses.

Of course, our first step in trying to figure out which of many competing theories is the most likely is to look for more evidence. As covered in Section II, it's often easiest to look for negative evidence. We rule out or reduce the probability of some of these competing theories by finding failed predictions. This allows us to narrow down the number of competing hypotheses.

But note that this problem (PPH) arises *even in ideal circumstances*, where we have access to all the relevant data/evidence. But real life is not ideal, and inquirers *never* have all the relevant data. We can never do all the experiments we could, we can never look at every example: we simply don't have the time, money, or even faculties to do so (some theories are not about the physical world, after all, but instead are about immaterial truths and objects--how could we ever have all the data on those?). In real inquiry, we will always have multiple hypotheses and, even though they make different predictions, our limitations prevent us from using those differences to conclude in favor of just one. Let's call this the Realistic Problem of Proliferation (RPP):

> ***RPP:*** *Because of human limitations, for any given data, there will always be multiple hypotheses that have the same level of confirmation, and the same level of disconfirmation, given the evidence/data we have access to. So, we will not be able to decide in favor of one or the other on the grounds of confirmation and disconfirmation.*

The point is this: we've hit a wall, a hurdle, a stopping point--at least two theories are tied in terms of the evidence, and we need a tie breaker. We may be crippled, unable to decide which theory is correct. So what method, what principles, can we appeal to in order to continue moving forward? This will be the topic of the next section…

**Exercise 1:** Come up with three equally powerful explanations for the following explanandum. Keep in mind that equally powerful *does not mean* equally probable.

> E: *At 8:30am, a loud 'bang' was heard on 6th and Broadway. Police arrived on the scene, and found a crashed car, the driver dead. The windshield had been broken, and there was a hole in the windshield on the driver's side.*

**Exercise 2:** What sort of evidence would you look for to decide between each of your three hypotheses constructed above? That is, what kinds of data would you need to find in order to decide which theory was the most likely to be correct? Explain.

**Exercise 3:** Take a guess--before reading the next section, how do you think we might get past the problem of proliferation? That is, do you have any ideas for how we can narrow down many theories to just the correct one? Share your thoughts. Again, do this before reading the next section!

## Section IV: Inference to the Best Explanation

**IBE in General:**

The difficulty created by the proliferation problem for selecting a theory as 'true' requires some sort of methodological tool that could help us decide which explanation we should prefer over the others. Enter Inference to the Best Explanation (IBE).[40]

IBE is a way of *comparing* multiple theories; it is not a tool for creating theories, or even testing individual theories, but a method or *process** for comparing multiple theories in the hopes that one emerges as "the best."

> *NOTE: IBE contains within it methods of relating individual propositions, posits, hypotheses, etc. to some body of evidence. It is not a simple method or single, unified kind of reasoning, but a series of rational processes taken together. It is *not* a kind of discrete inference/argument, but a *process*, a "step" in the overall process of inquiry that is itself composed of many other processes and discrete inferences. (This is, I take it, very similar to how C.S. Peirce conceived of the matter; see the anthology section entry on C.S. Peirce's theory of inquiry.) Some, like Gilbert Harman, have taken IBE to be a particular form of argument/discrete inference similar to the form I give below. However, this form of argument is best understood as a *summary* of the process of IBE, not the process itself. I will repeat this note below, because it is critical!

IBE comes into play when we already have a number of theories. The general idea is this: theories have a variety of strengths and weaknesses--"best-making features"--that make a theory more or less probable, more or less reasonable, better or worse in comparison to other theories. If I am left with a number of hypotheses, I must have some way to compare them, to decide which, among its competitors, is the most likely to be true, or the most reasonable to adopt, given our interests. There are important philosophical differences between calling a hypothesis "the most probable" and "the most reasonable." To avoid overcomplicating things, proponents of IBE simply say that IBE is a way of deciding upon the "best theory" out of some set of competing theories. Then, it is a separate question whether "best" means "probable" or "reasonable," and what the distinctions between those two concepts are. It is also a separate question whether the "best" theory should be the one that we believe.

But what are the details of IBE? How does it work? And what does it mean for a theory to be "the best?" There are many ways to construe the reasoning processes we engage in when trying to establish a theory as the "best one." As Lipton notes, for most philosophers, IBE seems to function as merely a general slogan--IBE is a vague way of thinking about these reasoning processes, but most have not really

[40] As far as I am aware, Gilbert Harman, writing in 1965, was the first to use the phrase "Inference to the Best Explanation" as the name for a distinct kind of inferential process. Harman offers little analysis of IBE in his article, however. He simply says that in making an IBE, one infers "from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis." (Harman, "The Inference to the Best Explanation".) As mentioned previously, by 1910 Dewey had described, though not in very great detail, the reasoning processes that we now call IBE (Dewey, John. *How We Think*, Ch. VI).

thought very deeply about what goes on in them. IBE without more detail is vague, too vague to be useful, because our reasoning processes are complex, and aren't always aimed at the same goals.

The reality is that our criteria of "best" will change depending on our context. So finding the "best theory" will certainly mean different things at different times, depending on our interests. And thus the reasoning processes required to find the "best" theory will differ given the different senses of "best."

Let's try to avoid confusing ourselves, and say that IBE includes all the kinds of pro/con, strength/weakness reasoning we use when trying to decide on the "best" *explanatory* theory. Call this kind of reasoning "weighing" or "balancing" reasoning. IBE is just a general term, then, for reasoning that involves weighing alternative explanatory theories. Thus, IBE will very likely include many different kinds of reasoning processes, with many different goals/conclusions. Certainly these processes will be similar to one another, and have overlapping features. They are all, after all, part of the same family of reasoning (IBE).

We can put it simply: in my particular context, what am I trying to conclude about my theories by weighing them against each other? We then analyze the sort of weighing-reasoning pattern (the sort of IBE) we need to reach our conclusion, and implement it. What does it mean, in my context, to find the "best" theory? What sorts of best-making features will I need to consider, and weigh, to find the best theory?

Let us simply avoid confusion, and avoid the trap of expecting to find a single model for every kind of weighing process.[41] We can recognize that there may be many different kinds of IBE. Each kind could be differentiated according to its goal, as well as by the sorts of considerations required to reach that goal. We could give a tentative, partial list of the different kinds of IBE below...

IBE1. Inference to the Most Probably Correct Explanation (out of some finite set of competitors)
IBE2. Inference to the Most Empirically Adequate Explanation
IBE3. Inference to the Most Economical Explanation (out of some finite set of competitors)
IBE4. Inference to the Most Useful Explanation (out of some finite set of competitors)
IBE5. Inference to the Most Cogent Explanation/Prediction (out of some finite set of competitors)
IBE6. Inference to the Most Rational Explanation/Prediction (out of some finite set of competitors)
IBE7. Inference to the 'Loveliest' Explanation (out of some finite set of competitors)[42]

---

[41] As Dewey earlier noted, each question we ask may require slightly different methods, or, at the least, different ways of applying a single method. The fine details of how, say, IBE is applied to answer particular questions may vary according to the content of the question. What is important is to have a general idea of theoretical method, and to cultivate skill in applying, and modifying, that method to meet our goals in particular cases.

Dewey writes: "The disciplined, or logically trained...is able to judge how far each of these steps needs to be carried in any particular situation. No cast-iron rules can be laid down. Each case has to be dealt with as it arises, on the basis of its importance and of the context in which it occurs. To take too much pains in one case is as foolish--as illogical--as to take too little in another. At one extreme, almost any conclusion that insures prompt and unified action may be better than any long delayed conclusion; while at the other, decision may have to be postponed for a long period--perhaps for a lifetime. The trained mind is the one that best grasps the degree of observation, forming of ideas, reasoning, and experimental testing required in any special case, and that profits the most, in future thinking, by mistakes made in the past." (*How We Think*, Ch. VI).

[42] Lipton develops the idea of the 'loveliest' explanation in detail. See: (Lipton, Peter. *Inference to the Best Explanation,* Ch. 2.)

Each of these is a different way of weighing best-making features with the aim of finding the one that is "best." But what counts as "best" will determine what we count as "best-making features," and that will in turn depend on our context. For instance, C.S. Peirce would likely rank "economy" as a best-making feature. The economy of a theory does not effect it's probability of being true, accurate, etc. Rather, it is how economical the theory is for us to test given our finite monetary and material resources (IBE3). Another feature might be cogency, or accessibility--how easy a theory is for someone (or some group) to understand (IBE5). These both may have overlap with "usefulness," or how useful a theory would be if we adopted it (IBE4). Or, we could perhaps look for the "best" theory that has the best balance of all or some of these features--the 'loveliest' theory (IBE7).

Finally, but perhaps most importantly, some philosophers have taken the goal of science to be discovering "empirically adequate" theories--theories that have predictive power (can predict future experience), and fit all our past and future observations. The "best" theory would be the most likely to be empirically adequate (IBE2). A perfectly successful scientific theory would, on this view, be one that coheres perfectly with experience, and can be used to predict future experience. Empirical adequacy is distinct from truth--it is conceivable that two incompatible theories might both perfectly cohere with experience, and be equally powerful predictors. Clearly, they cannot both be "true." In trying to find and compare the probability of theories, we are not trying to discover *how likely they are to be true.* Rather, we are trying to find *how likely they are to be successful,* i.e., consistent with and good predictors of future experience.*

*NOTE: Here, I am only giving a brief overview of the first kind of IBE (IBE1)--inference to the most probably correct explanation. However, much of what we discuss here can be applied to the other forms of IBE, especially IBE2. For instance, a pragmatist like van Fraassen may object to my interpretation of epistemic probability, and will add that scientists often go beyond IBE1 to make scientific progress. For if we are to ever go beyond the extraordinarily weak and tentative conclusion that IBE1 leaves us with, we will have to discover more powerful methods.

In any case, it seems to me that such a view will merely adopt a different conception of "best" that includes, but is not limited to, "the most probable," and thereby posit additional theoretical/explanatory virtues than the list I give for IBE1. IBE1 is, in my view, the most barebones conception of IBE that one could give, and it should, I think, be able to be subsumed under more rigorous forms of IBE and explanatory paradigms. That is, IBE1 is really just an attempt to capture the probabilistic elements of our pro/con reasoning about theories. There is, surely, much more that is involved in actual scientific and everyday practice.

**IBE1 in more Detail:**

For our purposes, I will assume that "best" means the most probably correct explanation *out of some finite set of competitors*. However, this is a simplification for our class--"best" is a value judgement, and other considerations besides probability might make one theory "better" than others.[43]

In simple terms, IBE1 is an act of weighing pros and cons--a procedure any student should already be familiar with from ordinary life. However, the pros and cons weighed in IBE are not purely pragmatic, but have to do with considerations that shape the probability of each theory.

IBE1 compares competing theories by looking at some of their "best-making" features, called "explanatory virtues," weighing their comparative strengths and weaknesses, and concluding that, because one theory has the best overall balance of these best-making features, it is the "best." Many philosophers, including myself, think that the "best" theory is just the theory with the highest overall probability, *in comparison to its competitors.* This is not to say that it is the *most probable explanation possible,* only that it is the most probably correct *out of the set of theories we are considering.* * IBE1, then, is fairly modest.

*NOTE: IBE1 comes to a very weak conclusion, and this must be kept in mind. Many philosophers who take IBE to be a key step in the process of inquiry take IBE to reach a stronger conclusion than that of IBE1: that the hypothesis is *probable overall*, out of all possible hypotheses, and not merely as compared to the competitors we have thought of. Those who take this view would say that the conclusion of IBE1--that *h* is the most probable of the competitors we have thought of--also warrants a further conclusion: that *h* is probable overall, or rational to accept. However, that *h* is the most probable out of its competitors, and that *h* is probable *simpliciter,* are distinct claims, and these philosophers have the burden of showing us that the former implies or makes probable the latter. See the anthology section paper, "Where Do We Go From Here? The Limit of IBE"

Something critical for understanding IBE1, and which is often overlooked, is the nature of its conclusion. Again, the conclusion is that some hypothesis H is the "most probable explanation out of its competitors." This involves two claims. First, that the posits of H (i.e. the story, events, entities, etc H predicts to exist) are more probable than the posits of all its other competitors. Second, that the posits of H are the most probable 'reasons why' the puzzling event occurred. Think about the difference: I am not pregnant. There are two explanations: I am taking my wife's birth control pills, or I am a biological male. The first story could be true--for some odd reason, I am taking the pills. Yet, this would not be the correct explanation. The correct explanation tells us the *real reasons why*. So, it is one thing to say that a theory is true, in that its story and posits have obtained, and another to say that it is the correct explanation (gives the real reasons why).

So, IBE1 does not merely conclude that the story H tells is more probably true than the others we know of. In addition, IBE1 is meant to conclude that the story H tells specifies the reasons why the explanandum obtained: it is arguing for a conclusion about 'causality,' widely construed, or the 'true

[43] Pragmatists like Bas van Fraassen contend that there is more, much more, to the process of rational belief formation/theory confirmation and adoption than mere probabilistic concerns.

causes' of the puzzling event. This is what it means to be a "correct" explanation--it is to be an accurate story that specifies the real reasons why E occurred.

One way of summarizing the overall reasoning in IBE1 is to cast it as a deductive argument as follows:[44]

1.) Data E1 obtains.
2.) H1 is the best explanation of E1 out of its competitors.

---

∴ *H1 is the most probably correct explanation of E out of its competitors.*

There is likely a tacit premise hidden in our argument for IBE, so that, really, we may be arguing that:

1.) Data E1 obtains.
2.) H1 is the best explanation of E1 out of its competitors.
3.) If any hypothesis is the best explanation of some E out of its competitors, then it is the most probably correct explanation of E out of its competitors.

---

∴ *H1 is the most probably correct explanation of E out of its competitors.*

The only difference between these two forms of IBE is that one makes explicit the link between "best" and "most probably correct explanation."

Again, since we say that a theory is best just in case it has the best overall balance of best-making features (explanatory virtues), in comparison to its competitors, we can give more detail to the model of IBE1 above like so:

1.) Data E obtains.
2.) Out of its competitors, H1 has the best overall balance of best-making features in regards to E.
3.) If any hypothesis has the best overall balance of best-making features in regards to E out of its competitors, then it is the most probably correct explanation of E out of its competitors.

---

∴ *H1 is the most probably correct explanation of E out of its competitors.*

Each of these models is highly similar, and nothing significant depends on what model we prefer (to my knowledge). They are each useful and appear to be equivalent, just divided up in different ways.

*NOTE: IBE contains within it methods of relating individual propositions, posits, hypotheses, etc. to some body of evidence. It is not a simple method or single, unified kind of reasoning, but a series of rational processes and inferences taken together. It is *not* a kind of discrete inference/argument, but a *process*, a "step" in the overall process of inquiry that is itself composed of many other processes and discrete inferences. (This is, I take it, very similar to how C.S. Peirce conceived of the matter; see the anthology section entry on C.S. Peirce's theory of inquiry.) Some, like Gilbert Harman, have taken IBE to be a particular

[44] This is, roughly, how Gilbert Harman conceived of IBE. (Harman, "The Inference to the Best Explanation")

form of argument/discrete inference similar to the form I have just given above. However, this form of argument is best understood as a *summary* of the process of IBE, not the process itself.

Note also that I have not specified which type of probability we are discussing in the conclusion. In my view, the conclusion of IBE1 has to do with the noetic-confidence probability of hypotheses, not mere statistical probabilities. That is, IBE is not merely a way of calculating statistical probabilities, but is concluding something like: "we ought to have a certain degree of confidence towards the proposition that 'H1 is the correct explanation of E' " or "we have most reason to accept H1 as the correct explanation of E out of its competitors." Or even "we should be more confident that H1 is the correct explanation than H2...Hn." However, I will leave this question open for now.[45]

**Best-Making Features, or Explanatory Virtues**

But just what are the "best making" features of theories? Between two (or more) theories, what features would make one "better" (more probable) than the other(s)? On what grounds can we compare theories to each other?

Though there is disagreement about the details of each consideration, and about the exact list of relevant considerations, the following features are widely cited in philosophical practice as "best-making features" or explanatory virtues. Fine-grained theories of IBE may revise this list, but we can at least use the following as a starting point:

**Explanatory Virtues**

- **(B1)** Logical Coherence (Intrinsic Probability)
- **(B2)** Explanatory Power
  - (a) Reasons Relevance
  - (b) Statistical Relevance
- **(B3)** Cumulative Effect of Positive and Negative Evidence (Conditional Probability)
- **(B4)** Theoretical Simplicity (Intrinsic Probability)
  - (a) Number of Posits and Kinds of Posits
  - (b) Degree of Falsifiability/Explanatory Scope (Number of Predictions and Specificity of Posits)
  - (c) Symmetry/Fit with Background Evidence

This list is not to be taken as final (especially as regards Theoretical Simplicity, of which there are other kinds not listed here). For instance, we may take (B1), logical coherence, as a *precondition* for being admitted to the process of IBE, since any hypotheses we would normally consider should already be coherent. In other words, there may be other ways of dividing up this list, and of conceiving of our act of weighing. Further, much more can be said on each of these best-making features. For our purposes, best-making features 1-5 should be understood well enough to allow us to get a general idea of "weighing" the strengths and weaknesses of theories against each other.

[45] Lipton puts the issue like this: "Inference to the Best Explanation does better [than other models of explanation], since it brings in competition ***selection***" (*Inference to the Best Explanation,* 67). That is, IBE allows us to conclude that a particular theory (or set of theories) is the one we should accept, or think true with a certain degree of confidence. This is a normative conclusion, not a conclusion about mere statistical probabilities.

Note that the phrase "all else being equal" is used frequently below. This phrase means that, between two hypotheses, *if* they are equal in regards to all other best-making features, one will be better/more probable than the other in virtue of some other best making feature.

**(B1) Logical Coherence:** First, a theory must be internally, logically coherent: it must not contain logical contradictions. Each of our theories being weighed are ordinarily already presumed to be *coherent* prior to being compared. Still, sometimes a hypothesis contains a hidden contradiction within its explanans, and we only notice this *after* weighing it against its competitors. So, we include it tentatively in our list of best-making features. If a hypothesis is incoherent, then this would reduce its intrinsic/prior probability to 0% (for a contradiction cannot possibly be true). As mentioned above, we could conceive of IBE as only occurring *after* all our incoherent theories have been discarded, so that B1 is a precondition for IBE, not part of the IBE process.

**(B2) Explanatory Power:** Remember that 'explanatory power' is an ambiguous term, used often to refer to at least two distinct concepts: statistical and reasons relevance. A theory is reasons relevant just in case it specifies at least some of the *reasons why* the explanandum obtained. As covered in Section I, if we discover that H1 fails to meet the reasons relevance requirement, H1 is ruled out as the correct explanation of E. For a correct explanation is just a story that specifies some of the real reasons why E occurred! Likewise, if H1 has no statistical relevance, it does not make the explanandum any more likely than it ordinarily would be without H1. Thus, it cannot be the correct, much less the best, explanation. So, we can use the statistical and reasons relevance requirement (P1) to rule against one hypothesis, in favor of another.

Keep in mind, however, that in real inquiry, we are almost never in a position to judge the *reasons* relevance of some condition/event C1 to another, E, at least not directly. We are mostly unaware of whether some *particular,* posited condition C1 actually does have reasons relevance to the *particular* puzzling event we want to explain E. If we knew this, then we would already know that C1 is a correct explanation of E! So, in most cases, we will not be in a position to directly compare the particular, posited conditions of hypotheses in terms of their reasons relevance to the explanandum. Instead, we will have to discover the statistical relevance of *kinds* of posited events/conditions to *kinds* of puzzling events. We will look for *past* data, regarding *other* events of the same *kind* or *general category* as our posited explanans. Then, we will look for whether or not those *kinds* of events were correlated with the same *kind* of event as our explanandum. If there is a positive correlation between *kinds*, then that is itself evidence for thinking that there would be a positive correlation (statistical relevance) between the *particular,* posited events and our particular explanandum. Thus, at least in the realm of causal explanation, past, *general* statistical data about *kinds* of events are really one of our only measures for assessing the statistical relevance of *particular* explanans to *particular* explananda. Further, since statistical relevance is typically our only indication of causal relevance, and since *particular* statistical relevance is only assessable through *general* statistical relevance, the clearest way to assess causal-reasons relevance is often through general statistical relevance. Data gathering is thus very important![46]

[46] We might also add that statistical relevance to the explanandum has a positive effect on the probability of H1. It did not seem necessary to add this as a separate best making feature, however, because it simply reduces to positive evidence (B3). The degree to which a hypothesis makes the explanandum likely affects its probability. For if H1 has statistical relevance/explanatory power for E, then it makes E more likely than it would have been by itself (without H1). If the explanatory power of H1 makes E greater than 50% likely, then H1 predicts that E will probably occur. We know that the explanandum obtains, because it is the puzzling event we want to understand, which means it must have obtained in order to puzzle us in the first place. So, the explanandum E always obtains, just by being the explanandum. And if H1 makes E >50% likely, then E counts as a successful prediction of H1. Successful

Bas van Fraassen, writing on the concept of 'Laws of Nature,' comes to a similar conclusion (which I will oversimplify and distort, a bit):[47] when we are trying to establish some theory about what causes what, we posit certain entities, events or laws, and are attempting to figure out whether these posits are, in fact, causes (reasons why) the puzzling event occurred. If we knew ahead of time that these general or particular kinds of posits were really *causes* of (reasons relevant to) these general or particular kinds of puzzling events, we would already have the answer to our question, "why did this puzzling event *E* occur?" The entire process of inquiry, in order to be more than definitional, is attempting to discover reasons for why things happen. So, we cannot hope to debunk (disconfirm) nor verify (confirm) a theory simply by directly 'checking' whether or not its posits are 'true causes.' You cannot appeal to explanatory (reasons) relevance to confirm or disconfirm a hypothesis directly--we will never actually be in such a position to do so, as that would essentially assume that we are already in position to know the truth straightaway.

Going beyond van Fraassen now: we must appeal first to statistical and observational regularities, or other ways of estimating probability relations. This sort of data can be an indicator of explanatory (reasons) relevance, though not a sure guarantee. That is, statistical and observational regularities can suggest, and make probable, propositions like "*H* is reasons relevant to *E,"* or, more specifically, "events of kind *H* are causes of events of kind *E*." However, propositions like "*H* is reasons relevant to *E*" or "*H* is the cause of *E*" cannot ever raise the probability of *H* given *E,* nor *E* given *H,* beyond what statistical or observational regularities can do! This is not an *in principle* limit, but one arising from our limited perspective. If we knew all the true causes of the universe, like God, we could, of course, know that P(*H*|*E*) is very high *precisely because H*s are normally *causes* of *E*, and this causal (explanatory) consideration might raise up P(*H*|*E*) above what mere statistical regularities could. If we could know that "*H*s cause *E*s 80% of the time," that would make P(*H*|*E*) higher than if we only knew that "*Hs* are correlated with *E*s 80% of the time!" However, we are not in a position to know "*H*s cause *E*s 80% of the time" *prior* to establishing that "*Hs* are correlated with *E*s 80% of the time," etc. So an appeal to causal, or reasons relevance does not give us a further boost of confidence in *H* given *E*, in addition to our knowledge of statistical relevance. We (attempt to) establish reasons relevance on the basis of statistical relevance. And so the P(*H*|*E*) cannot be boosted beyond what the statistical data (observational regularities, etc) has already shown.

In the end, however, reasons relevance is very important, since it plays a negative role. If our statistical data suggests that *H* does not have reasons relevance to *E,* then that is good reason to doubt that the theory is the *correct* explanation for the puzzling event! For part of being a *correct* explanation is to

---

predictions increase the probability of any hypothesis in proportion to both the degree to which the hypothesis makes E likely, and the likelihood of E regardless of the hypothesis. So, H1's degree of explanatory power above 50% increases the probability of H1 just because it is a special case of evidential confirmation. The very fact that H has explanatory power (statistical relevance) that raises the probability of E above 50% increases the probability of H! So, all else being equal, H1 is more probable than H2 if H1 has more explanatory power for E than H2 does.

[47] (van Fraassen, Bas. *Laws and Symmetry*, 181.) Note, here, that I am attempting to reconcile two views, though, perhaps, in a way that neither side will be happy with. Philosophers like Bas van Fraassen assert that explanatory concerns are irrelevant to probability, while those like Peter Lipton and Richard Swinburne give them a primary place in their probability calculations. How can these brilliant minds be in such tension? The reason, I think, is because van Fraassen overlooks the possibility that explanatory concerns can, *in principle,* boost probability above statistical concerns, while those like Lipton and Swinburne overlook the fact that we do not have access to these explanatory (causal) facts from our limited, human perspective. Explanatoriness (i.e., reasons relevance) *can* boost the probability of a hypothesis. However, we are not ordinarily in a position to assess reasons relevance, and thus not in a position to factor that boost into our probability estimates.

give the *true* or *real* reasons why! If we can cast doubt on *H* by showing that it is, probably, not providing the real reason for *E*, we thereby cast doubt on *"H* is the correct explanation of *E*." This can be done by showing that there is no profound correlation between *H* and *E,* for instance, or by other methods, as discussed in the anthology section under "How Do We Establish Causal/Reasons Relevance?"

All in all, appeals to explanatory concerns (reasons relevance) cannot confirm (boost) the probability of a theory in any way not already achievable by appeals to observational or logical concerns (statistical relevance, etc). However, appeals to explanatory concerns can disconfirm (reduce) the probability of a theory, as described above.

To summarize *explanatory power* in slightly more colloquial terms: if we are entertaining two competing hypotheses, H1 and H2, and H2, even if adopted as true, would leave more questions unanswered than H1, then H1 has more explanatory power than H2. Further, if H1 and H2 are equal in every other regard, but H1 would make our explanandum *E* less puzzling or unexpected than H2 would, then H1 has more explanatory power than H2. In either case, one hypotheses has more power to answer our questions (leave less unexplained) than the other, and has more power to clear up our puzzlement at the unexpected (predict *E* with greater probability).

**(B3) Cumulative Effect of Positive and Negative Evidence:** If we take our various bits of evidence/data together, we can get at least a rough idea of how the evidence impacts the probability of the story posited by each individual theory. Successful predictions besides the prediction of the explanandum increase the probability that the posits of the hypothesis obtain (P2b). Unsuccessful predictions reduce the probability of the hypothesis (P3b). If we take a single, unified body of evidence, and look at how that body of evidence affects the probability of each hypothesis, we can then compare the probability of each hypothesis given our evidence. All else being equal, the hypothesis with the greatest degree of confirmation given our body of total evidence will be the most probable. (I am leaving out the qualifications of "in proportion to the degree to which H1 makes prediction E likely, etc, but these still very much apply. See the principles of confirmation given above.)

**(B4) Theoretical Simplicity:** Though we have not yet covered the role of simplicity in theorization, we will in the next part of this book (in Richard Swinburne's contributed section, *Simplicity as Evidence of Truth*).[48] For now, suffice it to say that each hypothesis has a degree of *intrinsic probability* or likelihood. This intrinsic probability is often taken to be correlative with its degree of simplicity: as the saying goes, the simplest explanation is probably the correct one. We can compare the simplicity of multiple theories against one another. Perhaps one involves many more assumptions than another, in which case the one is more complex, or less simple, than the other. If, as we will see in the next section, simplicity increases the probability of a hypothesis, then we can say that, all else being equal, the simpler of two hypotheses is the most probable.

Theoretical Simplicity, however, is very complicated, and any particular understanding or application of theoretical simplicity will have passionate detractors. There are, as covered by Richard Swinburne, many senses of 'simplicity.' The three most easily graspable, and least contentious senses of theoretical simplicity are as follows.

First, we have simplicity as the ***number of posits or kinds of posits***. The more entities, events, causal powers, relationships, laws of nature, etc. that H posits in order to explain something, the less probable it is. That is, the more posits it takes to explain some event, the less simple, and less probable that hypothesis is. Another way to put the same idea: the more complex a theory is, the less probable it is:

[48] Note that I do not carve up the kinds of simplicity exactly as Swinburne does, but I do agree with him substantially.

more assumptions we have to make to explain E, the more improbable our story is. This is, in common language, called *"Ockham's Razor."* Recall the various banking hypotheses in Section III: it seems to me that the simplest of these theories, in terms of the number of transactional steps involved, would be the most plausible, barring all outside evidence. Further, the more *kinds* of entities, laws of nature, properties, etc that H posits, the less likely H is. If we can explain some puzzling event by only positing *two* kinds of entities, rather than *ten* kinds of entities, we should prefer the former. The more kinds of posits a theory puts forward, the more that theory multiplies the entities which, supposedly, exist, their relationships, and the laws governing them: it increases the complexity of our ontology (an ontology is a list of all the things and kinds of things that exist). A theory that simplifies our 'ontological list' is more probable than one that over-complicates it! Everything else being equal, the hypothesis with fewer posits and/or kinds of posits is the more probable.

Second, we have simplicity as the ***degree of falsifiability.*** 'Degree of Falsifiability' is simply how easy the theory is to disconfirm, or show to be false. Overall, the more easily falsifiable, the less probable the theory, and the less easily falsifiable, the more probable the theory. There are two components of the degree of falsifiability, or two factors affecting how falsifiable a theory is: *number of predictions* and *specificity of predictions/posits*. We might also call this ***explanatory scope*****,** or "informativeness."

It is intuitively obvious that the more claims someone makes about the world, the more likely they are to be wrong. For each claim someone makes is another opportunity to be mistaken, so that, the greater number of potential mistakes someone can make, the greater the probability that they are mistaken. Say that we have two theories, H1 and H2. H1 makes five predictions, while H2 makes five hundred predictions. Assuming that everything else is equal, which is more probable? Most of us, I think, will say H1, precisely because, given the far greater number of predictions, it is very likely that we will find that one of them fails.

Further, the more specific (precise) predictions H makes, and the more precise posits, the less probable that theory is. Conversely, the less specific H's predictions and posits, the more probable H is (all else being equal). Specificity refers to the amount of precise detail given by a theory. This is easiest to grasp by example. Imagine that you see a car drive by very fast. Your friend estimates that the car must have been going 75mph (and on a back road, at that!). You, however, insist that the car was going between 60-80mph. Which of your theories are more probable? Clearly the more imprecise, the theory which predicts with less narrow values. One reason for this is that, the more narrow, precise values a theory predicts, the more likely it is to be inaccurate. All it would take to disconfirm your friends theory is to show that the car was traveling less than or greater than 75mph: even 74.6543mph would be incompatible with your friend's theory, strictly speaking. However, your theory admits of a great range of possibilities: the true value could lie anywhere between 60-80mph, and so finding the values 74mph, 73mph, 74.6543mph, would not defeat your theory. If you've ever played a 'bean-guessing' game, where there are a certain number of jelly beans in a jar, and you are tasked with guessing how many there are, this should be familiar to you. You could guess that there are between one and five hundred jellybeans in the jar. Of course, the rules would prevent you from doing this, and will force you to guess a more narrow range of beans. Why? Because surely everyone who guesses such a broad range will be correct, and the game would result in a tie. The interesting part of the game comes down to guessing more narrow values, precisely because it is very improbable that any specific person guesses it correctly.

Likewise, if a theory posits very specific properties of objects, or very specific laws of nature governing the world, that theory is less simple than a less precise theory. Imagine, as we did in section II, that I am trying to solve a murder. We previously had two theories: Mr Pimplewise was killed with a gun,

or Mr. Pimplewise was killed with a tool like an ice-pick. I could come up with another theory, however: Mr. Pimplewise was killed with a gun made either by Sig Sauer, Glock, or Smith and Wesson. Which is more probable, if we have no other evidence, the broader, less precise gun-theory, or the narrower? I can narrow it even further: Mr Pimplewise was killed with a gun made by Smith and Wesson. Now which is more probable, in the absence of any evidence? We would surely need some compelling evidence to feel comfortable affirming either of these narrower theories. But why? It seems to me that this is due to the fact that the more specific, precise, or narrow the traits of a posited subject are, the less probable that theory is, all else being equal. So, we feel that we need some strong evidence to warrant the more specific theory, in order to overcome its *intrinsic* implausibility (compared to the more vague theory).

Third, we have simplicity as ***symmetry,*** or ***fit with background evidence***. In real inquiry, inquirers do not pursue understanding in a vacuum. Rather, each of us is preceded by thousands of years of prior research, in which thousands of scholars have put forward, and argued for, many diverse theories: we are but stepping stones. Especially today, given our advanced technologies and well-confirmed theories of mechanistic science, there are a large number of very plausible theories, so well established as to be considered by many "the truth."

When we encounter some new phenomena, or an old phenomena that puzzles us, it is best to try to explain it in a way that would *fit with* these well established theories we already accept. That is to say, the more a theory *fits* with pre-existing, pre-established theories, the more probable it is, and the less a theory first with these, the less probable it is. If we want to explain E, we could do so by completely inventing an entirely new framework for understanding the world, or we can work within the already established framework. Theories 'fit' with others by having similar, or the same, posits (explanatory story): a theory that makes up entirely new entities, laws of nature, or causal powers in order to explain some strange occurance is much more *ad hoc* than a theory that uses already verified, already understood posits. Sometimes, none of our pre-established theories are relevant to the puzzling data we want to explain, so we will need to posit very different sorts of subjects. Other times, entire explanatory systems, called *worldviews* or *paradigms,*[49] are debunked, and a radically different view of the world may need to replace it. Still, for the most part, we should seek to maintain a degree of continuity and similarity in our explanations. For in so doing, our entire view of the world is simplified--we reduce the number of posits required to understand the world *overall*. Our particular theory or explanation is one part of our entire worldview, and the better that theory fits with others, the simpler the overall worldview. The less 'fit' between our new theory and our body of background data (including previously accepted theories), the more complex, overall, our systematic view of the world. Our theories should be, as far as we can reasonably make them, like to our old theories! As Bas van Fraassen states, "essentially similar problems have essentially similar," or *symmetrical*, "solutions."[50]

Though this is controversial, it seems to some that the better a theory fits (the more symmetrical a theory is), the more plausible it is. Indeed, if simplicity raises the probability of a theory, then fit might also. For the better a theory fits with our background evidence, the simpler our overall view of the world.

Despite the complexities involved, simplicity is a crucial, indispensable ingredient of theorization, for only with it do we have any hope of overcoming the proliferation problem. As discussed in Section IV, it is easy to come up with many (even infinitely many) equally powerful hypotheses. We will always face at least two competing hypotheses that are equal, at least as far as we can tell, in regards

[49]

[50] Note, however, that van Fraassen does not think that symmetry boosts the probability of a theory. Rather, it makes the theory more likely to be (*Laws and Symmetry*, 235.)

to (B1-B3). To decide between them, we will need some other factors we can look to. Simplicity (B4) plays this role--when we reach a stalemate, we can decide on the best theory by choosing the simplest from among the competition. When comparing theories, the simpler is the more probable, all else being equal.

At the end of the day, we consider these various best-making attributes of theories, and try to decide which theory has the best overall balance of them. We are, in effect, trying to take all the different considerations which affect the relevance and probability of a hypothesis into account, "add them up," and see which hypothesis comes out as both relevant and with the highest probability. That theory will be the "best" (among its competitors). Many times, however, we won't be able to narrow the competition down to a single theory. For it is possible, and happens often, that multiple theories will emerge as equal in all respects (B1-B4): they will end up in a tie. The best we can do, then, is to keep an open mind, and continue looking for evidence. Over time, we continually find more evidence, and continue to re-weigh and re-evaluate our hypotheses. Inquiry is never finished, and probably will never be final.

**A Note for Advanced Students--IBE is a Compound Method/How IBE Relates to Confirmation Theory and the Probability Calculus:**

Note that I am not construing IBE1 as a way of figuring out the probability of a theoretical prediction or degree of confirmation of a prediction given some evidence. IBE1 is a way of weighing various explanations against one another, not a way of figuring out how some evidence E affects the probability of P (it is not a model of conditional probability). This will require development of a theory of confirmation and conditional probability, which we will introduce later. Of course, we can only compare the probability (degree of confirmation) of two explanations H1 and H2 by understanding how some body of evidence E has shaped the probability of H1, and how E has shaped H2 (how much E confirms or disconfirms H1 compared to how much if confirms or disconfirms H2). This requires us to understand conditional probabilities on some level. Then we compare the probdabilities of H1 and H2 given our evidence E to each other, which is part of IBE (B3).

Think of it this way: IBE is how we sum up, and come to a conclusion, after looking at all the details of how evidence shapes probability for each individual hypothesis. We can understand this way of summing up on an intuitive level first, then go back to work out the details later (i.e. we understand how to use IBE1 intuitively, then ask for more detail on how the individual parts of IBE1, like confirmation, work).

Put another way, it is not very clear how our best-making features determine the overall probability of an individual hypothesis. Similarly, it is not very clear how our best-making features allow us to conclude that "H is the most probably correct explanation of E." This will become a topic in itself, under the header of "Confirmation Theory," in later sections.[51]

Overall, IBE1 *contains within it* methods of relating individual propositions, posits, hypotheses, etc to some body of evidence. It is not a simple method or single, unified kind of reasoning, but a series of rational processes taken together for pragmatic purposes. It is *not* a kind of discrete inference/argument, but a *process*, a "step" in the overall process of inquiry that is itself composed of many other processes and discrete inferences. (This is, I take it, very similar to how C.S. Peirce conceived of the matter; see the anthology section.) For example, in order to fully understand IBE1, we must investigate the following topics:

[51] Lipton writes, "What reason is there to believe that the explanation that would be loveliest, if it were true, is also the explanation that is most likely to be true?" (*Inference to the Best Explanation,* Ch. 3)

- **Theory of Probability:** An analysis of the term 'probability,' and how it is linked to the aims of inquiry (i.e. truth, accuracy, empirical adequacy, etc).
- **Theory of Conditional Confirmation/Probability:** Normative principles for calculating or assessing probabilities of some hypothesis, proposition or posit P in relation to some evidence E. (To set rules for assessing B3)
- **Theory of Intrinsic Confirmation/Probability:** Normative principles for calculating, assessing or quantifying the probability of a hypothesis or prediction P in virtue of itself. (To set rules for assessing B1, B4)
- **Theory of Comparative Confirmation/Probability:** Normative principles for comparing the conditional and intrinsic probabilities of multiple hypotheses, propositions, etc., and for what to infer from those comparisons. (To set rules for comparing competing theories with respect to B1, B3, B4)
- **Theory of Causal/Reasons Triangulation**: A method for how to discover whether a given posit or kind of posit is reasons relevant, particularly causally relevant, to some explanandum. (To set rules for assessing B2)[52]
- **Development of a Probability Calculus:** A symbolic language for neatly and clearly calculating, expressing and making use of the confirmation principles, similar to a system of symbolic, deductive logic. (To more efficiently and clearly go through the process of B1-B4)

You see, I hope, how complex IBE1, or IBE in general, can be. IBE is a large, messy method that proceeds as a holistic, organic process.[53] It contains within it methods of probability calculation, probability assignment, comparative weighing, and explanatory evaluation (causal or reasons relevance triangulation), not to mention the foundational problem: analyzing the very concept of 'probability' that we have taken for granted, and its relationship to 'truth' or the aims of inquiry.

**Hopeful, Concluding Remarks:**

We have seen how complicated Inference to the Best Explanation can become--it is not a single kind of reasoning, but an amalgamation of reasonings, methods, ideas into one process. There is some good news, however: we can take comfort in the apparent successes of science, of the humanities, and of philosophy. If this complex, messy, sometimes convoluted way of reasoning (IBE) is correct, and if it accurately describes how researchers often reason, then clearly they have been implementing IBE well enough without our help. It seems to me that, on some intuitive level, human beings, when they are being careful,

[52] I borrow the term "causal triangulation" from Peter Lipton. (Ibid)

[53] For instance, some philosophers think that IBE is really just a useful, more intuitive way of doing complex calculations under the Bayesian probability calculus. This seems to be the view of Peter Lipton and Tim Mcgrew. Lipton writes, "We are not good at probabilistic reasoning, so we use other methods or heuristics… one heuristic, inference to the best explanation, in part is a way of helping us to respect the constraints of Bayes' theorem." (Lipton 112).

Mcgrew, in the same paper, also raises the problem of disentangling and clarifying all the different best-making features, and how those relate to probability and confirmation, as I have mentioned above. Mcgrew writes: " But when explanationists try to spell out inference to the best explanation (IBE) in a more precise form and show how it provides a distinctive form of non-deductive inference, they find themselves hard pressed. The standard model cites various 'theoretical virtues' such as simplicity, consilience and precision that make some hypotheses more 'lovely' than others… But in many cases it is difficult to determine in a principled way whether a given virtue makes its contribution to the prior probability of a hypothesis or to its likelihood; there is no straightforward sense in which those categories match the virtues." (Mcgrew, Ibid.)

and reasoning from within careful communities (read: communities which carry out some form of peer review), get this substantially right. In fact, this entire section has been an act of *analytic* philosophy. I have taken what I think are intuitive concepts and methods already in play, analyzed them, clarified them, and given them a structure in a model. It is conceivable that, even without an analysis such as this, human beings would still make theoretical progress. Messy progress, of course, and progress that is not linear, but progress nonetheless. The use of a structural model is to clarify, to understand more deeply, and to train others in the good habits (the intellectual virtues) that have contributed to our success. It also provides us with a deeper understanding of ourselves, of our methods, and of our practices. So rest assured that, despite how complex these topics are, they need not *paralyze* us intellectually. In fact, it is precisely because I have found myself paralyzed, and am now free from such paralysis, that I write this to you now.

Simplicity and explanatory scope, as well as the details of how confirmation (conditional probabilities) works, are left for the Parts II and III of this book. The topics are complex, and will involve some level of symbolic, quasi-mathematical reasoning (which is very off-putting to most people). But for now, let us sum up what we've covered with a discussion of the process of inquiry--how we inquire, or seek answers to questions, over time.

## Section V: The Process of Inquiry

Introductory science texts often describe the special effectiveness of the "scientific method." As STEM students mature and learn more about their field, they tend to be disappointed that this highly idealized model of inquiry does not accurately describe their own practices. Here, we are going to cover the process of inquiry--which includes the "scientific method"--in light of what we have covered in the previous sections.

When we inquire, we are trying to find and understand the truth. But inquiry does not only involve straightforward argumentation. For inquiry proceeds through question-asking. We ask questions, and seek answers to them; the hope of philosophers, scientists, scholars and laypeople is that, by asking and answering questions, we eventually get to the truth (or something approximating the truth).[54] Inquiry thus involves explanations, predictions, and IBE in addition to straightforward, deductive argumentation. And we engage in these different kinds of reasoning over a period of time.

How, though, do we inquire? We have given details on the various parts of inquiry, but how does this process unfold over time? How does inquiry *actually look* in real life?

In what follows, I give one model for understanding how we inquire: a Peircean model, which I cover in the Advanced Readings section in detail, though with slight modifications.[55] However, this is an idealized model (though not as highly idealized as the "scientific method"). As such, it gives a false impression of linearity. In reality, inquiry will proceed in haphazard ways (for even scientists and scholars tend not to be trained in confirmation theory--if you get through this book, consider yourself a member of a very small class of lucky people). And, given the discussion of the constant need to revise our conclusions derived from IBE in Section IV, inquiry will involve a back and forth, revisionary motion, with multiple applications of IBE at every stage. Worse, when we complete all the steps of inquiry described below, we can easily find new evidence, and be required to start the process all over again, potentially reaching radically different conclusions. And on and on, until we die (or get bored).

Overall, the steps of inquiry on this model are as follows:

1. Observation + Puzzlement
2. Question Asking
3. The Proliferation of Hypotheses (Theorization)
4. Evaluation
    a. Testing
        i. Narrowing via Disconfirmation
        ii. Probabilistic Confirmation and Disconfirmation (positive and negative evidence)
    b. Inference to the Best Explanation
        i. Weighing of evidence, simplicity, relevance, coherence
5. Tentative Conclusion

---

[54] Approximation of the truth is, again, complex; see (Duhem, Pierre. *To Save the Phenomena,* 47-52.)

[55] For a historical overview of C.S. Peirce's, see the article attached in the advanced readings section. Dewey, a student of Peirce, by 1910 gives a very similar model to the one laid out here. He says: "Upon examination, each instance reveals, more or less clearly, five logically distinct steps : (i) a felt difficulty ; (ii) its location and definition ; (iii) suggestion of possible solutions; (iv) development by reasoning of the bearings of the suggestion ; (v) further observation and experiment leading to its acceptance or rejection; that is, the conclusion of belief or disbelief. (*How We Think*, Ch. VI).

We begin inquiry with sensory experience and observation. In experiences over time, we encounter various puzzling events and subjects, subjects that prompt us to ask "why?" questions. These puzzling subjects need not be just puzzling *events*, or puzzling *experiences*. Many, if not most, puzzling subjects are events and experiences in the physical, material world. However, philosophers often discover puzzling subjects that do not clearly involve material events or empirical observations. For instance, a philosopher might ask questions like,"just *what is* the number two?" "What makes something valuable?" Or, as we have been discussing, "what is explanation? How does it work?" These are not necessarily *causal* questions about the physical, observable world. Still, the philosopher concerned with abstract questions begins asking these questions *as a result of experience.* Human beings are fundamentally reactive: we have certain experiences, think about those experiences, and find various aspects of them puzzling. A philosopher may be asking about the nature of logic, ethics or explanation--these are not subjects they experience directly--precisely because they find themselves, in experience, using certain concepts, reasoning in certain ways, believing certain things. We experience ourselves reasoning, and then inquire into reasoning. We experience ourselves as having moral beliefs, and then inquire into the nature of those moral beliefs. Yet, **all inquiry, all question asking, begins with experience**, because puzzlement begins with experience, even if the puzzling subjects are not apprehended directly in experience.

Once we have found some puzzling subject, and **formulate a question** about that subject, we then begin to **create a host of potential explanations** (theories/hypotheses). If we are honest, and avoid hyperfocusing on one single explanation, we will find that our hypotheses proliferate (as discussed in Section III): we can create many equally powerful, potentially equally relevant, hypotheses. And so, we will not be able to select the "correct" or even "most plausibly correct" explanatory hypothesis.

How then are we to proceed past the proliferation stage? How are we to select the best, most probable, most likely correct, hypothesis (explanation or prediction)? We must figure out some way of differentiating between multiple equally powerful and (from our limited perspective) relevant theories.

This task will come in a few stages. We can divide them up for clarity, or lump them into one big process of IBE. That is, we might conceive of this last, evaluative step as IBE, with IBE containing testing and evaluation of intrinsic probability, alongside evaluation of explanatory concerns (relevance, simplicity, scope). Does it matter? Probably not. Think of it this way: we can conceive of IBE as itself a long, complex process of reasoning. The logic of probability, testing, and weighing all occur within, and together constitute, this complex process. IBE, in this sense, is not an argument, nor a discrete kind of reasoning: it is a compound, complex process made of many kinds of reasoning. Of course, IBE can also be spoken of as an argument, or a family of different kinds of arguments, for the conclusion that some H is 'the best' or 'most probable' or 'most lovely,' etc. These are not incompatible ways of speaking about IBE--IBE as a process will contain IBE as an argument. Unfortunately, many philosophers focus on IBE as a straightforward argument, with a single argumentative form/structure, and lose sight of the larger picture. Let us allow for these various ways of using words, and call this step the "**evaluation**" step.

First, we will *test* each theory. **Theoretical testing** involves disconfirmation and confirmation: we will look for evidence other than the explanandum that affects the probability of our hypotheses. Most commonly, we will do this by using our theories to make theoretical predictions. That is, we will draw out the implications or predictions of each of our hypotheses. Then, we will check if those implications--the predictions each theory makes--obtain. If H1 makes a prediction that E1 will occur with 100% confidence, and E1 does not occur, then we can rule out H1 (P3c). This allows us to *narrow down* the number of competing hypotheses by *ruling some of them out.*

Next, we continue to collect more data, and we will **look for other predictions and implications of each theory**, and, to the best of our abilities, check whether these predictions are successful or unsuccessful. Depending on how strongly H1 predicts E, E obtaining or failing to obtain will increase or decrease the probability of H1 (P2b-P3c). We examine all the successful and failed implications of each theory, and take these successes and failures as evidence for each theory. Further, there may be independent pieces of evidence that affect the probability of our hypotheses without counting as successful or failed predictions; these pieces of evidence should be taken into account here, if they exist. When we have done the best we can, we take our entire body of evidence (E1...En), and see how much that evidence confirms or disconfirms each theory (H1...Hn)

However, we quickly run into the problem of proliferation again: even after looking at all the evidence, we will still be left with multiple equally confirmed *and* equally powerful theories. And, because we are not sure *which* hypothesis (if any) specifies the *real* reasons why E occurred, we won't be able to rule out others for being "causally" (reasons) irrelevant. For if we could know which hypothesis specified the *real* reasons for E, then we would already know which hypothesis was correct.

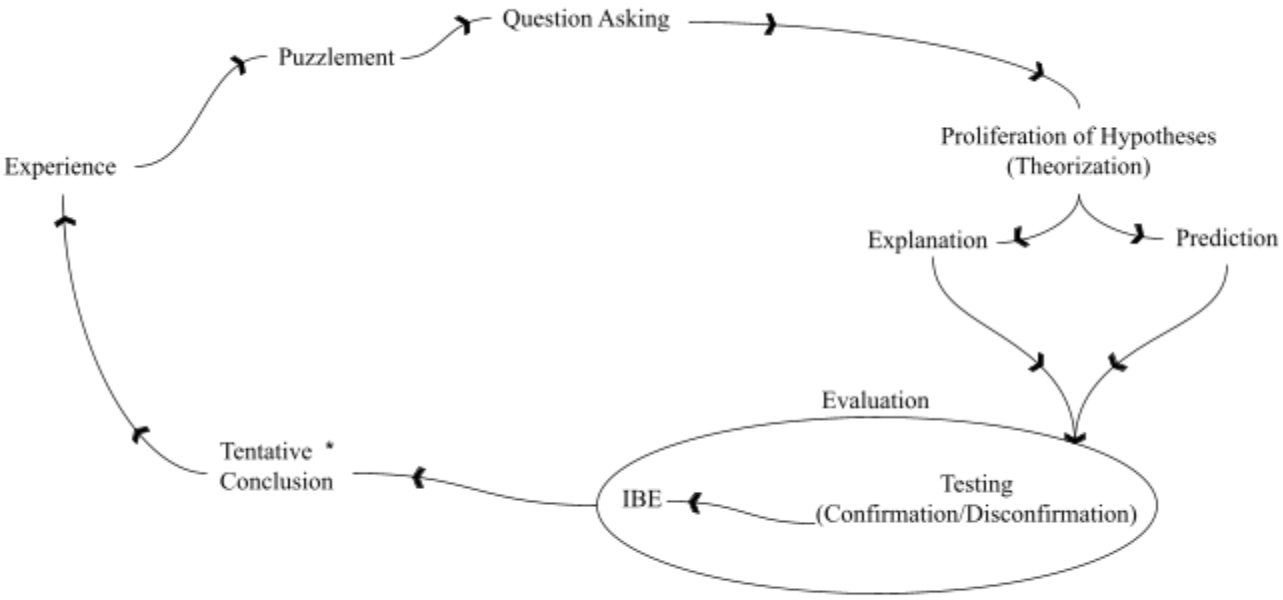
So, we will be left with our final step of inquiry: to appeal to simplicity and explanatory scope, weighing them against one another. We will integrate all the information we've gained prior--information about statistical relevance, reasons relevance, confirmation and disconfirmation--and combine it with considerations of simplicity and explanatory scope. That is to say, we will now combine all our *extrinsic* probability information (confirmation/disconfirmation) with information about the *intrinsic* probabilities of each hypothesis (simplicity and explanatory scope). With all this information in hand, we commence **weighing all the various best-making features (IBE)**, and see if this process of weighing results in one hypothesis emerging as the most probably correct explanation. Of course, other kinds of IBE may be required for our goals, and I do not wish to rule them out as part of the process of inquiry.

However, our job is not done--it is never done. IBE does not give us a final answer. Instead, IBE concludes by selecting a single hypothesis (or a number of hypotheses) as the most probable *given our current evidence,* and merely *in comparison to all competing hypotheses.* It is highly likely that someone will come up with a new hypothesis or find some new evidence, thereby forcing us to restart the testing and IBE processes. It is crucial to keep this in mind. As Bas van Fraassen reminds us,[56] IBE is very limited. Our conclusions reached via IBE are extraordinarily modest: "H1 is the best/most plausible theory among its known competitors, H2...H*n*." However, the conclusion, "we should believe H1," or "H1 is probable overall," is something more than this. So, even if IBE works, it cannot get us to full, unconditional belief in a theory, only a belief that this theory is best/most probable out of its competitors. That is why I term this step in any inquiry, the formation of a **tentative conclusion.** \

Inquiry, then, involves constant revision, tweaking theories, bodies of evidence, and our evaluation of theories as we expand our observations of the world. This may sound like a burden--and it is--but the revisable, continual nature of inquiry is a powerful failsafe against complacency, and prevents inquirers from becoming too attached to any specific theory. This ever-seeking, ever-updating activity gives science it's power to make progress and revise its current views….

[56] (Van Fraassen, Bas. "Inference to the Best Explanation: Salvation by Laws?")

**The Process of Inquiry**




** Hypothesis H is the best explanation/prediction out of its current competitors, given our current body of evidence.*

**Concluding Remarks:**

Where are we left, then, in our study of inquiry? Students should now have some understanding, albeit a rough understanding, of the process and nature of inquiry: seeking truth by asking questions and pursuing answers via explanation and prediction over time. We have also looked at the basics of confirmation theory, and described some principles of confirmation: principles which tell us when, and to what degree, evidence increases the probability of some hypothesis or proposition H. Finally, we have offered a brief sketch of a powerful tool for understanding the selection of the most probably correct explanation--Inference to the Best Explanation--and have signaled to students the areas where work remains to be done. This should, I hope, give students a better picture of how to properly inquire into truth, especially truth that is not directly observable or verifiable, without being too technical.

But again, there is much work to be done. We are left wanting more details, for instance, on the precise nature of the best-making features, and their relationship to probability.

- How can we disentangle all the overlapping best-making features, and understand their impact on probability precisely?
- Why does the best overall balance of best-making features guarantee or indicate that a hypothesis is the most probably correct?
- How do we precisely calculate conditional probabilities? Can we do so?
- How are we to evaluate the probability of *predictions*, rather than explanations? (No method has been given for this yet.)
- Just what is "probability?"
- What is the relationship between noetic-confidence probabilities and statistical probabilities?

- Is it possible to quantify or calculate noetic-confidence probabilities, as we can with statistical probabilities?

So many questions to work on, and so little time. However, you are now in a position to understand what these questions mean, and to see their interrelations. This puts you far ahead of many of your peers, and even most scholars, who have been unintentionally kept apart from the direct study of inquiry. In the following sections, we will develop these ideas more fully, though many questions will, necessarily, remain unanswered, or not fully answered. For *no one* has reached a final understanding of these issues, as of yet.

# Bibliography

Dewey, John. *How We Think*. 1st ed. D.C. Heath and Co, 1910. https://www.gutenberg.org/files/37423/37423-h/37423-h.htm

Dykhuizen, George. *The Life and Mind of John Dewey.* Southern Illinois Univ. Press, 1973.

Fann, K. T. *Peirce's Theory of Abduction*. The Hague: Martinus Nijhoff (1970).

Frankfurt, Harry G. "Peirce's Notion of Abduction." Journal of Philosophy 55 (1958): 593-597.

Goodman, Nelson. *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press, 1983.

Griffiths, P. E. "The Historical Turn in the Study of Adaptation." The British Journal for the Philosophy of Science 47, no. 4 (1996): 511-32.

Harman, Gilbert. "The Inference to the Best Explanation." Philosophical Review 74 (1965): 88-95.

Hempel, Carl. "Explanation in Science and History," In *Frontiers of Science and Philosophy,* ed. R.G. Colodny. London: Allen and Unwin and University of Pittsburgh Press, 1962.

Hempel, Carl G. and Paul Oppenheim, "Studies in the Logic of Explanation", *Philosophy of Science*, 15(2): 135–175. Reprinted in Hempel 1948 (1965): 245–290.

Hickman, Cleveland P., Susan L. Keen, David J. Eisenhour, Allan Larson, and Helen I'Anson. *Integrated Principles of Zoology*. New York, NY: McGraw Hill, 2024.

Mcauliffe, William H. B. "How did Abduction Get Confused with Inference to the Best Explanation?" Transactions of the Charles S. Peirce Society 51 (2015): 300-319.

McGrew, T., 2003. "Confirmation, Heuristics, and Explanatory Reasoning," British Journal for the Philosophy of Science, 54: 553–567.

Patten, Steven C. "Carl Hempel: Explanations by Reasons." *Canadian Journal of Philosophy,* Vol. II, No. 4 (1973).

Peirce, C.S. "On the Logic of Drawing History from Ancient Documents." Vol. 2, in *The Essential Peirce*, edited by the Peirce Edition Project. Bloomington: Indiana University Press, 1998.

Peirce, C.S. "Pragmatism as the Logic of Abduction." Vol. 2, in *The Essential Peirce*, edited by the Peirce Edition Project. Bloomington: Indiana University Press, (1903) 1998.

Popper, Karl. *The Logic of Scientific Discovery.* London:

Routledge Classics, 2002.

Salmon, Wesley. *Statistical Explanation and Statistical Relevance*, Pittsburgh, PA: University of Pittsburgh Press, 1971.

Salmon, Wesley. "Causality and Explanation: A Reply to Two Critiques", *Philosophy of Science*, 64(3): 461–477.

Swinburne, Richard. Simplicity as Evidence of Truth. Milwaukee: Marquette University Press, 1997

Swinburne, Richard. *Introduction to Confirmation Theory*. London: Methuen & CO LTD, 1973.

Swinburne, Richard (ed). *Bayes Theorem*. Oxford University Press, 2002.

Swinburne, Richard. *Epistemic Justification*. Oxford University Press, 2001.

van Fraassen, Bas. "Inference to the Best Explanation: Salvation by Laws?" In *Laws and Symmetry,* Ch. 6. Oxford: Clarendon Press, 1989.

Veen, Peter van der, Robert Deutsch and Gabriel Barkay. "Reconsidering the Authenticity of the Berekhyahu Bullae: A Rejoinder." *Antiguo Oriente* 14 (2016): 99-136.

Walton, Douglas. *Abductive Reasoning.* University of Alabama Press, 2005.

Wong, Julia C. "Qanon Explained: The Antisemitic Conspiracy Theory Gaining Traction around the World." The Guardian. Guardian News and Media, August 25, 2020.