Chapter 3

Induction and Innateness

One of the deepest problems in philosophy concerns how we learn about the world, and whether there are right or wrong ways to go about it. In this chapter I introduce this problem—the "problem of induction"—and describe its relevance to understanding learning in intelligent agents, and brains in particular. One consequence of the problem of induction is that there can be no such thing as a universal learning machine; it is not even possible that brains could enter the world as blank slates equipped with universal learning algorithms. The goal of the chapter is to provide a kind of solution to the problem of induction, and also to put forth something I call a theory of innateness. The latter would be a mathematical framework in which we are able to make sense of the kinds of structures that must be innately generated in a brain in order for that brain to have its own innate way of learning in the world. I present a theory called *Paradigm Theory* (Changizi and Barber, 1998) that purports to do these things.

What is induction?

"John is a man. All men are mortal. Therefore, John is mortal." This argument from two premises to the conclusion is a *deductive* argument. The conclusion logically follows from the premises; equivalently, it is logically impossible for the conclusion not to be true *if* the premises are true. Mathematics is the primary domain of deductive argument, but our everyday lives and scientific lives are filled mostly with another kind of argument.

Not all arguments are deductive, and 'inductive' is the adjective labelling any non-deductive argument. Induction is the kind of argument in which we typically engage. "John is a man. Most men die before their 100th birthday.

Therefore John will die before his 100th birthday." The conclusion of *this* argument can, in principle, be false while the premises are true; the premises do not logically entail the conclusion that John will die before his 100th birthday. It nevertheless is a pretty good argument.

It is through inductive arguments that we learn about our world. Any time a claim about infinitely many things is made on the evidence of only finitely many things, this is induction; e.g., when you draw a best-fit line through data points, your line consists of infinitely many points, and thus infinitely many claims. Generalizations are kinds of induction. Even more generally, any time a claim is made about more than what is given in the evidence itself, one is engaging in induction. It is with induction that courtrooms and juries grapple. When simpler hypotheses are favored, or when hypotheses that postulate unnecessary entities are *dis*favored (Occam's Razor), this is induction. When medical doctors diagnose, they are doing induction. Most learning consists of induction: seeing a few examples of some rule and eventually catching on. Children engage in induction when they learn the particular grammatical rules of their language, or when they learn to believe that objects going out of sight do not go out of existence. When rats or pigeons learn, they are acting inductively. On the basis of retinal information, the visual system generates a percept of its guess about what is in the world in front of the observer, despite the fact that there are always infinitely many ways the world could be that would lead to the same retinal information—the visual system thus engages in induction.

If ten bass are pulled from a lake which is known to contain at most two kinds of fish—bass and carp—it is induction when one thinks the next one pulled will be a bass, or that the probability that the next will be a bass is more than $1/2$. Probabilistic conclusions are still inductive conclusions when the premises do not logically entail them, and there is nothing about having fished ten or one million bass that logically entails that a bass is more probable on the next fishing, much less some specific probability that the next will be a bass. It is entirely possible, for example, that the probability of a bass is now *decreased*—"it is about time for a carp."

What the problem is

Although we carry out induction all the time, and although all our knowledge of the world depends crucially on it, there are severe problems in our understanding of it. What we would *like* to have is a theory that can do the following. The theory would take as input (i) a set of hypotheses and (ii) all the evidence

known concerning those hypotheses. The theory would then assign each hypothesis a probability value quantifying the degree of confidence one logically *ought* to have in the hypothesis, given all the evidence. This theory would interpret probabilities as *logical probabilities* (Carnap, 1950), and might be called a theory of logical induction, or a theory of logical probability. (Logical probability can be distinguished from other interpretations of probability. For example, the *subjective* interpretation interprets the probability as how confident a person actually is in the hypothesis, as opposed to how confident the person ought to be. In the *frequency* interpretation, a probability is interpreted roughly as the relative frequency at which the hypothesis has been realized in the past.)

Such a theory would tell us the proper method in which to proceed with our inductions, i.e., it would tell us the proper "inductive method." [An *inductive method* is a way by which evidence is utilized to determine *a posteriori* beliefs in the hypotheses. Intuitively, an inductive method is a box with evidence and hypotheses as input, and *a posteriori* beliefs in the hypotheses as output.] When we fish ten bass from the lake, we could use the theory to tell us exactly how confident we should be in the next fish being a bass. The theory could be used to tell us whether and how much we should be more confident in simpler hypotheses. And when presented with data points, the theory would tell us which curve ought to be interpolated through the data.

Notice that the kind of theory we would like to have is a theory about what we *ought* to do in certain circumstances, namely inductive circumstances. It is a *prescriptive* theory we are looking for. In this way it is actually a lot like theories in ethics, which attempt to justify why one ought or ought not do some act.

Now here is the problem: *No one has yet been able to develop a successful such theory!* Given a set of hypotheses and all the known evidence, it sure *seems* as if there is a single right way to inductively proceed. For example, if all your data lie perfectly along a line—and that is *all* the evidence you have to go on—it seems intuitively obvious that you should draw a line through the data, rather than, say, some curvy polynomial passing through each point. And after seeing a million bass in the lake—and assuming these observations are all you have to help you—it has just *got* to be right to start betting on bass, not carp.

Believe it or not, however, we are still not able to defend, or justify, why one really ought to inductively behave in those fashions, as rational as they seem. Instead, there are multiple inductive methods that seem to be just as

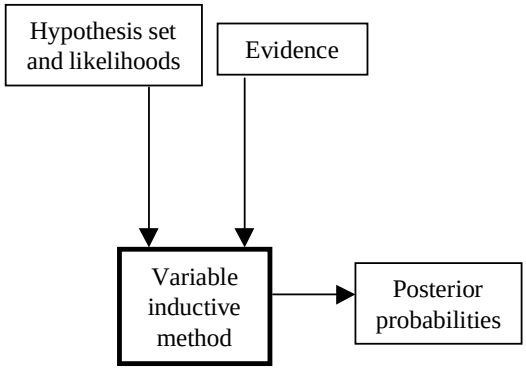


Figure 3.1: The purpose of an inductive method is to take a set of hypotheses and the evidence as input, and output the degree to which we should believe in each hypothesis in light of the evidence, i.e., output the posterior probability distribution over the set of hypotheses. Inductive methods may, in principle, be any function from the hypotheses and evidence to a posterior probability distribution, but some inductive methods seem better than others. Which one ought we use? That is the riddle of induction. An ideal answer would be a theory of logical probability that tells us, once and for all, which inductive method to use. But there is no such ideal theory.

good as one another, in terms of justification. Figure 3.1 depicts the problem. The hypothesis set and evidence need to be input into some inductive method in order to obtain beliefs in light of the evidence. But the inductive method is, to this day, left variable. Different people can pick different inductive methods without violating any mathematical laws, and so come to believe different things even though they have the same evidence before them.

But do we not use inductive methods in science, and do we not have justifications for them? Surely we are not picking inductive methods willy nilly! In order to defend inductive methods as we actually use them today, we make *extra* assumptions, assumptions going beyond the data at hand. For example, we sometimes simply assume that lines are more *a priori* probable than parabolas (i.e., more probable *before* any evidence exists), and this helps us conclude that a line through the data should be given greater confidence than the other curves. And for fishing at the lake, we sometimes make an *a priori* assumption that, if we pull n fish from the lake, the probability of getting n bass and no

carp is the same as the probability of getting $n - 1$ bass and one carp, which is the same as the probability of getting $n - 2$ bass and two carp, and so on; this assumption makes it possible (as we will see later) to begin to favor bass as more and more bass, and no carp, are pulled from the lake. Making different *a priori* assumptions would, in each case, lead to different inductive methods, i.e., lead to different ways of assigning inductive confidence values, or logical probabilities, to the hypotheses.

But what justifies our making these *a priori* assumptions? That's the problem. If we had a theory of logical probability—the sought-after kind of theory I mentioned earlier—we would not have to make any such undefended assumption. We would know how we logically ought to proceed in learning about our world. By making these *a priori* assumptions, we are just *a priori* choosing an inductive method; we are not bypassing the problem of justifying the inductive method.

I said earlier that the problem is that "no one has yet been able to develop a successful such theory." This radically understates the dilemma. It suggests that there could really *be* a theory of logical probability, and that we have just not found it yet. It is distressing, but true, that there simply *cannot* be a theory of logical probability! At least, not a theory that, given only the evidence and the hypotheses as input, outputs the degrees of confidence one really "should" have. The reason is that to defend any one way of inductively proceeding requires adding constraints of some kind—perhaps in the form of extra assumptions—constraints that lead to a distribution of logical probabilities on the hypothesis set even *before* any evidence is brought to bear. That is, to get induction going, one needs something equivalent to *a priori* assumptions about the logical probabilities of the hypotheses. But how can these hypotheses have degrees of confidence that they, *a priori*, simply *must* have. Any theory of logical probability aiming to once-and-for-all answer how to inductively proceed must essentially make an *a priori* assumption about the hypotheses, and this is just what we were hoping to avoid with our theory of logical probability. That is, the goal of a theory of logical induction is to explain why we are justified in our inductive beliefs, and it does us no good to simply assume inductive beliefs in order to explain other inductive beliefs; inductive beliefs are what we are trying to explain.

Bayesian formulation of the problem

We have mentioned probabilities, but it is important to understand a simple, few-centuries-old theorem of Bayes. Using Bayes' Theorem it will be possible to understand inductive methods more deeply. As set up thus far, and as depicted in Figure 3.1, the inductive method is left entirely variable. Any way of using evidence to come to beliefs about hypotheses can fill the 'inductive method' role. Different inductive methods may utilize evidence in distinct ways to make their conclusions. Bayes' Theorem allows us to lay down a fixed principle dictating how evidence should modify our beliefs in hypotheses. The variability in inductive methods is constrained; inductive methods cannot now differ in regards to how evidence supports hypotheses. As we will see, the Bayesian framework does not dictate a single unique inductive method, however; the variability is pushed back to prior probabilities, or the degrees of confidence in the hypotheses before having seen the evidence. Let me first explain Bayes' Theorem and the framework.

First, the Bayesian framework is a probabilistic framework, where degrees of confidence in hypotheses are probabilities and must conform to the axioms of Probability Theory. The axioms of probability are these: (i) Each probability is in the interval [0,1]. (ii) The sum of all the probabilities of the hypotheses in the hypothesis set must add to 1. (iii) The probability of no hypothesis being true is 0. And (iv), the probability of two possibilities A and B is equal to the sum of their individual probabilities minus the probability of their co-occurrence. We will be assuming our hypotheses in our hypothesis sets to be mutually exclusive, and so no two hypotheses can possibly co-occur, making axiom (iv) largely moot, or trivially satisfied for us.

Suppose the probability of event A is $P(A)$, and that for event B is $P(B)$. What is the probability of *both* A and B. We must first consider the probability that A occurs, $P(A)$. Then we can ask, given that A occurs, what is the probability of B; this value is written as $P(B|A)$. The probability of A and B occurring is the product of these two values. That is, we can conclude that

$$P(A\&B) = P(A) \cdot P(B|A).$$

But note that we could just as well have started with the probability that B occurs, and then asked, given that B occurs, what is the probability of A. We would then have concluded that

$$P(A\&B) = P(B) \cdot P(A|B).$$

The right hand sides of these two equations differ, but the left hand sides are the same, so we may set them equal to one another, resulting in

$$P(A) \cdot P(B|A) = P(B) \cdot P(A|B).$$

This is essentially Bayes' theorem, although it is usually manipulated a little.

To see how it is usually stated, let us change from A and B to h and e, where h denotes some hypothesis, and e denotes the evidence. The formula now becomes

$$P(h) \cdot P(e|h) = P(e) \cdot P(h|e).$$

What do these values mean?

- $P(h)$ stands for the probability of hypothesis h *before* any evidence exists. It is called the *prior probability* of h. Each hypothesis might have its own distinct prior probability.
- $P(h|e)$ is the probability of hypothesis h *after* the evidence has been considered; it is the hypothesis' probability given the evidence. Accordingly, it is called the *posterior probability* of h. Each hypothesis might have its own distinct posterior probability.
- $P(e|h)$ is the probability of getting the evidence *if* hypothesis h were true. It is called the *likelihood*. Each hypothesis might have its own distinct likelihood, and its likelihood is usually determinable from the hypothesis.
- $P(e)$ is the probability of getting that evidence. This value does not depend on the hypothesis at issue. It may be computed from other things above as follows:

$$P(e) = \sum_h [P(h)P(e|h)].$$

Ultimately, the value that we care about most of all is $P(h|e)$, the posterior probability. That is, we want to know how much confidence we should have in some hypothesis given the evidence. So, let us solve for this term, and we get a formula that is the traditional way of expressing Bayes' Theorem.

$$P(h|e) = \frac{P(h) \cdot P(e|h)}{P(e)}.$$

Since $P(e)$ does not depend on which hypothesis is at issue, it is useful to simply forget about it, and write Bayes' Theorem as

$$P(h|e) \sim P(h) \cdot P(e|h).$$

That is, the posterior probability is proportional to the prior probability times the likelihood. This makes intuitive sense since how much confidence you have

in a hypothesis should depend on both how confident you were in it before the evidence—the prior probability—and on how much that hypothesis is able to account for the evidence—the likelihood.

Using the evidence to obtain posterior probabilities is the aim of induction. Figure 3.2 shows the material needed to obtain posterior probabilities within the Bayesian framework. As in Figure 3.1, the hypothesis set (along with the likelihoods) and the evidence are inputs to the inductive method (which may be of many different kinds, and is thus variable), which outputs posterior probabilities. But now the inductive method box has some boxes within it; inductive methods are now determined by variable prior probability distributions and the fixed Bayes' Theorem.

Consider an example first. I present to you a coin, and tell you it is possibly a trick coin. I tell you that there are three possibilities: it is fair, always-heads, or always-tails. These three possibilities comprise the hypothesis set. Your task is to flip the coin and judge which of these three possibilities is true. Your evidence is thus coin flip outcomes. Your likelihoods are already defined via the decision to consider the three hypotheses. For example, suppose two heads are flipped. The likelihood of getting two heads for the coin-is-fair hypothesis is $(1/2)^2 = 1/4$. The likelihood for the always-heads hypothesis is $1^2 = 1$, and for the always-tails hypothesis it is $0^2 = 0$. What is the posterior probability for these three hypotheses given that the evidence consists of the two heads? To answer this, we still need prior probability values for the hypotheses. This is where things get hairy. In real life, we may have experience with tricksters and coins with which we can make guesses as to the prior (i.e., the prior probability distribution). But the point of this example is to imagine that you have no experience whatsoever with tricksters or coins, and you somehow need to determine prior probabilities for these three hypotheses. Let us suppose you declare the three to be equally probable, *a priori*. Now you can engage in induction, and the posterior probabilities are as follows:

- $P(\text{fair}|\text{two heads}) = \frac{\frac{1}{3}\cdot\frac{1}{4}}{\frac{1}{3}\cdot\frac{1}{4}+\frac{1}{3}\cdot 1+\frac{1}{3}\cdot 0} = \frac{1}{5}$.
- $P(\text{always-heads}|\text{two heads}) = \frac{\frac{1}{3}\cdot 1}{\frac{1}{3}\cdot\frac{1}{4}+\frac{1}{3}\cdot 1+\frac{1}{3}\cdot 0} = \frac{4}{5}$.
- $P(\text{always-tails}|\text{two heads}) = \frac{\frac{1}{3}\cdot 0}{\frac{1}{3}\cdot\frac{1}{4}+\frac{1}{3}\cdot 1+\frac{1}{3}\cdot 0} = 0$.

Different prior probability assignments would have led to different posterior probability assignments; i.e., led to different inductive conclusions.

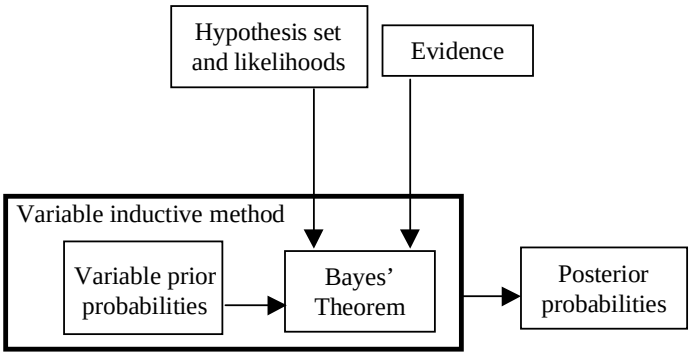


Figure 3.2: To acquire beliefs about the world, evidence and a set of hypotheses must be input into an inductive method, whose job it is to output the degrees of belief about those hypotheses one ought to have given the evidence. In the Bayesian framework, the inductive method is determined by a choice of prior probabilities over the hypothesis set. This variable prior is put, along with the evidence, into Bayes' Theorem, which outputs the posterior probabilities. Now it is not the case that any old inductive method is justified, unlike in Figure 3.1. However, there is still tremendous variability in the possible inductive methods due to the variability in the choice of prior. One of the nice things about this is that the variability no longer concerns how evidence is brought to bear on hypotheses; this is kept constant by the use of Bayes' Theorem. All the variability in inductive methods is reduced to just one kind of thing: one's degrees of belief in the hypotheses before having seen the evidence. Also, note that the Bayesian framework is also a probabilistic framework, which constrains the numerical degrees of confidence in hypotheses to satisfy the axioms of Probability Theory; this constraint is not depicted in the figure.

What does the Bayesian framework for induction buy us? After all, we still have many possible inductive methods to choose from; we have not solved the problem of the variability, or indeterminacy, of inductive methods. For one thing, it rules out whole realms of possible inductive methods; inductive methods must now fit within the framework. Algorithmic learning rules that take evidence and assign probabilities to the hypotheses are not allowable inductive methods if they cannot be obtained by starting with a prior probability distribution and grinding it through Bayes' Theorem. The second nice thing about the Bayesian framework is that it gets inside inductive methods and helps to distinguish between two things an inductive method needs in order to do its job: evidence principles and prior probabilities. Any inductive method needs "evidence principles," principles by which it employs the evidence to affect the degrees of confidence in the hypotheses. For example, if I fish one more bass, is this good or bad for the hypothesis that the next fish will be a bass? The Bayesian framework encapsulates its evidence principle in Bayes' Theorem, effectively declaring that all inductive methods must use this same evidence principle. Whatever variability in inductive method choice is left is not, then, due to differences in evidence principles. The second thing the Bayesian framework serves to distinguish is the prior probability distribution. This is left indeterminate, but any inductive method within the Bayesian framework requires some setting for this variable. All the variability in inductive methods is, then, reduced to one kind: one's *a priori* degrees of confidence in the hypotheses. A final important thing about the Bayesian framework is that the evidence principle is not just any old evidence principle; it is justifiable in the sense that it follows from probability axioms that everyone believes. Not only does "everyone believe" the probability axioms, they are, in a certain clear sense, principles a reasoner ought to hold. This is due to the fact that if someone reasons with numerical confidences in hypotheses that do not satisfy the probability axioms, then it is possible to play betting games with this fellow and eventually take all his money. This is called the *Dutch Book Theorem*, or the Ramsey-de Finetti Theorem (Ramsey, 1931; de Finetti, 1974; see also Howson and Urbach, 1989, pp. 75–89 for discussion). And Bayes' Theorem follows from these axioms, so this evidence principle is rational, since to not obey it would lead one to being duped out of one's money.[1]

[1]Things are actually a bit more complicated than this. Using Bayes' Theorem as our principle of evidence (or our "principle of conditionalization," as it is sometimes said) is *the* rational principle of evidence—i.e., in this case because any other will lead you to financial ruin—*if*, upon finding evidence e, e does not entail that your future degree of confidence in the hypoth-

With this machinery laid before us, the riddle of induction can now be stated more concisely as, "What prior probability distribution ought one use?" By posing induction within the Bayesian framework, one cannot help but see that to have a theory of logical induction would require a determinate "best" choice of prior probabilities. And this would be to make an *a priori* assumption about the world (i.e., an assumption about hypotheses concerning the world). But our original hope was for a theory of logical induction that would tell us what we ought to do *without* making *a priori* assumptions about the world.

What would a theory of logical probability look like?

There is, then, no solution to the riddle of induction, by which we mean there is no theory of logical probability which, given just a set of hypotheses and some evidence, outputs *the* respectable inductive method. There simply *is no* unique respectable inductive method.

If one tries to solve a problem, only to eventually realize that it has no solution, it is a good idea to step back and wonder what was wrong with the way the problem was posed. The problem of induction must be ill posed, since it has no solution of the strong kind for which we were searching. Let us now step back and ask what we want out of a theory of logical probability.

The Bayesian framework serves as a strong step forward. Within it, we may make statements of the form,

> If the prior probability distribution on H is $P(h)$, then, given the evidence, the posterior probability distribution ought to be given by $P(h|e)$, as dictated by Bayes' Theorem.

There are a number of advantages we mentioned earlier, but a principal downside remains. What the inductive method is depends entirely on the prior probability distribution, but the prior probability distribution comprises a set of beliefs about the degrees of confidence in the hypotheses. That is, prior probabilities are judgements about the world. Thus, the Bayesian statement becomes something like,

> If one has certain beliefs about the world before the evidence, then he should have certain other beliefs about the world after the evidence.

esis given e will be different from that given by Bayes' Theorem. That is, if, intuitively, the evidence does not somehow logically entail that Bayes' Theorem is inappropriate in the case at issue. This 'if' basically makes sure that some very weird scenarios are not occurring; no weird circumstances... Bayes' Theorem is the rational principle of evidence. See Howson and Urbach (1989, pp. 99–105) for details.

But one of the goals of a logical theory of induction is to tell us which beliefs about the world we ought to have. The Bayesian framework leaves us unsatisfied because it does not tell us which *a priori* beliefs about the world we should have. Instead, it leaves it entirely open for us to believe, *a priori*, anything we want!

In our move from the pre-Bayesian framework (Figure 3.1) to the Bayesian framework (Figure 3.2), we were able to encapsulate a fixed evidence principle, and were left with variable prior probabilities. Now *I submit that the task of a theory of logical probability is to put forth fixed principles of prior probability determination, and to have left over some variable that does* not *possess information about the world (unlike prior probabilities)*. Just as the left over variable in the Bayesian framework was non-evidence-based, the variable left over within this new framework will be non-induction-based, or non-inductive, or not-about-the-world. If we had something like this, then we could make statements like,

> If one has non-inductive variable Q, then one ought to have prior probability distribution $P_Q(h)$, as dictated by the principles of prior probability determination.

It should also be the case that the non-inductive variable has some coherent (non-inductive) interpretation, lest one not know how anyone would ever pick any value for it. The principles of prior probability determination would possess a few things that one ought to do when one picks prior probabilities given the non-inductive variable. In this way, we would have reduced all oughts found in induction to a small handful of principles of ought, and no undefended assumptions about the world would need to be made in order to get different inductive methods up and going.

Figure 3.3 is the same as Figure 3.2, but now shows the kind of machinery we need: (i) some fixed, small number of axioms of *a priori* logical probability determination, in the form of rationality principles, and (ii) some variable with a meaningful interpretation, but not with any inductive significance.

The bulk of this chapter consists of the development and application of a theory of logical induction aiming to fill these shoes. The theory is called *Paradigm Theory.* Three abstract symmetry-related principles of rationality are proposed for the determination of prior probabilities, and a kind of non-inductive variable—called a "paradigm"—is introduced which is interpreted as a conceptual framework, capturing the kinds of properties of hypotheses one acknowledges. A paradigm and the principles together entail a prior probability distribution; the theory allows statements of the form,

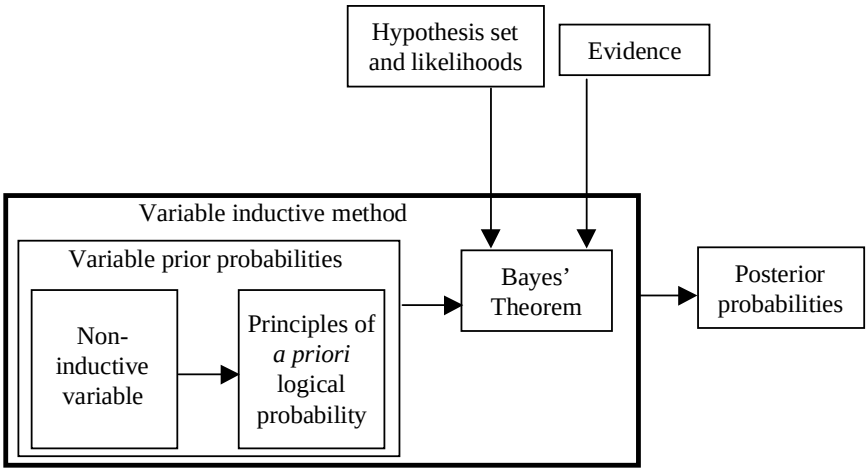


Figure 3.3: The structure of a sought-after theory of logical induction. The prior probability distribution should follow from the combination of a small number of rationality principles—things a rational agent ought to do—and some non-inductive variable with a meaningful interpretation.

> If one has paradigm Q, then one ought to have prior probability distribution $P_Q(h)$, as dictated by the symmetry-related principles of prior probability determination.

Innateness

The brain learns. It therefore entertains hypotheses, and implements inductive methods. What do we mean by a hypothesis in regards to the brain? Here are a couple examples. The human brain quickly learns the grammar of natural language, and there are (infinitely) many possible hypotheses concerning which grammar is the correct one. Kids eventually converge to the correct grammar; i.e., after sufficient accumulation of evidence, they impart the highest degree of belief to the correct (or nearly correct) grammatical hypothesis. Another example is vision. The retina is a two-dimensional sheet, and the world is three-dimensional, with many properties such as reflectance and object type. The information on the retina cannot uniquely specify the thing in the world that caused it, the reason being that there are infinitely many things in the world that may have caused it. Each possible cause of the retinal stimulus is a "perceptual hypothesis," and after acquiring experience in the world, upon presentation of a stimulus, the visual system typically finds one perceptual hypothesis to be more probable than the others, which is why we see just one scene at a time most of the time. When the probabilities are tied between two perceptual hypotheses, we jump back and forth between them, as in bistable stimuli such as the Necker Cube (which is just a line drawing of a cube, which can be seen in one of two orientations). These two examples for hypotheses entertained by the brain do not even scratch the surface; the brain indeed is a kind of learning machine, and thus entertains possible hypotheses at every turn.

Not only does the brain learn, but it is thought by many to enter the world a blank slate, and to be endowed with powerful and general learning abilities. One can get the impression that the cortex is some kind of universal learning engine. The relatively homogenous nature of the anatomy and connectivity of the cortex is one reason scientists come away with this impression: the cortex is a few millimeter thick sheet (its exact thickness depending on the brain's size), with six layers, and with statistically characterized connectivity patterns for the neurons within it. Roughly, the cortex seems to be built from many repeating units called "minicolumns." And although the cortex is divided up into distinct areas, connecting to other areas primarily via myelinated white matter axons, and although the areas often have distinguishing anatomical features,

they appear to be fairly similar in basic design. The high degree of plasticity of the cortex also suggests that it is a general learning machine, not a prisoner to instinct. When, for example, a limb is lost, somatosensory areas formerly devoted to the limb sometimes become employed by other areas. Also, the basic connectivity features of our cortex do not appear much different than that of monkey or cat, animals leading drastically different lives. The intuitive conclusion sometimes drawn is that we differ from monkeys merely in that our brains are relatively much larger, and that our ecologies and thus experiences are different. Our perception of the world appears to rely on general learning strategies by the visual system. People raised in non-carpentered environments, like Bushmen, do not experience some of the classical geometrical illusions that we find illusory (Segall et al., 1966). Even thirst and hunger, two appetitive states one might imagine would be innate if anything is innate, appear to be learned (Changizi et al., 2002b): rats do not know to orient toward a known water source when cellularly dehydrated unless they have experienced dehydration paired with drinking water, and similarly they do not know to orient toward a known food source when food restricted unless they have experienced food restriction paired with eating.

But being highly homogeneous and plastic does not entail that the brain does not possess innate content, or knowledge. Whatever is innate could well be wrapped up in the detailed connectivity patterns in the brain. And a strong role for experience does not mean the cortex is a universal learning machine. Even those scientists that are fans of a strongly innate brain, such as those that believe that grammar is innate (e.g., Chomsky, 1972; Pinker, 1994), obviously believe in an immense role for learning.

With an understanding of the riddle of induction under our belts, we can say, without knowing anything about the particulars of our brains, that we *must* enter the world with innate knowledge. There is no universal learning machine. There are, instead, just lots and lots of different inductive methods. Whatever our brains are doing when they learn, they are engaging in an inductive method (although perhaps a different inductive method for different kinds of learning). As discussed earlier, the brain must therefore, in effect, be making an assumption about the world in the form of a prior probability distribution over the possible hypotheses. That is, in order to learn, brains must enter the world with something equivalent to preconceptions about the degrees of confidence of all the possible hypotheses. Brains are not blank slates; they are born with what are, in effect, *a priori* assumptions about the world.

What would a theory of innateness be?

Brains, then, come furnished with an inductive method; i.e., they have some way by which they take evidence and determine the posterior probabilities of hypotheses. Different brain types—e.g., human versus cat—may employ different inductive methods, and these differences are innate. We will assume that the principal innate differences between brain types are due to their instantiating different inductive methods. (They may also differ in their choice of what hypotheses to consider in the first place, and they may well differ concerning what things matter to them (i.e., utilities).)

What I wish to consider here is a *theory of innateness*, a theory aimed at characterizing the nature of the information that must be innately generated. How much must innately differ between two kinds of brain (or two parts of the same brain) in order for them to possess distinct inductive methods? The theory of innateness I seek is not one that actually claims that brains conform to the theory. Rather, the aim is to construct a mathematical theory, or framework, within which we can conceptually distinguish among the kinds of structure required for an innate inductive method. With a theory of innateness in hand, we will *then* have the conceptual apparatus to begin to speak about the principles governing how brains—or any intelligent learning agents—have innate inductive methods.

Here is one thing that we would like out of a theory of innateness. I have already mentioned that brains of different kinds have a lot in common. It would accordingly be useful to find a theory of innateness that postulates no greater innate differences than are absolutely necessary to account for the different inductive methods used. We would like to be able to model brains of different types—i.e., brains employing different inductive methods—as following the same underlying principles, principles used in determining their inductive method. All these brains are, after all, brains, and the way they go about their learning should be describable using universal principles. Furthermore, these principles should be rationality principles of some kind, or principles stating what a rational agent would do. We would then be able to model brains having different innatenesses as nevertheless being similar to the extent that they follow the same rationality principles. We would be able to say that all these kinds of brains may be different in *some* regard that specifies what is innate, but that in all other ways we may model these brains identically.

The second aspect of our theory of innateness that requires concern is the distinguishing feature between brains of different types—the feature that is the

possessor of the innate information. There must be some variable property of brains, distinct settings of the variable which lead to distinct inductive methods used by the brain. As mentioned earlier, the theory of innateness for which we search would postulate no greater innate differences than are absolutely necessary to account for the different inductive methods used. Accordingly, we want the variable that determines the innate differences to be as weak as possible. Furthermore, innate content is derided by many because it seems absurd that, say, natural language grammar could be encoded into the brain at birth. Surely it is an incredible claim that brains enter the world with *a priori* beliefs, or assumptions. With this in mind, we would also like the "innateness variable" to say as little as possible about the world; i.e., to be non-inductive. Finally, this innateness variable should be interpretable in some plausible fashion; if it has no interpretation, then one begins to suspect that it is just a stand-in for an *a priori* inductive assumption.

In short, we would like a theory of innateness that models brains, or any intelligent agent, as following fixed principles of rationality in their learning, and models the differences with an innateness variable that is weak, non-inductive, and has a meaningful interpretation.

If you recall the earlier subsection on what we want out of a theory of logical probability, you will notice a close connection to that and to what we here want out of a theory of innateness. This is not a coincidence. The discovery of a theory of logical probability of the kind described would state how, through a fixed set of prior probability determination principles, a rational agent with a setting of the non-inductive variable should proceed in assigning his prior probabilities, and consequently what inductive method he ought to follow. On the one hand, the theory would tell us what we ought to do, but on the other hand, the theory tells us what a rational agent will, in fact, do—since this agent will do what he should. If our interest is in modeling innateness in assumed-to-be-rational brains, then the theory of logical probability can be used to describe brains, not just to say how brains ought to perform.

Let us go through the connection between a theory of induction and a theory of innateness more slowly, beginning with the earlier Figure 3.1. One way to treat innateness differences in different brain types is to postulate that they are governed by entirely different principles altogether. Brains of different types are, in regards to learning, just (computable) functions of any old kind taking evidence and outputting posterior probabilities. Each brain would innately make different assumptions about the world, and follow different rules concerning how evidence supports hypotheses (i.e., follow different evidence

principles). But we wish to be able to retain the view that brains and other intelligent agents learn in a rational fashion, and thus are all fundamentally similar, following identical principles, and differing only in regard to some small, interpretable, weak variable that does not correspond to an assumption about the world.

Bayesianism provides a great first step toward satisfying these demands, just as it provided a great first step for a theory of logical induction. Figure 3.2 was our corresponding figure for the problem of induction, and it is apt to look at it again for innateness. Bayes' Theorem is helpful toward a theory of innateness and learning because it allows us to treat all agents, or brains, as following Bayes' Theorem in their modification of their degrees of confidence in hypotheses in the light of evidence. And the Bayesian evidence principle is not just any old evidence principle, it seems to be the right principle—it is the way one should use evidence to modify the degree of confidence in hypotheses. This is why Bayesian approaches have been so popular in the psychological, brain and decision sciences. This Bayesian framework is used to model the visual system, memory, learning, behavior, economic agents and hosts of other cases where there is some kind of "agent" dealing with an uncertain world. The argument goes something like this: (a) These agents have probably been selected to learn in an optimal, or rational, manner. (b) The optimal learning manner is a Bayesian one. (c) Therefore, we may treat these agents as following Bayesian principles. The Bayesian framework also severely constrains the space of possible inductive methods, from anything-goes down to only those using its evidence principle.

As powerful as the Bayesian framework is, it leaves us with some residual dissatisfaction concerning a theory of innateness. The Bayesian framework has prior probabilities that differ between agents that follow different inductive methods. A prior probability distribution, then, is the innateness variable. Brains that differ in innateness would be postulated to enter the world with different *a priori* beliefs about the degree of confidence in the hypotheses. As discussed earlier, this is one thing we want to avoid with a theory of innateness. We would like it to be the case that innateness can be much more subtle than *a priori* beliefs about the world in the head. Perhaps there are further principles—principles beyond Bayes' Theorem—that an optimally engineered agent will follow, so that two such agents might innately differ in some non-inductive fashion, yet by following these fixed principles they come to have different prior probabilities. Figure 3.3 from earlier is again appropriate, for it shows what we are looking for. Such a theory would even further constrain the

space of possible inductive methods, from any-prior-probability-distribution-goes down to only those using the fixed principles of prior probability determination.

That is our goal for a theory of innateness. The theory that I will propose in the next section—called *Paradigm Theory*—consists of fixed symmetry and symmetry-like principles of rationality—or principles of non-arbitrariness—which I argue any rational agent will follow. The non-inductive variable is something I call a "paradigm", which is just the kinds of hypotheses the agent acknowledges; for example, you and I might possess the same hypothesis set, but I may carve it up into kinds differently than do you. The intuition is that we have different conceptual frameworks, or belong to different (Kuhnian) paradigms. Innateness differences, then, would be attributable to differences in the conceptual frameworks they are born with. But in all other regards agents, or brains, of different innatenesses would be identical, having been selected to follow fixed optimal, or rational, principles, both of prior probability determination and of evidence.

Are there really innate paradigms in the head? I don't know, and at the moment it is not my primary concern. Similarly, the Bayesian framework is widely considered a success, yet no one appears particularly worried whether there is any part of the developing brain that corresponds to prior probabilities. The Bayesian framework is a success because it allows us to model brains as if they are rational agents, and it gives us the conceptual distinctions needed to talk about evidence principles and *a priori* degrees of belief in hypotheses. Similarly, the importance of Paradigm Theory in regards to innateness will be that it allows us to model brains as if they are rational agents, giving us more conceptual distinctions so that, in addition to evidence principles and *a priori* degrees of belief in hypotheses, we can distinguish between principles of non-arbitrariness for prior probability determination and *a priori* conceptual frameworks. Paradigm Theory gives us the power to make hypotheses we otherwise would not be able to make: that brains and intelligent learners could have their innate inductive methods determined by innate, not-about-the-world paradigms, along with a suite of principles of rationality. Whether or not brains actually utilize these possibilities is another matter.

3.1 Paradigm Theory

In this section I introduce a theory of logical probability (Changizi and Barber, 1998), with the aim of satisfying the criteria I put forth in the previous section. The plan is that it will simultaneously satisfy the demands I put forward for a theory of innateness. The theory's name is "Paradigm Theory," and it replaces prior probabilities with a variable that is interpreted as a conceptual framework, and which we call a "paradigm." A paradigm is roughly the way an agent "carves up the world"; it is the kinds of hypotheses acknowledged by the agent.

[The idea that induction might depend on one's conceptual framework is not new. For example, Harsanyi (1983, p. 363) is sympathetic to a dependency on conceptual frameworks for simplicity-favoring in induction. Salmon (1990) argues for a Kuhnian paradigmatic role for prior probabilities. Earman (1992, p. 187) devotes a chapter to Kuhnian issues including paradigms. Di Maio (1994, especially pp. 148–149) can be interpreted as arguing for a sort of conceptual framework outlook on inductive logic. DeVito (1997) suggests this with respect to the choice of models in curve-fitting. Also, Gärdenfors (1990) develops a conceptual framework approach to address Goodman's riddle, and he attributes a conceptual framework approach to Quine (1960), Carnap (1989) and Stalnaker (1979).]

Having a paradigm, or conceptual framework, cannot, all by itself, tell us how we ought to proceed in our inductions. Oughts do not come from non-oughts. As discussed in the previous section, we are looking for principles of ought telling us how, given a paradigm, we should assign *a priori* degrees of belief in the hypotheses. I will put forward three symmetry-related principles that enable this.

Before moving to the theory, let us ask where the hypothesis set comes from. This is a difficult question, one to which I have no good answer. The difficulty is two-fold. First, what hypotheses should one include in the hypothesis set? And second, once this set is chosen, how is that set parameterized? I make some minimal overtures toward answering this in Changizi and Barber (1998), but it is primarily an unsolved, and possibly an unsolvable problem. I will simply assume here that the hypothesis set—a set of mutually exclusive hypotheses—and some parameterization of it is a given.

3.1.1 A brief first-pass at Paradigm Theory

Before presenting Paradigm Theory in detail, I think it is instructive to give a short introduction to it here, with many of the intricacies missing, but nevertheless capturing the key ideas. Paradigm Theory proposes to replace the variable prior probabilities of the Bayesian framework with variable "paradigms," which are interpreted as comprising the inductive agent's way of looking at the set of hypotheses, or the agent's conceptual framework. For example, you and I might share the same hypothesis set, but I might acknowledge that there are simple and complex hypotheses, and you might, instead, acknowledge that some are universal generalizations and some are not. More generally, a paradigm consists of the kinds of hypotheses one acknowledges. One of the most important aspects of paradigms is that they do not make a claim about the world; they are non-inductive. If, in complete ignorance about the world, I choose some particular paradigm, I cannot be charged with having made an unjustifiable assumption about the world. Paradigms are just a way of carving up the space of hypotheses, so they make no assumption. Prior probabilities, on the other hand, are straightforwardly claims about the world; namely, claims about the *a priori* degree of confidence in the hypotheses. The justification of induction in Paradigm Theory rests not on a variable choice of prior probabilities as it does in the Bayesian framework, but, instead, on a variable choice of a non-inductive paradigm. Paradigm Theory puts forth three principles which prescribe how prior probabilities ought to be assigned given that one possesses a paradigm. Different inductive methods differ only in the setting of the paradigm, not on any *a priori* differences about claims about the world or about how one ought to go about induction.

To understand the principles of prior probability determination, we have to understand that any paradigm naturally partitions the hypothesis set into distinct sets. [A *partition* of a set B is a set of subsets of B, where the subsets do not overlap and their union is B.] The idea is this. From the point of view of the paradigm—i.e., given the properties of hypotheses acknowledged in the paradigm—there are some hypotheses which cannot be distinguished using the properties in the paradigm. Hypotheses indistinguishable from one another are said to be *symmetric*. Each partition consists of hypotheses that are symmetric to one another, and each partition is accordingly called a *symmetry type*. Hypotheses in distinct partitions *can* be distinguished from one another. Since hypotheses cannot be distinguished within a symmetry type, the symmetry types comprise the kinds of hypothesis someone with that paradigm can distinguish.

Note that the symmetry types may well be different than the properties in the paradigm; the properties in the paradigm imply a partition into symmetry types of distinguishable (from the paradigm's viewpoint) types of hypotheses.

With an understanding that paradigms induce a natural partition structure onto the hypothesis set, I can state Paradigm Theory's principles for how one should assign prior probabilities. One principle states that each distinguishable type of hypothesis should, *a priori*, receive the same degree of confidence; this is the *Principle of Type Uniformity*. The intuition is that if one is only able to distinguish between certain types of hypotheses—i.e., they are the kinds of hypotheses one is able to talk about in light of the paradigm—and if there is no apparatus within the paradigm with which some of these distinguished types can *a priori* be favored (and there is no such apparatus), then it would be the height of arbitrariness to give any one type greater prior probability than another. The second principle states that hypotheses that are symmetric to one another—i.e., the paradigm is unable to distinguish them—should receive the same probability; this is the *Principle of Symmetry*. The motivation for this is that it would be entirely arbitrary, or random, to assign different *a priori* degrees of confidence to symmetric hypotheses, given that the paradigm has no way to distinguish between them; the paradigm would be at a loss to explain why one gets a higher prior probability than the other. There is one other principle in the full Paradigm Theory, but it is less central than these first two, and we can skip it in this subsection.

The Principle of Type Uniformity distributes equal shares of prior probability to each symmetry type, and the Principle of Symmetry distributes equal shares of the symmetry type's probability to its members. In this way a prior probability distribution is determined from a paradigm and the principles. Paradigms leading to different symmetry types usually lead to different prior probability distributions. Justifiable inductive methods are, then, all the same, in the sense that they share the Bayesian principle of evidence, and share the same principles of prior probability determination. They differ only in having entered the world with different ways of conceptualizing it. I can now make claims like, "If you conceptualize the world in fashion Q, then you ought to have prior probabilities $P_Q(H)$ determined by the principles of Paradigm Theory. This, in turn, entails a specific inductive method you ought to follow, since you ought to follow Bayes' Theorem in the application of evidence to your probabilities."

The remainder of this subsection, and the next subsection, develop this material in detail, but if you wish to skip the details, and wish to skip example

applications of Paradigm Theory (to enumerative induction, simplicity favoring, curve-fitting and more), you may jump ahead to Section 3.3.

3.1.2 Paradigms, Symmetry and Arbitrariness

In the next subsection I will present the principles of prior probabilities determination, i.e., principles of ought which say what one's prior probabilities should be given that one has a certain paradigm. But first we need to introduce paradigms, and to motivate the kinds of symmetry notions on which the principles will rest.

Paradigms

Let us begin by recalling that we are assuming that we somehow are given a hypothesis set, which is a set filled with all the hypotheses we are allowed to consider. The hypotheses could concern the grammar of a language, or the curve generating the data, and so on. The hypothesis set comprises an inductive agent's set of all possible ways the world could be (in the relevant regard).

Now, what is a paradigm? A paradigm is just a "way of thinking" about the set of hypotheses. Alternatively, a paradigm is the kinds of similarities and differences one appreciates among the hypotheses. Or, a paradigm stands for the kinds of hypotheses an inductive agent acknowledges. A paradigm is a kind of conceptual framework; a way of carving up the set of hypotheses into distinct types. It is meant to be one way of fleshing out what a Kuhnian paradigm might be (Kuhn, 1977). If the hypothesis set is the "universe," a paradigm is the properties of that "universe," a kind of ontology for hypothesis sets. When an inductive agent considers there to be certain kinds of hypotheses, I will say that the agent *acknowledges* those kinds, or *acknowledges* the associated properties. I do not mean to suggest that the agent would not be able to discriminate, or notice, other properties of hypotheses; the agent can presumably tell the difference between any pair of hypotheses. The properties in the paradigm, however, are the only properties that are "sanctioned" or endowed as "genuine" properties in the ontology of that universe of hypotheses.

For example, suppose the hypothesis set is the set of six outcomes of a roll of a six-sided die. One possible paradigm is the one that acknowledges being even and being odd; another paradigm is the one that acknowledges being small (three or less) and big (four or more). Or, suppose that the hypothesis set is the set of all points in the interior of a unit circle. One possible paradigm is the one that acknowledges being within distance 0.5 from the center. Another possible

paradigm would be the one acknowledging the different distances from the center of the circle; that is, points at the same radius would be of the same acknowledged kind. For a third example, suppose the hypothesis set is the set of all possible physical probabilities p of a possibly biased coin; i.e., the hypothesis set is $H = [0, 1]$, or all the real numbers from 0 to 1 included. One possible paradigm is the one that acknowledges the always-heads ($p = 0$) and always-tails ($p = 1$) hypotheses, and lumps the rest together. Another paradigm on this hypothesis set could be to acknowledge, in addition, the coin-is-fair hypothesis ($p = 1/2$).

For each of these examples, there is more than one way to carve up the hypothesis set. One person, or inductive community, might acknowledge properties that are not acknowledged by another person or community. Where do these properties in the paradigm come from? From Paradigm Theory's viewpoint it does not matter. The properties will usually be interpreted as if they are subjective. There are two kinds of subjective interpretations: in the first kind, the properties in the paradigm have been consciously chosen by the inductive agent, and in the second kind, the properties are in the paradigm because the inductive agent has evolved or been raised to acknowledge certain properties and not others.

Recall that our aim for a theory of logical probability was to have an interpretable, non-inductive variable to replace prior probabilities. In Paradigm Theory, the variable is the paradigm, and we have just seen that paradigms are interpreted as conceptual frameworks. But we also want our variable—namely, paradigms—to also be non-inductive, or not-about-the-world. (And, similarly, for our hoped-for theory of innateness, the innate content was to have some interpretable, non-inductive variable.)

Are paradigms about the world? A paradigm is just the set of properties acknowledged, and there is no way for a paradigm to favor any hypotheses over others, nor is there any way for a paradigm to favor any properties over others—each property is of equal significance. Paradigms cannot, say, favor simpler hypotheses, or disfavor hypotheses inconsistent with current ontological commitments; paradigms can *acknowledge* which hypotheses are simpler than others, and *acknowledge* which hypotheses are inconsistent with current ontological commitments. Paradigms make no mention of degrees of belief, they do not say how inductions ought to proceed, and they do not presume that the world is of any particular nature. Do not confuse a paradigm with information. Being unbiased, the properties in the paradigm give us no information about the success or truth of any hypothesis, and in this sense the paradigm is

not information. Therefore, paradigms are non-inductive.

To help drive home that paradigms are non-inductive, suppose that an agent discounts certain hypotheses on the basis of something not measured by the paradigm (e.g., "too complex") or favors some properties over others. Paradigm Theory is not then applicable, because the inductive agent now effectively already has prior probabilities. Paradigm Theory's aim is to attempt to defend inductive beliefs such as priors themselves. If an agent enters the inductive scenario with what are in effect prior probabilities, then Paradigm Theory is moot, as Paradigm Theory is for the determination of the priors one should have. Consider the following example for which Paradigm Theory is inapplicable. A tetrahedral die with sides numbered 1 through 4 is considered to have landed on the side that is face down. Suppose one acknowledges that one of the sides, side 4, is slightly smaller than the others, and acknowledges nothing else. The paradigm here might seem to be the one acknowledging that side 4 is a unique kind, and the others are lumped together. If this were so, Paradigm Theory would (as we will see) say that 4 should be preferred. But side 4 should definitely *not* be preferred! However, Paradigm Theory does not apply to cases where one begins with certain inductive beliefs (e.g., that smaller sides are less likely to land face down). Paradigm Theory is applicable in those kinds of circumstances where one has not yet figured out that smaller sides are less likely to land face down. [There may remain an issue of how we assign a precise prior probability distribution on the basis of an imprecise inductive belief such as "smaller sides are less likely to land face down," but this issue of formalization of imprecise inductive beliefs is a completely different issue than the one we have set for ourselves. It is less interesting, as far as a theory of logical probability goes, because it would only take us from imprecise inductive beliefs to more precise inductive beliefs; it would not touch upon the justification of the original imprecise inductive belief.]

I now have the concept of a paradigm stated, but I have not quite formally defined it. Here is the definition.

Definition 1 A *paradigm* is any set of subsets of the hypothesis set that is closed under complementation. The complements are presumed even when, in defining a paradigm, they are not explicitly mentioned. △

Recall that when you have a set of objects of any kind, a property is just a subset of the set: objects satisfying the property are in the set, and objects not satisfying the property are not in the set (i.e., are in the complement of the set). The definition of a paradigm just says that a paradigm is a set of subsets, or

properties; and it says that for any property P in the paradigm, the property of not being P is also in the set. And that is all the definition says.

Being Symmetric

We now know what a paradigm is: it is the non-inductive variable in our theory of logical probability that I call Paradigm Theory, and paradigms are interpreted as conceptual frameworks, or ways of conceptualizing the set of hypotheses. Our goal is to present compelling principles of rationality which prescribe how one ought to assign prior probabilities *given* one's paradigm; we would thereby have fixed principles of prior probability determination that all rational agents would follow, and all justifiable differences in inductive methods would be due to differences in the way the inductive agent carved up the world before having known anything about it.

Before we can understand Paradigm Theory's principles of prior probability determination, we must acquire a feel for the intuitive ideas relating to symmetry, and in this and the following subsubsection I try to relate these intuitions.

One of the basic ideas in the rational assignment of prior probabilities will be the motto that *names should not matter*. This motto is, generally, behind every symmetry argument and motivates two notions formally introduced in this subsubsection. The first is that of a *symmetry type*. Informally, two hypotheses are of the same symmetry type if the only thing that distinguishes them is their names or the names of the properties they possess; they are the same type of thing but for the names chosen. One compelling notion is that hypotheses that are members of smaller symmetry types may be chosen with less arbitrariness than hypotheses in larger symmetry types; it takes less arbitrariness to choose more unique hypotheses. The principles of Paradigm Theory in the Subsection 3.1.3, founded on different intuitions, respect this notion in that more unique hypotheses should receive greater prior probability than less unique hypotheses. The second concept motivated by the "names should not matter" motto, and presented in the next subsubsection, is that of a *defensibility hierarchy*, where picking hypotheses higher in the hierarchy is less arbitrary, or more defensible. The level of defensibility of a hypothesis is a measure of how "unique" it is. Subsection 3.1.3 describes how the principles of rationality of Paradigm Theory lead to a prior probability assignment which gives more defensible types of hypotheses greater prior probability. Onward to the intuition pumping.

Imagine having to pick a kitten for a pet from a box of five kittens, numbered 1 through 5. Imagine, furthermore, that you deem no kitten in the litter to be a better or worse choice for a pet. All these kittens from which to choose, and you may not wish to pick randomly. You would like to find a reason to choose one from among them, even if for no other reason but that one is distinguished in some way. As it turns out, you acknowledge some things about these kittens: the first four are black and the fifth is white. These properties of kittens comprise your paradigm. Now suppose you were to pick one of the black kittens, say kitten #1. There is no reason connected with their colors you can give for choosing #1 that does not equally apply to #2, #3 and #4. "I'll take the black kitten" does not pick out #1. Saying "I'll take kitten #1" picks out that first kitten, but these number-names for the kittens are arbitrary, and had the first four kittens been named #2, #3, #4 and #1 (respectively), "I'll take kitten #1" would have picked out what is now called the fourth kitten. #1 and #4 are the same (with respect to the paradigm) save their arbitrary names, and we will say that they are symmetric; in fact, any pair from the first four are symmetric.

Imagine that the five kittens, instead of being just black or white, are each a different color: red, orange, yellow, green and blue, respectively. You acknowledge these colors in your paradigm. Suppose again that you choose kitten #1. Unlike before, you can at least now say that #1 is "the red one." However, why is redness any more privileged than the other color properties acknowledged in this modified paradigm? 'red' is just a name for a property, and had these five properties been named 'orange', 'yellow', 'green', 'blue' and 'red' (respectively), "the red one" would have picked out what is now called the blue one. #1 and #5 will be said to be symmetric; in fact, each pair will be said to be symmetric.

For another example, given an infinite plane with a point above it, consider the set of all lines passing through the point. If the plane and point "interact" via some force, then along which line do they do so? This question was asked by a professor of physics to Timothy Barber and myself as undergraduates (we shared the same class), and the moral was supposed to be that by symmetry considerations the perpendicular line is the only answer, as for every other line there are lines "just as good." In our theoretical development we need some explicit paradigm (or class of paradigms) before we may make conclusions. Suppose that you acknowledge the properties of the form "having angle θ with respect to the plane," where a line parallel to the plane has angle 0. Any pick of, say, a parallel line will be arbitrary, as one can rotate the world

about the perpendicular line and the parallel line picked would become another one. Each parallel line is symmetric to every other. The same is true of each non-perpendicular line; for any such line there are others, infinitely many others, that are the same as far as the paradigm can tell. The perpendicular line is symmetric only with itself, however.

In the remainder of this subsubsection we make the notion of symmetry precise, but there is no real harm now skipping to the next subsubsection if mathematical details bother you. The following defines the notion of being symmetric.

Definition 2 Fix hypothesis set H and paradigm Q. h_1 and h_2 are *Q-symmetric* in H if and only if it is possible to rename the hypotheses respecting the underlying measure such that the paradigm is unchanged but the name for h_1 becomes the name for h_2. Formally, for $p : H \rightarrow H$, if $X \subseteq H$ then let $p(X) = \{p(x)|x \in X\}$, and if Q is a paradigm on H, let $p(Q) = \{p(X)|X \in Q\}$. h_1 and h_2 are *Q-symmetric* in H if and only if there is a measure-preserving bijection $p : H \rightarrow H$ such that $p(Q) = Q$ and $p(h_1) = h_2$. $\triangle$

In the definition of 'Q-symmetric' each measure-preserving bijection $p : H \rightarrow H$ is a renaming of the hypotheses. Q represents the way the hypothesis set H "looks," and the requirement that $p(Q) = Q$ means that the renaming cannot affect the way H looks. For example, if $H = \{h_1, h_2, h_3\}$ with names 'a', 'b', and 'c', respectively, and $Q = \{\{h_1, h_2\}, \{h_2, h_3\}\}$, the renaming $p_1 : (a, b, c) \rightarrow (c, b, a)$ preserves Q, but the renaming $p_2 : (a, b, c) \rightarrow (c, a, b)$ gives $p_2(Q) = \{\{h_3, h_1\}, \{h_1, h_2\}\} \neq Q$. Suppose we say, "Pick a." We are referring to h_1. But if the hypotheses are renamed via p_1 we see H in exactly the same way yet we are referring now to h_3 instead of h_1; and thus h_1 and h_3 are Q-symmetric. Two hypotheses are Q-symmetric if a renaming that swaps their names can occur that does not affect the way H looks. Only arbitrary names distinguish Q-symmetric hypotheses; and so we say that Q-symmetric hypotheses cannot be distinguished non-arbitrarily. Another way of stating this is that there is no name-independent way of referring to either h_1 or h_3 because they are the same symmetry type. h_1 and h_3 are of the same type in the sense that each has a property shared by just one other hypothesis, and that other hypothesis is the same in each case.

But cannot one distinguish h_1 from h_3 by the fact that they have different properties? The first property of Q is, say, 'being red,' the second 'being short.' h_1 is red and not short, h_3 is short and not red. However, so the intuition goes, just as it is not possible to non-arbitrarily refer to h_1 because of the "names

should not matter" motto, it is not possible to non-arbitrarily refer to the red hypotheses since $p_1(Q) = Q$ and $p_1(\{h_1, h_2\}) = \{h_3, h_2\}$ (i.e., $p_1(red) = short$). Our attempt to refer to the red hypotheses by the utterance "the red ones" would actually refer to the short hypotheses if 'red' was the name for short things. The same observation holds for, say, $Q' = \{\{h_\alpha\}, \{h_\beta\}, \{h_\gamma\}\}$. The fact that each has a distinct property does not help us to refer to any given one non-arbitrarily since each pair is Q'-symmetric.

Consider h_2 from above for a moment. It is special in that it has the unique property of being the only hypothesis having both properties. I say that a hypothesis is Q-*invariant* in H if and only if it is Q-symmetric only with itself. h_2 is invariant (the white kitten was invariant as well, as was the perpendicular line). Intuitively, invariant hypotheses can be non-arbitrarily referred to.

Three other notions related to 'symmetric' we use later are the following: First, $I(Q, H)$ is the set of Q-invariant hypotheses in H, and $\neg I(Q, H)$ is its complement in H. Above, $I(Q, H) = \{h_2\}$, and $\neg I(Q, H) = \{h_1, h_3\}$. Second, a paradigm Q is called *totally symmetric* on H if and only if the hypotheses in H are pairwise Q-symmetric. Q' above is totally symmetric (on $\{h_\alpha, h_\beta, h_\gamma\}$). Third, t is a *Q-symmetry type* in H if and only if t is an equivalence class with respect to the relation 'Q-symmetric'. $\{h_2\}$ and $\{h_1, h_3\}$ are the Q-symmetry types. In each of the terms we have defined, we omit Q or H when either is clear from context.

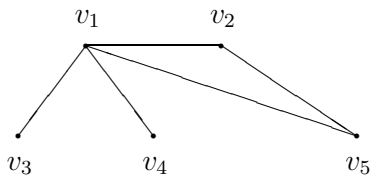
The Q-symmetry types are the most finely grained objects one can speak of or distinguish via the paradigm Q. One can distinguish between *no* hypotheses when the paradigm is totally Q-symmetric. When we say that a property is "acknowledged" we mean that the property is in the paradigm. Acknowledging a property does not mean that it is distinguishable, however, as we saw above with Q'. When we say that a property is "distinguishable" we mean that it is a symmetry type (but not necessarily a set appearing in the paradigm). $\{h_1, h_2\}$ is acknowledged in Q above but is not distinguishable. $\{h_2\}$ is distinguishable but not acknowledged in the paradigm.

Invariant hypotheses, then, can be non-arbitrarily referred to—non-invariant hypotheses cannot. From the point of view of the paradigm, invariant hypotheses can be "picked for a reason," but non-invariant hypotheses cannot. In this sense to pick an invariant hypothesis is to make a non-random choice and to pick a non-invariant hypothesis is to make a random choice; however I will try to avoid using this terminology for there are already many rigorous notions of randomness and this is not one of them. Any "reason" or procedure that picks a non-invariant hypothesis picks, for all the same reasons, any other hypothesis

in its symmetry type; where "reasons" cannot depend on names. We say that invariant hypotheses are more *defensible,* or less *arbitrary,* than non-invariant ones. Picking a hypothesis that is not invariant means that had it been named differently you would have chosen something else; this is bad because surely a defensible choice should not depend on the names. Invariant hypotheses would therefore seem, *a priori*, favorable to non-invariant hypotheses. More generally, the intuition is that hypotheses that are members of larger symmetry types are less preferred, as picking one would involve greater arbitrariness. These intuitions are realized by the rationality principles comprising Paradigm Theory (as we will see later).

Consider the following example. $H_a = \{h_0, h_1, h_2, h_3\}$, $Q_a = \{\{h_0\}, \{h_1\}, \{h_2\}, \{h_2, h_3\}\}$. The reader may check that h_0 is symmetrical to h_1, and that h_2 and h_3 are each invariant. Suppose one chooses h_0. Now suppose that the hypotheses h_0, h_1, h_2, h_3 are renamed h_1, h_0, h_2, h_3, respectively, under the action of p. Since $p(Q_a) = Q_a$, the choice of hypotheses is exactly the same. However, this time when one picks h_0, one has really picked h_1. h_3 is invariant because, intuitively, it is the only element that is not in a one-element set. h_2 is invariant because, intuitively, it is the only element occurring in a two-element set with an element that does not come in a one-element set.

One way to visualize paradigms of a certain natural class is as an undirected graph. Hypothesis set H and paradigm Q are *associated with* undirected graph G with vertices V and edges $E \subset V^2$ if and only if there is a bijection $p : V \rightarrow H$ such that $Q = \{\{p(v)\} | v \in V\} \cup \{\{p(v_1), p(v_2)\} \mid (v_1, v_2) \in E\}\}$. This just says that a graph can represent certain paradigms, namely those paradigms that (i) acknowledge each element in H and (ii) the other sets in Q are each composed of only two hypotheses. Consider the following graph.



The associated hypothesis set is $H_b = \{v_1, \ldots, v_5\}$ and the associated paradigm is $Q_b = \{\{v_1\}, \ldots, \{v_5\}\} \cup \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_1, v_4\}, \{v_1, v_5\}, \{v_2, v_5\}\}$. Notice that $\{v_1\}$, $\{v_2, v_5\}$, and $\{v_3, v_4\}$ are the Q_b-symmetry types; so only v_1 is Q_b-invariant—informally, it is the vertex that is adjacent to every

other vertex. When visualized as graphs, one is able to *see* the symmetry.

Defensibility Hierarchy and Sufficient Reason

In the previous subsubsection I introduced the notion of a symmetry type, and we saw that a paradigm naturally induces a partition on the hypothesis set, where each partitions consists of hypotheses that are symmetric to one another. The symmetry types are the kinds of hypotheses that the inductive agent can distinguish, given his paradigm. Hypotheses that are members of smaller symmetry types can intuitively be chosen with less arbitrariness, as there are fewer hypotheses just like it as far as the paradigm is concerned. An invariant hypothesis—a hypothesis that is all alone in its symmetry type—can be chosen with the least arbitrariness since there are no other hypotheses symmetrical to it. Invariant hypotheses can, intuitively, be picked for a reason.

Although an invariant hypothesis may be able to be picked for a reason and is thus more defensible than non-invariant hypotheses, if there are one hundred other invariant hypotheses that can be picked for one hundred other reasons, how defensible can it be to choose that hypothesis? Why *that* reason and not any one of the others? Among the invariant hypotheses one may wonder if there are gradations of invariance. The way this may naturally be addressed is to restrict the hypothesis set to the invariant hypotheses, consider the induced paradigm on this set (we discuss what this means in a moment), and again ask what is invariant and what is not. Intuitively, concerning those hypotheses that can be picked for a reason, which of these *reasons* is justifiable? That is to say, which of these hypotheses can *now* be picked for a reason?

For the remainder of this subsubsection we say how to make this precise, but if you wish to skip the details, it will serve the purpose to simply know that there is a certain well-motivated, well-defined sense in which a paradigm induces a hierarchy of more and more defensible hypotheses, where being more defensible means that it can, intuitively, be picked with less arbitrariness.

A paradigm Q is just the set of acknowledged properties of the hypotheses in H. If one cares only about some subset H' of H, then the *induced paradigm* is just the one that acknowledges the same properties in H'. Formally, if $H' \subseteq H$, let $Q \sqcap H'$ denote $\{A \cap H' | A \in Q\}$, and call it the *induced paradigm* on H'. $Q \sqcap H'$ is Q after throwing out all of the hypotheses in $H - H'$. For example, let $H_d = \{h_0, h_1, h_2, h_3, h_4\}$ and $Q_d = \{\{h_0, h_2\}, \{h_1, h_2\}, \{h_3\}, \{h_2, h_3, h_4\}\}$. h_0 and h_1 are the non-invariant hypotheses; h_2, h_3 and h_4 are the invariant hypotheses. Now let H'_d be the set of invariant hypotheses, i.e.,

$H'_d = I(Q_d, H_d) = \{h_2, h_3, h_4\}$. The induced paradigm is $Q'_d = Q_d \sqcap H'_d = \{\{h_2\}, \{h_3\}, \{h_2, h_3, h_4\}\}$.

Now we may ask what is invariant at this new level. h_2 and h_3 are together in a symmetry type, and h_4 is invariant. h_4 is the least arbitrary hypothesis among H'_d; and since H'_d consisted of the least arbitrary hypotheses from H_d, h_4 is the least arbitrary hypothesis of all. This hierarchy motivates the following definition.

Definition 3 Fix hypothesis set H and paradigm Q. $H^0 = H$, and for any natural number n, $Q^n = Q \sqcap H^n$. For any natural number n, $H^{n+1} = I(Q^n, H^n)$, which just means that H^{n+1} consists of the invariant hypotheses from H^n. This hierarchy is the *defensibility hierarchy*, or the *invariance hierarchy*. $\triangle$

For instance, for H_d and Q_d above we had:

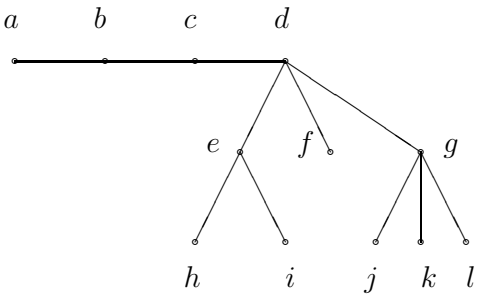
- $H^0_d = \{h_0, h_1, h_2, h_3, h_4\}$, $Q^0_d = \{\{h_0, h_2\}, \{h_1, h_2\}, \{h_3\}, \{h_2, h_3, h_4\}\}$.
- $H^1_d = \{h_2, h_3, h_4\}$, $Q^1_d = \{\{h_2\}, \{h_3\}, \{h_2, h_3, h_4\}\}$.
- $H^2_d = \{h_4\}$, $Q^2_d = \{\{h_4\}\}$.
- $H^3_d = \{h_4\}$, $Q^3_d = \{\{h_4\}\}$.
- etc.

For any hypothesis set H and paradigm Q there is an ordinal number $\alpha(Q, H)$ such that $H^\alpha = H^{\alpha+1}$; this is the *height* of the defensibility hierarchy of Q on H.[2] We say that a hypothesis h is *at level* m in the defensibility hierarchy if the highest level it gets to $\leq \alpha$ is the m^{th}. For H_d/Q_d, h_2 is at level 1, or the second level of the defensibility hierarchy; h_4 is at level 2, or the third level. We let Δ_m denote the set of hypotheses at level m. Hypotheses at higher levels in the hierarchy are said to be *more defensible*. This defines 'defensibility' respecting our intuition that, other things being equal, the more defensible a hypothesis the less arbitrary it is. h_4 is the lone maximally defensible hypothesis, and the intuition is that it is the most non-arbitrary choice and should, *a priori*, be favored over every other hypothesis.

[2]When H is infinite it is possible that the least ordinal number α such that $H^\alpha = H^{\alpha+1}$ is transfinite. To acquire hypothesis sets H^β when β is a limit ordinal we must take the intersection of H^γ for all $\gamma < \beta$. $Q^\beta = Q \sqcap H^\beta$ (as usual).

For H_d/Q_d above, notice that h_2 and h_3 are similar in that, although they are not symmetric with each other at level 0, they *are* symmetric at level 1. We will say that they are Q_d-equivalent. Generally, two hypotheses are *Q-equivalent* in H if and only if at some level H^n they become symmetric (i.e., there is a natural number n such that they are $Q \sqcap H^n$-symmetric). Two invariant hypotheses may therefore be Q-equivalent but not Q-symmetric. d is a *Q-equivalence type* in H if and only if d is an equivalence class of Q-equivalent hypotheses. $\{h_0, h_1\}$, $\{h_2, h_3\}$ and $\{h_4\}$ are the Q_d-equivalence types, whereas $\{h_0, h_1\}$, $\{h_2\}$, $\{h_3\}$ and $\{h_4\}$ are the symmetry types. The equivalence types are therefore coarser grained than the symmetry types. Two members of an equivalence type are equally defensible. For Q-equivalence types d_1 and d_2, we say that d_1 is *less Q-defensible than* d_2 if and only if for all $h \in d_1$ and $h' \in d_2$, h is less Q-defensible than h'. Our central intuition was that hypotheses that are more unique are to be preferred, *a priori*. Similarly we are led to the intuition that more defensible types of hypotheses are to be preferred, *a priori*. Paradigm Theory's rationality principles, presented in the next section, result in higher (actually, not lower) prior probability for more defensible equivalence types.

As an example, consider the paradigm represented by the following graph, where $H_f = \{a, \ldots, l\}$.



The symmetry types are $\{h, i\}$, $\{j, k, l\}$ and every other vertex is in a singleton symmetry type. The defensibility types are $\{h, i\}$, $\{j, k, l\}$, $\{e, f, g\}$, $\{a, d\}$ and $\{b, c\}$. The defensibility levels are $\Delta^0 = \{h, i, j, k, l\}$, $\Delta^1 = \{e, f, g\}$, and $\Delta^2 = \{a, b, c, d\}$.

We noted earlier that invariant hypotheses can be picked "for a reason," and this is reminiscent of Leibniz's Principle of Sufficient Reason, although

not with his metaphysical import,[3] which says, in Leibniz's words, "we can find no true or existent fact, no true assertion, without there being a sufficient reason why it is thus and not otherwise…" (Ariew and Garber, 1989, p. 217.) Rewording our earlier intuition, we can say that invariant hypotheses can be picked "for sufficient reason." The problem with this statement is, as we have seen, that there may be multiple invariant hypotheses, and what sufficient reason can there be to pick from among them? This subsubsection's defensibility hierarchy answers this question. It is perhaps best said that lone maximally defensible hypotheses may be picked "for sufficient reason." More important is that the defensibility hierarchy is a natural formalization and generalization of Leibniz's Principle of Sufficient Reason (interpreted non-metaphysically only), giving a more finely grained breakdown of "how sufficient" a reason is for picking a hypothesis: hypotheses in smaller symmetry types possess more sufficient reason, and hypotheses higher in the hierarchy possess (other things equal) more sufficient reason. Paradigm Theory, further, quantifies the degree of sufficiency of reason with real numbers in [0,1], as we will soon see.

3.1.3 Paradigm Theory's principles

In this subsection I present the guts of Paradigm Theory: its principles of ought. Let us first, though, sum up the previous subsection: I showed how acknowledging *any* set of subsets of a hypothesis set—i.e., a paradigm—naturally determines a complex hierarchical structure. We saw that the "names should not matter" motto leads to a partition of the hypothesis set into types of hypotheses: the symmetry types. Among those hypotheses that are the lone members of their symmetry type—i.e., the invariant (or "unique") hypotheses—there may be some hypotheses that are "more" invariant, and among *these* there may some that are "even more" invariant, etc. This led to the defensibility, or invariance, hierarchy. Hypotheses that "become symmetric" at some level of the hierarchy are equivalent, and are said to be members of the same equivalence type.

We also noted in Subsection 3.1.2 the following related intuitions for which we would like principled ways to quantitatively realize: *a priori*, (i) hypotheses in smaller symmetry types are more favorable; or, more unique hypotheses are to be preferred as it takes less arbitrariness to choose them, (ii) (equivalence)

[3]Leibniz believed that Sufficient Reason arguments actually determine the way the world must be. However, he did seem, at least implicitly, to allow the principle to be employed in a purely epistemic fashion, for in a 1716 letter to Newton's friend and translator Samuel Clarke, Leibniz writes, "has not everybody made use of the principle upon a thousand occasions?" (Ariew and Garber, 1989, p. 346).

types of hypotheses that are more defensible are more favorable, and (iii) the lone most defensible hypothesis—if there is one—is most favorable (this follows from (ii)). Each is a variant of the central intuition that less arbitrary hypotheses are, *a priori*, more preferred.

These intuitions follow from the three rationality principles concerning prior probabilities I am about to present. The principles are conceptually distinct from these intuitions, having intuitive motivations of their own. The fact that two unrelated sets of intuitions converge in the way we see below is a sort of argument in favor of Paradigm Theory, much like the way different intuitions on computability leading to the same class of computable functions is an argument for Church's Thesis. The motivations for stating each principle is natural and intuitive, and the resulting prior probability distributions are natural and intuitive since they fit with intuitions (i), (ii) and (iii).

The first subsubsection presents the three principles of rationality, the next discusses the use of "secondary paradigms" to acquire more detailed prior probability distributions, and the final subsubsection sets forth the sort of explanations Paradigm Theory gives.

The Principles

Paradigm Theory consists of three principles of rationality that, from a given paradigm (and a hypothesis set with a finite measure), determine a prior probability distribution. Paradigm Theory as developed in this section is only capable of handling cases where there are finitely many symmetry types.[4] We

[4]If one begins with H and Q such that there are infinitely many symmetry types, one needs to restrict oneself to a proper subset H' of H such that there are only finitely many symmetry types with respect to the induced paradigm. There are some compelling rationality constraints on such a restriction that very often suffice: (i) any two members of the same equivalence type in H either both appear in H' or neither, (ii) if an equivalence type from H appears in H', then (a) all more defensible equivalence types appear in H', and (b) all equally defensible equivalence types in H that are the same size or smaller appear in H'. These constraints on hypothesis set reduction connect up with the observation that we do not seriously entertain all logically possible hypotheses. This is thought by F. Suppe (1989, p. 398) "to constitute one of the deepest challenges we know of to the view that science fundamentally does reason and proceed in accordance with inductive logic." These rationality constraints help guide one to focus on the *a priori* more plausible hypotheses, ignoring the rest, and is a first step in addressing this challenge. These constraints give us the ability to begin to break the bonds of a logic of discovery of a prior assessment sort, and claim some ground also as a logic of discovery of a hypothesis generation sort: hypotheses are generated in the first place by "shaving off" most of the other logically possible hypotheses.

will assume from here on that paradigms induce just finitely many symmetry types.[5]

Assume hypothesis set H and paradigm Q are fixed. $P(A)$ denotes the probability of the set A. $P(\{h\})$ is often written as $P(h)$.

Principle of Type Uniformity

Recall that the symmetry types are precisely the types of hypotheses that can be referred to with respect to the paradigm. Nothing more finely grained than symmetry types can be spoken of. *Prima facie*, a paradigm gives us no reason to favor any symmetry type (or "atom") over any other. To favor one over another would be to engage in arbitrariness. These observations motivate the first principle of Paradigm Theory of Induction.

Principle of Type Uniformity: *Every (symmetry) type of hypothesis is equally probable.*

There are other principles in the probability and induction literature that are akin to the Principle of Type Uniformity. For example, if the types are taken to be the complexions (where two strings are of the same complexion if they have the same number of each type of symbol occurring in it), then Johnson's Combination Postulate (Johnson, 1924, p. 183) says to set the probability of the complexions equal to one another. Carnap's m^* amounts to the same thing.

The (claimed) rationality of the Principle of Type Uniformity emanates from the seeming rationality of choosing a non-arbitrary prior; to choose a non-uniform prior over the symmetry types would mean to give some symmetry types higher probability for no good reason. Is favoring some symmetry types over others necessarily arbitrary? Through the eyes of a paradigm the symmetry types are distinguishable, and might not there be aspects of symmetry types that make some, *a priori*, favorable? If any are favorable, it is not because any is distinguished among the symmetry types; each is equally distinguished. Perhaps some could be favorable by virtue of having greater size? Size is, in fact, relevant in determining which sets are the symmetry types. Actually, though, it is size *difference*, not size, that is relevant in symmetry type determination. Paradigms are not capable of recognizing the size of symmetry types; symmetry types *are* the primitive entities, or atoms, in the paradigm's

[5]This restriction ensures that the height of the defensibility hierarchy is finite (although having infinitely many symmetry types does not entail a transfinite height).

ontology. From the paradigm's point of view, symmetry types cannot be favored on the basis of their being larger. Given that one possesses a paradigm and nothing else (like particular inductive beliefs), it is plausible that anything but a uniform distribution on the symmetry types would be arbitrary.

Now, perhaps one could argue that the weakness of paradigms—e.g., their inability to acknowledge larger symmetry types—counts against Paradigm Theory. Paradigm Theory aims to be a "blank slate" theory of induction, taking us from innocuous ways of carving the world to particular degrees of belief. Paradigms are innocuous in part because of their weakness. Strengthening paradigms to allow the favoring of symmetry types over others would have the downside of decreasing the explanatory power; the more that is packed into paradigms, the less surprising it is to find that, given them, they justify particular inductive methods. That is my motivation for such a weak notion of paradigm, and given only such a weak paradigm, the Principle of Type Uniformity is rational since to not obey it is to engage in a sort of arbitrariness.

Principle of Symmetry

The second principle of rationality is a general way of asserting that the renaming of objects should not matter (so long as the paradigm Q is unaltered). Recall the convention that the underlying measure on H is finite.

Principle of Symmetry: *Within a symmetry type, the probability distribution is uniform.*

For finite H this is: hypotheses of the same type are equally probable, or, hypotheses that can be distinguished only by their names or the names of their properties are equally probable. Unlike the Principle of Type Uniformity whose intuition is similar to that of the Classical Principle of Indifference (which says that if there is no known reason to prefer one alternative over another, they should receive equal probability), the Principle of Symmetry is truly a symmetry principle. Violating the Principle of Symmetry would result in a prior probability distribution that would not be invariant under renamings that do not alter the paradigm; names would suddenly matter. Violating the Principle of Type Uniformity, on the other hand, would *not* contradict the "names should not matter" motto (and is therefore less compelling).

If one adopts the Principle of Symmetry without the Principle of Type Uniformity, the result is a Generalized Exchangeability Theory. Each paradigm induces a partition of symmetry types, and the Principle of Symmetry, alone,

requires only that the probability within a symmetry type be uniform. When the hypothesis set is the set of strings of outcomes (0 or 1) of an experiment and the paradigm is such that the symmetry types are the complexions (see Q_L then the Principle of Symmetry just *is* Johnson's Permutability Postulate (Johnson, 1924, pp. 178–189), perhaps more famously known as de Finetti's Finite Exchangeability.

The Basic Theory

Carnap's m^*-based theory of logical probability (Carnap, 1950, p. 563)—which I will call *Carnap's logical theory*—uses versions of the Principles of Type Uniformity and Symmetry (and leads to the inductive method he calls c^*). His "structure-descriptions," which are analogous to complexions, are given equal probability, which amounts to the use of a sort of Principle of Type Uniformity on the structure-descriptions. Then the probabilities are uniformly distributed to his "state-descriptions," which are analogous to individual outcome strings of experiments, which amounts to a sort of Principle of Symmetry. But whereas Carnap (and Johnson) is confined to the case where the partition over the state-descriptions is given by the structure-descriptions (or for Johnson, the partition over the outcome strings is given by the complexions), Paradigm Theory allows the choice of partition to depend on the choice of paradigm and is therefore a natural, powerful generalization of Carnap's m^*-based Logical Theory. The paradigm determines the symmetry types, and the symmetry types play the role of the structure-descriptions. When the hypothesis set is totally symmetric, one gets something akin to Carnap's $m^\dagger$-based logical theory (which he calls $c^\dagger$).

It is convenient to give a name to the theory comprised by the first two principles alone.

Basic Theory: *Assign probabilities to the hypothesis set satisfying the Principles of Type Uniformity and Symmetry.*

Applying the Basic Theory to H_a and Q_a from Subsection 3.1.2, we get $P(h_0) = P(h_1) = 1/6$ and $P(h_2) = P(h_3) = 1/3$. Applying the Basic Theory to H_b and Q_b from the same subsection, we get $P(v_1) = 1/3$, and the remaining vertices each receive probability $1/6$. Applying it to H_d and Q_d, $P(h_0) = P(h_1) = 1/8$ and $P(h_2) = P(h_3) = P(h_4) = 1/4$.

Notice that since the underlying measure of the hypothesis set is finite, the

probability assignment for the Basic Theory is unique. For $h \in H$ let $c(h)$ be the cardinality of the symmetry type of h. Let w denote the number of symmetry types in H. The following theorem is obvious.

Theorem 1 *Fix finite H. The following is true about the Basic Theory. For all $h \in H$, $P(h) = \frac{1}{w \cdot c(h)}$.* $\triangle$

Theorem 1 may be restated more generally to include infinite hypothesis sets: for any measure μ and all $A \subseteq H$ with measure μ that are a subset of the same symmetry type, $P(A) = \frac{1}{w\mu}$.

We see that the probability of a hypothesis is inversely proportional to both the number of symmetry types and the number (or measure) of other hypotheses of the same symmetry type as itself. The fraction $1/w$ is present for every hypothesis, so $c(h)$ is the variable which can change the probabilities of hypotheses relative to one another. The more hypotheses in a type, the less probability we give to each of those hypotheses; this fits with our earlier intuition number (i) from the beginning of this section. The following corollary records that the Basic Theory fits with this intuition and the intuition that invariant hypotheses are more probable. The corollary is true as stated no matter the cardinality of the hypothesis set.

Theorem 2 *The following are true about the Basic Theory.*

1. *Hypotheses in smaller symmetry types acquire greater probability.*
2. *Each invariant hypothesis receives probability $1/w$, which is greater than (in fact, at least twice as great as) that for any non-invariant hypothesis.* $\triangle$

The Basic Theory is not Paradigm Theory, although when the defensibility hierarchy has no more than two levels the two theories are equivalent. The Basic Theory does not notice the hierarchy of more and more defensible hypotheses, and noticing the hierarchy will be key to providing a general explanation for why simpler hypotheses ought to be favored. When I say things like, "only the Basic Theory is needed to determine such and such probabilities," I mean that the probabilities are not changed upon the application of the third principle (to be stated below) of Paradigm Theory.

Principle of Defensibility

The third principle of rationality is, as far as I know, not similar to any previous principle in the induction and probability literature. It encapsulates an intuition

similar to that used when I discussed gradations of invariance in Subsection 3.1.2. I asked: Among the invariant elements, which are more defensible? Now I ask: Among the invariant elements, which are more probable? From the viewpoint of the entire hypothesis set H the invariant hypotheses seem equally and maximally defensible. But when focusing only on the invariant hypotheses we see further gradations of defensibility. Similarly, from the viewpoint of the entire hypothesis set H the invariant hypotheses look equally and maximally probable. But when focusing only on the invariant hypotheses we see further gradations of probability. The third principle of rationality says to refocus attention on the invariant hypotheses.

Principle of Defensibility: *Reapply the Principles of Type Uniformity, Symmetry, and Defensibility to the set of invariant hypotheses* $(H' = I(Q, H))$ *via the induced paradigm* $(Q \sqcap H')$.

Since the Principle of Defensibility is one of the three rationality principles mentioned in its own statement, it applies to itself as well. I have named the principle the Principle of Defensibility because it leads to the satisfaction of intuition (ii) from the beginning of this section, i.e., to more defensible types of hypotheses acquiring higher prior probability. However, neither the intuitive motivation for the principle nor the statement of the principle itself hints at this intuition. The principle only gets at the idea that there is structure among the invariant hypotheses and that it should not be ignored.

Paradigm Theory

With the three principles presented I can state Paradigm Theory.

Paradigm Theory: *Assign probabilities to the hypothesis set satisfying the Principles of Type Uniformity, Symmetry, and Defensibility.*

The proposal for the prior probability assignment is to use the principles in the following order: (i) Type Uniformity, (ii) Symmetry, and (iii) Defensibility (i.e., take the invariant hypotheses and go to (i)). These principles amount to a logical confirmation function, as in the terminology of Carnap, but ours is a function of a hypothesis h, evidence e, *and* paradigm Q; i.e., $c(h, e, Q)$.

Paradigm Theory is superior to the Basic Theory in the sense that it is able to distinguish higher degrees of defensibility. Paradigm Theory on H_a/Q_a and

H_b/Q_b from Section 3.1.2 behaves identically to the Basic Theory. Applying Paradigm Theory to H_d and Q_d is different, however, than the Basic Theory's assignment. First we get, as in the Basic Theory, $P(h_0) = P(h_1) = 1/8$, $P(h_2) = P(h_3) = P(h_4) = 1/4$. Applying the Principle of Defensibility, the probability assignments to h_0 and h_1 remain fixed, but the $3/4$ probability assigned to the set of invariant hypotheses is to be redistributed among them. With respect to $\{h_2, h_3, h_4\}$ and the induced paradigm $\{\{h_2, h_3\}, \{h_4\}\}$, the symmetry types are $\{h_2, h_3\}$ and $\{h_4\}$, so each symmetry type receives probability $(3/4)/2 = 3/8$. The probabilities of $h_0, \ldots h_4$ are, respectively, $2/16$, $2/16$, $3/16$, $3/16$, $6/16$. Recall that h_4 is the lone most defensible element but the Basic Theory gave it the same probability as h_2 and h_3; Paradigm Theory allows richer assignments than the Basic Theory.

It is easy to see that since the underlying measure of the hypothesis set is finite and there are assumed to be only finitely many symmetry types, Paradigm Theory assigns a unique probability distribution to the hypothesis set, and does so in such a way that each hypothesis receives positive prior probability density (i.e., priors are always "open-minded" within Paradigm Theory). Theorem 14 in the appendix at the end of this chapter examines some of its properties. Unlike the Basic Theory, Paradigm Theory respects the intuition (number (ii)) that more defensible (less arbitrary) implies higher probability by giving the more defensible equivalence types not less probability than the less defensible equivalence types. Also, unlike the Basic Theory, Paradigm Theory respects the intuition (number (iii)) that if a hypothesis is lone most defensible (the only least arbitrary one) then it receives higher probability than every other hypothesis. The following theorem states these facts; the proofs along with other properties are given in the appendix to this chapter.

Theorem 3 *The following are true about Paradigm Theory.*

1. *For all equivalence types d_1 and d_2, if d_1 is less defensible than d_2, then $P(d_1) \leq P(d_2)$.*

2. *For all hypotheses h; h is the lone most defensible if and only if for all $h' \neq h$, $P(h') < P(h)$.* $\triangle$

Theorem 3 is an argument for the superiority of Paradigm Theory over the Basic Theory.

Secondary Paradigms

Suppose we have found the prior probability distribution on H given a paradigm Q, and, say, half of the hypotheses end up with the same probability; call this subset H^*. Now what if we acknowledge other properties concerning H^*, properties which are, in some sense, *secondary* to the properties in the original paradigm? May H^*'s probabilities be validly redistributed according to this secondary paradigm? After all, cannot any hypothesis set and paradigm be brought to Paradigm Theory for application, including H^* and this secondary paradigm? The problem is that to do this would be to modify the original, or *primary* probability distribution, and this would violate the principles in the original application of Paradigm Theory.

Here is an example of the sort of thing I mean. Let $H = \{3, \ldots, 9\}$ and Q acknowledge the property of being prime. There are two symmetry types, $\{4, 6, 8, 9\}$ and $\{3, 5, 7\}$, each receiving probability $1/2$. Now suppose that there are secondary paradigms for each symmetry type, in each case acknowledging the property of being odd. The second symmetry type above remains unchanged since all are odd, but the first gets split into $\{4, 6, 8\}$ and $\{9\}$, each receiving probability $1/4$. Notice that this is different than what a primary paradigm that acknowledges both being prime and odd gives; in this case the probability of $\{3, 5, 7\}$, $\{4, 6, 8\}$ and $\{9\}$ are $1/3$, $1/3$, $1/3$ instead of, respectively, $1/2$, $1/4$, $1/4$, as before. The first method treats being prime as more important than being odd in the sense that primality is used to determine the large-scale probability structure, and parity is used to refine the probability structure. The second method treats being prime and being odd on a par. A more Kuhnian case may be where one allows the primary paradigm to acknowledge scope, and allows the secondary paradigm to acknowledge simplicity; this amounts to caring about scope first, simplicity second.

I generalize Paradigm Theory to allow such secondary paradigms in a moment, but I would first like to further motivate it. There is a sense in which Paradigm Theory, as defined thus far, is artificially weak. For simplicity consider only the Principles of Type Uniformity and Symmetry; i.e., the Basic Theory. These two principles are the crux of the probability assignment on the hypothesis set. Together they allow only two "degrees of detail" to probability assignments: one assignment to the symmetry types, and another to the particular hypotheses within the symmetry types. The Principle of Defensibility does allow further degrees of detail *for the invariant hypotheses*, and it accomplishes this without the need for secondary paradigms. But for non-invariant

hypotheses there are just two degrees of detail. Why two? This seems to be a somewhat artificial limit.

Allowing secondary paradigms enables Paradigm Theory to break this limit. Paradigm Theory is now generalized in the following way: *Secondary paradigms may modify the primary prior probability distribution by applying the three principles to any subset H^* such that the primary prior in H^* is uniform.* In other words, we are licensed to tinker with the primary prior using secondary paradigms, so long as we tinker only on subsets that were originally equiprobable. When H^* and a secondary paradigm Q^* are brought to Paradigm Theory for application, they can be treated as creating their own primary distribution within H^*. Secondary paradigms with respect to H^* and Q^* are *tertiary* paradigms with respect to the original hypothesis set H and paradigm Q. The point is that any degree of detail in the sense mentioned above is now sanctioned, so long as there are n^{th}-ary paradigms for large enough n.

All this increase in power may make one skeptical that one can create any prior one wants by an ad hoc tuning of the secondary (tertiary, and so on) paradigms. An explanation by Paradigm Theory is only as natural and explanatory as is the paradigm (primary, secondary, and so on) used (see Section 3.1.3). Ad hoc secondary paradigms create ad hoc explanations. The only use of paradigms in this chapter beyond primary ones are secondary ones. I use them later where they are quite explanatory and give Paradigm Theory the ability to generalize a certain logical theory of Hintikka's ($\alpha = 0$). I also note in Subsection 3.2.3 their ability to give a non-uniform prior over the simplest hypotheses. If in any particular application of Paradigm Theory there is no mention of secondary paradigms, then they are presumed not to exist.

The Paradigm Theory Tactic

In the following section Paradigm Theory is used to explain why certain inductive methods we tend to believe are justified are, indeed, justified. The general tactic is two-fold. First, a mathematical statement concerning the power of Paradigm Theory is given (often presented as a theorem). Second, an informal explanatory argument is given. Paradigm Theory's ability to justify induction is often through the latter.

Most commonly, the mathematical statement consists of showing that paradigm Q entails inductive method x. This alone only shows that inductive method x is or is not within the scope of Paradigm Theory; and this is a purely mathematical question. Such a demonstration is not enough to count as an ex-

planation of the justification of inductive method x. Although paradigm Q may determine inductive method x, Q may be artificial or ad hoc and thereby not be very explanatory; "who would carve the world *that* way?" If Q is very unnatural and no natural paradigm entails inductive method x, then this may provide an explanation for why inductive method x is disfavored: one would have to possess a very strange conceptual framework in order to acquire it, and given that we do not possess such strange conceptual frameworks, inductive method x is not justified. Typically, the paradigm Q determining inductive method x is natural, and the conclusion is that inductive method x is justified because we possess Q as a conceptual framework. I do not actually argue that we *do* possess any particular paradigm as a conceptual framework. Rather, "inductive method x is justified because we possess paradigm Q" is meant to indicate the form of a possible explanation in Paradigm Theory. A fuller explanation would provide some evidence that we in fact possess Q as conceptual framework.

A second type of mathematical statement is one stating that every paradigm entails an inductive method in the class Z. The explanatory value of such a statement is straightforward: every conceptual framework leads to such inductive methods, and therefore one cannot be a skeptic about inductive methods in Z; any inductive method not in Z is simply not rational. A sort of mathematical statement that sometimes arises in future sections is slightly weaker: every paradigm Q *of such and such type* entails an inductive method in the class Z. The explanatory value of this is less straightforward, for it depends on the status of the "such and such type." For example, open-mindedness is of this form for the Personalistic (of Subjective) Bayesian on the hypothesis set $H = [0, 1]$: every prior that is open-minded (everywhere positive density) converges in the limit to the observed frequency. If the type of paradigm is extremely broad and natural, and every paradigm not of that type is not natural, then one can conclude that inductive skepticism about inductive methods in Z is not possible, unless one is willing to possess an unnatural paradigm; inductive skepticism about inductive methods in Z is not possible because every non-artificial conceptual framework leads to Z. Similar observations hold for arguments of the form, "no paradigm Q of such and such type entails an inductive method in the class Y."

These claims of the "naturalness" of paradigms emanate from our (often) shared intuitions concerning what properties are natural. The naturalness of a paradigm is *not* judged on the basis of the naturalness of the inductive method to which it leads; this would ruin the claims of explanatoriness.

3.2 Applications

3.2.1 Simple preliminary applications

By way of example we apply the Basic and Paradigm Theories to some preliminary applications, first presented in Changizi and Barber (1998).

Collapsing to the Principle of Indifference

Paradigm Theory (and the Basic Theory) gives the uniform distribution when the paradigm is empty. This is important because, in other words, Paradigm Theory collapses to a uniform prior when no properties are acknowledged, and this is a sort of defense of the Classical Principle of Indifference: be ignorant *and* acknowledge nothing... get a uniform prior. More generally, a uniform distribution occurs whenever the paradigm is totally symmetric. Since being totally symmetric means that there are no distinctions that can be made among the hypotheses, we can say that *Paradigm Theory collapses to a uniform prior when the paradigm does not have any reason to distinguish between any of the hypotheses.* Only the Principle of Symmetry—and not the Principle of Type Uniformity—needs to be used to found the Principle of Indifference as a subcase of Paradigm Theory.

Archimedes' Scale

Given a symmetrical scale and (allegedly) without guidance by prior experiment Archimedes (*De aequilibro*, Book I, Postulate 1) predicts the result of hanging equal weights on its two sides. The hypothesis set in this case is plausibly the set of possible angles of tilt of the scale. Let us take the hypothesis set to include a finite (but possibly large) number, N, of possible tilting angles, including the horizontal, uniformly distributed over the interval $[-90°, 90°]$. Archimedes predicts that the scale will remained balanced, i.e., he settles on $\theta = 0°$ as the hypothesis. He makes this choice explicitly on the basis of the obvious symmetry; that for any $\theta \neq 0°$ there is the hypothesis $-\theta$ which is "just as good" as θ, but $\theta = 0°$ has no symmetric companion.

To bring this into Paradigm Theory, one natural paradigm is the one that acknowledges the amount of tilt but does not acknowledge which way the tilt is; i.e., $Q = \{\{-\theta, \theta\} | 0° \leq \theta \leq 90°\}$. $\theta = 0°$ is the only hypothesis in a single-element set in Q, and it is therefore invariant. Furthermore, every other hypothesis can be permuted with at least its negation, and so $\theta = 0°$ is the only

invariant hypothesis. With the paradigm as stated, any pair $-\theta, \theta$ (with $\theta > 0°$) can permute with any other pair, and so there are two symmetry types: $\{0°\}$ and everything else. Thus, $0°$ receives prior probability $1/2$, and every other hypothesis receives the small prior probability $1/(2 \cdot (N-1))$. Even if N is naturally chosen to be 3—the three tilting angles are $-90°$, $0°$ and $90°$—the prior probabilities are $1/4$, $1/2$ and $1/4$, respectively.

Now let the paradigm be the one acknowledging the property of being within $\theta°$ from horizontal, for every $\theta \in [0°, 90°]$. For each $\theta \in H$, $\{-\theta, \theta\}$ is a symmetry type, and this includes the case when $\theta = 0°$, in which case the symmetry type is just $\{0°\}$. Each symmetry type receives equal prior probability by the Principle of Type Uniformity, and by the Principle of Symmetry each $\theta \neq 0°$ gets half the probability of its symmetry type. $0°$ gets all the probability from its symmetry type, however, as it is invariant. Therefore it is, *a priori*, twice as probable as any other tilting angle. If N is chosen to be 3, the prior probabilities for $-90°$, $0°$ and $90°$ are as before: $1/4$, $1/2$ and $1/4$, respectively.

Explanations for such simple cases of symmetry arguments can sometimes seem to be assumptionless, but certain *a priori* assumptions are essential. Paradigm Theory explains Archimedes' prediction by asserting that he possessed one of the paradigms above as a conceptual framework (or some similar sort of paradigm). He predicts that the scale will remained balanced because, roughly, he acknowledges the angle of tilt but not its direction. Most natural paradigms will entail priors favoring $\theta = 0°$, and I suspect no natural paradigm favors any other.

Leibniz's Triangle

To a second historical example, I noted earlier the connection of Paradigm Theory to Leibniz's Principle of Sufficient Reason (interpreted non-metaphysically), and I stated that Paradigm Theory is a sort of generalization of the principle, giving precise real-valued degrees to which a hypothesis has sufficient reason to be chosen. Let us now apply Paradigm Theory to an example of Leibniz. In a 1680s essay, he discusses the nature of an unknown triangle.

> And so, if we were to imagine the case in which it is agreed that a triangle of given circumference should exist, without there being anything in the givens from which one could determine what kind of triangle, freely, or course, but without a doubt. There is nothing in the givens which prevents another kind of triangle from existing, and so, an equilateral triangle is not necessary. However,

> all that it takes for no other triangle to be chosen is the fact that in no triangle except for the equilateral triangle is there any reason for preferring it to others. (Ariew and Garber, 1989, p. 101.)

Here the hypothesis set is plausibly $\{\langle\theta_1,\theta_2,\theta_3\rangle|\ \theta_1+\theta_2+\theta_3 = 180^\circ\}$, where each 3-tuple defines a triangle, θ_i being the angle of vertex i of the triangle. Now consider the paradigm that acknowledges the three angles of a triangle, but does not acknowledge which vertex of the triangle gets which angle; i.e., $Q = \{\{\langle\theta_1,\theta_2,\theta_3\rangle, \langle\theta_3,\theta_1,\theta_2\rangle, \langle\theta_2,\theta_3,\theta_1\rangle, \langle\theta_3,\theta_2,\theta_1\rangle, \langle\theta_1,\theta_3,\theta_2\rangle, \langle\theta_2,\theta_1,\theta_3\rangle\} \mid \theta_1+\theta_2+\theta_3 = 180^\circ\}$. This natural paradigm, regardless of the hypothesis set's underlying measure, results in $\langle 60^\circ, 60^\circ, 60^\circ\rangle$ being the only invariant hypothesis. In fact, every other of the finitely many symmetry types is of the size continuum, and thus every hypothesis but the 60° one just mentioned receives infinitesimal prior probability. An explanation for why Leibniz believed the equilateral triangle must be chosen is because he possessed the conceptual framework that acknowledged the angles but not where they are.

Straight Line

Consider a hypothesis set H consisting of all real-valued functions consistent with a finite set of data falling on a straight line (and let the underlying measure be cardinality). It is uncontroversial that the straight line hypothesis is the most justified hypothesis. Informally, I claim that any natural paradigm favors—if it favors any function at all—the straight line function over all others, and that this explains why in such scenarios we all feel that it is rational to choose the straight line. For example, nothing but the straight line can be invariant if one acknowledges any combination of the following properties: 'is continuous', 'is differentiable', 'has curvature κ' (for any real number κ), 'has n zeros' (for any natural number n), 'has average slope of m' (for any real number m), 'changes sign of slope k times' (for any natural number k). One can extend this list very far. For specificity, if the curvature properties are acknowledged for each κ, then the straight line is the only function fitting the data with zero curvature, and for every other value of curvature there are multiple functions fitting the data that have that curvature; only the straight line is invariant and Paradigm Theory gives it highest probability. The same observation holds for the 'changes sign of slope k times' property. What is important is not any particular choice of natural properties, but the informal claim that any natural choice entails that the straight line is favored if any function is. The reader is challenged to think of a natural paradigm that results in some other function in

H receiving higher prior probability than the straight line.

Reference

For a consistent set of sentences, each interpretation of the language making all the sentences true can be thought of as a hypothesis; that is, each model of the set of sentences is a hypothesis. The question is: Which model is, *a priori*, the most probable? Consider the theorems of arithmetic as our consistent set of sentences. There is one model of arithmetic, called the "standard model," that is considered by most of us to be the most preferred one. That is, if a person having no prior experience with arithmetic were to be presented with a book containing all true sentences of arithmetic (an infinitely long book), and this person were to attempt to determine the author's interpretation of the sentences, we tend to believe that the standard model should receive the greatest prior probability as the hypothesis. Is this preference justified?

Suppose that one's paradigm acknowledges models "fitting inside" other models, where a model M_1 *fits inside* M_2 if the universe of M_1 is a subset (modulo any isomorphism) of that of M_2 and, when restricted to the universe of M_1, both models agree on the truth of all sentences.[6] Intuitively, you can find a copy of M_1 inside M_2 yet both satisfactorily explain the truth of each sentence in the set. As such, M_2 is unnecessarily complex.[7] Does this paradigm justify the standard model? The standard model of arithmetic has the mathematical property that it fits inside any model of arithmetic; it is therefore invariant for this paradigm. We do not know of a proof that there is no other invariant (for this paradigm) model of arithmetic, but it is strongly conjectured that there is no other (M. C. Laskowski, private communication). If this is so, then the standard model is the most probable one (given this paradigm).

Paradigm Theory can be used to put forth a conceptual framework-based probabilistic theory of reference in the philosophy of language: *to members of a conceptual framework represented by paradigm Q, the reference of a symbol in a language is determined by its interpretation in the most probable model, where the prior probabilities emanate from Q and are possibly conditioned via Bayes' Theorem if evidence (say, new sentences) comes to light.* (See Putnam (1980, 1981, p. 33) for some discussion on underdetermination of interpretation and its effect on theories of reference, and Lewis (1984) for some commentary and criticism of it.

[6] In logic it is said in this case that M_1 *embeds elementarily* into M_2.

[7] This is a sort of "complexification;" see Subsection 3.2.3.

3.2.2 Enumerative Induction

I consider *enumerative induction* on two types of hypothesis set: (i) the set of strings of the outcomes (0 or 1) of N experiments or observations, and I denote this set H_N; (ii) the set of possible physical probabilities p in $[0, 1]$ of some experiment, with the uniform underlying measure. Three types of enumerative induction are examined: no- , frequency- , and law-inductions. *No-induction* is the sort of inductive method that is completely rationalistic, ignoring the evidence altogether and insisting on making the same prediction no matter what. *Frequency-induction* is the sort of inductive method that converges in the limit to the observed frequency of experimental outcomes (i.e., the ratio of the number of 0s to the total number of experiments). *Law-induction* is the sort of inductive method that is capable of giving high posterior probability to laws. 'all 0s' and 'all 1s' are the laws when $H = H_N$, and '$p = 0$' and '$p = 1$' are the laws when $H = [0, 1]$.

For reference throughout this section, Table 3.1 shows the prior probability assignments for the paradigms used in this section on the hypothesis set H_4.

No-Induction

The sort of no-induction we consider proceeds by predicting with probability .5 that the next experimental outcome will be 0, regardless of the previous outcomes.

$H = H_N$

First we consider no-induction on the hypothesis set H_N, the set of outcome strings for N binary experiments. Table 3.1 shows the sixteen possible outcome strings for four binary experiments. The first column of prior probabilities is the uniform assignment, and despite its elegance and simplicity, it does not allow learning from experience. For example, suppose one has seen three 0s so far and must guess what the next experimental outcome will be. The reader may easily verify that $P(0|000) = P(1|000) = 1/2$; having seen three 0s does not affect one's prediction that the next will be 0. The same is true even if one has seen one million 0s in a row and no 1s. This assignment is the one Wittgenstein proposes (1961, 5.15–5.154), and it is essentially Carnap's $m^\dagger$ (Carnap, 1950) (or $\lambda = \infty$).

Recall that a totally symmetric paradigm is one in which every pair of hypotheses is symmetric. Any totally symmetric paradigm entails the uniform

Table 3.1: The prior probability assignments for various paradigms over the hypothesis set H_4 (the set of possible outcome strings for four experiments) are shown. Q_{law_L} is shorthand for Q_{law} with Q_L as secondary paradigm. The table does not indicate that in the Q_{law} cases the 'all 0s' and 'all 1s' acquire probability $1/4$ no matter *the value of N (in this case, $N = 4$); for the other paradigms this is not the case.*

string	Q_s	Q_L	Q_{rep}	Q_{law}	Q_{law_L}
0000	1/16	1/5	1/8	1/4	1/4
0001	1/16	1/20	1/24	1/28	1/24
0010	1/16	1/20	1/24	1/28	1/24
0100	1/16	1/20	1/24	1/28	1/24
1000	1/16	1/20	1/24	1/28	1/24
0011	1/16	1/30	1/24	1/28	1/36
0101	1/16	1/30	1/8	1/28	1/36
0110	1/16	1/30	1/24	1/28	1/36
1001	1/16	1/30	1/24	1/28	1/36
1010	1/16	1/30	1/8	1/28	1/36
1100	1/16	1/30	1/24	1/28	1/36
0111	1/16	1/20	1/24	1/28	1/24
1011	1/16	1/20	1/24	1/28	1/24
1101	1/16	1/20	1/24	1/28	1/24
1110	1/16	1/20	1/24	1/28	1/24
1111	1/16	1/5	1/8	1/4	1/4

assignment on H_N. Therefore, any totally symmetric paradigm results in no-induction on H_N. This is true because there is just one symmetry type for a totally symmetric paradigm, and so the Principle of Symmetry gives each string the same prior probability. I have let Q_s denote a generic totally symmetric paradigm in Table 3.1.

The uniform assignment on H_N is usually considered to be inadequate on the grounds that the resulting inductive method is not able to learn from experience. There is a problem with this sort of criticism: it attributes the inadequacy of a particular prior probability assignment to the inadequacy of the inductive method to which it leads. If prior probabilities are chosen simply in order to give the inductive method one wants, then much of the point of prior probabilities is missed. Why not just skip the priors altogether and declare the desired inductive method straightaway? In order to be explanatory, prior probabilities must be chosen for reasons independent of the resulting inductive method. We want to explain the lack of allure of the uniform prior on H_N *without* referring to the resulting inductive method.

One very important totally symmetric paradigm is the empty one, i.e., the paradigm that acknowledges nothing. If one considers H_N to be the hypothesis set, and one possesses the paradigm that acknowledges no properties of the hypotheses at all, then one ends up believing that each outcome string is equally likely. I believe that for H_N the paradigm that acknowledges nothing is far from natural, and this helps to explain why no-induction is treated with disrepute. To acknowledge nothing is to not distinguish between the 'all 0s' string and any "random" string; for example, 0000000000 and 1101000110. To acknowledge nothing is also to not acknowledge the relative frequency. More generally, any totally symmetric paradigm, no matter how complicated the properties in the paradigm, does not differentiate between any of the outcome strings and is similarly unnatural. For example, the paradigm that acknowledges every outcome string is totally symmetric, the paradigm that acknowledges every pair of outcome strings is totally symmetric, and the paradigm that acknowledges every property is also totally symmetric. No-induction is unjustified because we do not possess a conceptual framework that makes no distinctions on H_N. On the other hand, if one really does possess a conceptual framework that makes no distinctions among the outcome strings, then no-induction *is* justified.

There are some ad hoc paradigms that do make distinctions but still entail a uniform distribution over H_N. For example, let paradigm Q acknowledge $\{1\}$, $\{1,2\}$, $\{1,2,3\}$, ..., $\{1,\ldots,16\}$, where these numbers denote the corresponding strings in Table 3.1. Each string is then invariant, and therefore can

be distinguished from every other, yet the probability assignment is uniform by the Principle of Type Uniformity. For another example, let the paradigm consist of $\{1,\ldots,8\}$ and $\{1,\ldots,16\}$. There are two symmetry types, $\{1,\ldots,8\}$ and $\{9,\ldots,16\}$, each subset can be distinguished from the other, but the resulting prior probability assignment is still uniform. These sorts of paradigms are artificial—we have not been able to fathom any natural paradigm of this sort. The explanation for why no-induction is unjustified is, then, because we neither possess conceptual frameworks that make no distinctions nor possess conceptual frameworks of the unnatural sort that make distinctions but still give a uniform distribution.

$H = [0,1]$

Now we take up no-induction on the hypothesis set $H = [0,1]$, the set of physical probabilities p of a repeatable experiment. In no-induction it is as if one believes with probability 1 that the physical probability of the experiment (say, a coin flip) is $.5$, and therefore one is incapable of changing this opinion no matter the evidence. In fact this is *exactly* what the uniform probability assignment over H_N is equivalent to. That is, the prior on $[0,1]$ leading to no-induction gives $p = .5$ probability 1, and the probability density over the continuum of other hypotheses is zero. What was an elegant, uniform distribution on H_N has as its corresponding prior on $[0,1]$ an extremely inelegant Dirac delta prior. With $[0,1]$ as the hypothesis set instead of H_N, there is the sense in which no-induction is *even more* unjustified, since the prior is so clearly arbitrary. The reason for this emanates from the fact that $[0,1]$ is a "less general" hypothesis set than H_N, for, informally, $[0,1]$ lumps all of the outcome strings in a single complexion into a single hypothesis (recall, two strings are in the same complexion if they have the same number of 0s and 1s); H_N is capable of noticing the order of experiments, $[0,1]$ is not. This property of $[0,1]$, that it presumes exchangeability, severely constrains the sort of inductive methods that are possible and makes frequency-induction "easier" to achieve in the sense that any open-minded prior converges asymptotically to the observed frequency; no-induction is correspondingly "harder" to achieve in $[0,1]$.

In fact, within Paradigm Theory no-induction on [0,1] is impossible to achieve for the simple reason that paradigms always result in open-minded priors. The reason we believe no-induction is unjustified on [0,1] is because no paradigm leads to no-induction.

Frequency-Induction

If an experiment is repeated many times, and thus far 70% of the time the outcome has been 0, then in very many inductive scenarios most of us would infer that there is a roughly 70% chance that the next experiment will result in 0. This is frequency-induction, and is one of the most basic ways in which we learn from experience, but is this method justifiable? Laplace argued that such an inference is justified on the basis of his Rule of Succession. It states that out of $n + 1$ experiments, if 0 occurs r times out of the first n, then the probability that 0 will occur in the next experiment is $\frac{r+1}{n+2}$. As $n \to \infty$, this very quickly approaches $\frac{r}{n}$; and when $r = n$, it very quickly approaches 1. Derivations of this rule depend (of course) on the prior probability distribution; see Zabell (1989) for a variety of historical proofs of the rule. In this section we demonstrate how Paradigm Theory naturally leads to the Rule of Succession when $H = H_N$ and $H = [0, 1]$.

$H = H_N$

The second column of probabilities in Table 3.1, headed "Q_L," shows the probability assignment on H_4 needed to lead to Laplace's Rule of Succession.[8] Notice, in contrast to Q_s, that for this column $P(0|000) = (1/5)/(1/5 + 1/20) = 4/5$, and so $P(1|000) = 1/5$; it learns from experience. Laplace's derivation was via a uniform prior on the hypothesis set $H = [0, 1]$ (with uniform underlying prior), but on H_N something else is required. Johnson's Combination Postulate and Permutability Postulate (Johnson, 1924, pp. 178–189) together give the needed assignment. The Combination Postulate—which states that it is *a priori* no more likely that 0 occurs i times than j times in n experiments—assigns equal probability to each complexion, and the Permutability Postulate—which states that the order of the experiments does not matter—distributes the probability uniformly within each complexion. Carnap's logical theory with m^* (Carnap, 1950, p. 563) does the same by assigning equal probability to each structure-description (analogous to the complexions), and distributing the probability uniformly to the state-descriptions (analogous to the individual outcome strings) within each structure-description (see the earlier discussion of the "Basic Theory").

In order for Paradigm Theory to give this prior probability assignment it suffices to find a paradigm whose induced symmetry types are the complex-

[8] A discussion on the difference between Q_s and Q_L can be found in Carnap (1989).

ions. If a paradigm satisfies this, the Principle of Type Uniformity assigns each complexion the same prior probability, and the Principle of Symmetry uniformly distributes the probability among the outcome strings within each complexion. In other words, if one's conceptual framework distinguishes the complexions, then one engages in frequency-induction via the Rule of Succession. Explanatorily, the Rule of Succession is justified because we possess paradigms that distinguish the complexions.

For distinguishing the complexions it is not sufficient to simply acknowledge the complexions; if the paradigm consists of *just* the complexions, then there are three symmetry types in H_4 as in Table 3.1: $\{0000, 1111\}$, $\{0001, 0010, 0100, 1000, 1110, 1101, 1011, 0111\}$, and the "middle" complexion. There *are* very natural paradigms that do induce symmetry types equal to the complexions. One such paradigm is employed in the following theorem whose proof may be found in the appendix to this chapter.

Theorem 4 *Let Q_L ('L' for 'Laplace') be the paradigm containing each complexion* and *the set of all sequences with more 0s than 1s. The probability assignment of Q_L on H_N via Paradigm Theory is identical to that of Johnson, and so Q_L results in Laplace's Rule of Succession.* $\triangle$

Note that Q_L is quite natural. It is the paradigm that acknowledges the complexions, and in addition acknowledges the difference between having more 0s than 1s and not more 0s than 1s. An explanation for the intuitive appeal of the Rule of Succession is that we often acknowledge exactly those properties in Q_L, and from this the Rule of Succession follows.

Since there are only finitely many inductive methods that may result given H_N via Paradigm Theory, the theory is not capable of handling a continuum of frequency-inductive methods as in Johnson and Carnap's λ-continuum, which says if r of n outcomes have been 1 in a binary experiment, the probability of the next outcome being a 1 is $\frac{r+\lambda/2}{n+\lambda}$. I have not attempted to determine the class of all λ such that there exists a paradigm that entails the λ-rule, but it seems that the only two natural sorts of paradigms that lead to an inductive method in the λ-continuum with $H = H_N$ are totally symmetric paradigms and those that have the complexions as the symmetry types. The first corresponds to $\lambda = \infty$, and the second corresponds to $\lambda = 2$. Reichenbach's Straight Rule (where, after seeing r of n outcomes of 1 in a binary experiment, the probability that the next will be 1 is r/n), or $\lambda = 0$, does not, therefore, seem to be justifiable within Paradigm Theory.

Laplace's Rule of Succession needs the assumption on H_N that, *a priori*, it is no more likely that 1 is the outcome i times than j times in n experiments. Call a *repetition* the event where two consecutive experiments are either both 1 or both 0; two strings are in the same *repetition set* if they have the same number of repetitions. Why, for example, should we not modify Johnson's Combination Postulate (or Principle of Indifference on the complexions) to say that, *a priori*, it is no more likely that a repetition occurs i times than j times in n experiments? The prior probability assignment resulting from this does not lead to Laplace's Rule of Succession, but instead to the "Repetition" Rule of Succession. 'REP' denotes the assignment of equal probabilities to each repetition set, with the probability uniformly distributed among the strings in each repetition set; this is shown for H_4 in Table 3.1 under the heading Q_{rep}. If one has seen r repetitions of 1 thus far with n experiments, the probability the outcome of the next experiment will be the same as the last outcome, via REP, is $\frac{r+1}{n+1}$. The proof is derivable from Laplace's Rule of Succession once one notices that the number of ways of getting r repetitions in a length n binary sequence is $2C_r^{n-1}$; the proof is omitted. This result can be naturally accommodated within Paradigm Theory.

Theorem 5 *Let Q_{rep} be the paradigm that acknowledges the number of repetitions in a sequence as well as acknowledging the sequences with less than half the total possible number of repetitions. The probability assignment of Q_{rep} is identical to that of REP, and so Q_{rep} results in the Repetition Rule of Succession.* $\triangle$

Whereas all of the previously mentioned paradigms on H_N entail prior probability assignments that are de Finetti exchangeable, Q_{rep} does not. It *is* Markov exchangeable, however: where strings with both the same initial outcome and the same number of repetitions have identical prior probability. A conceptual framework that acknowledges both the number of repetitions and which (0 or 1) has the greater number of repetitions results in the Repetition Rule of Succession. When our inductive behavior is like the Repetition Rule, it is because we possess Q_{rep} (or something like it) as our conceptual framework.

Q_L and Q_{rep} generally give very different predictions. However, they nearly agree on the intuitively clear case where one has seen all of the experiments give the same result. For example, Laplace had calculated the probability that the sun will rise tomorrow with his Rule of Succession; "It is a bet of 1,826,214 to one that it will rise again tomorrow" (Laplace, 1820). The Repetition Rule of Succession says that the odds are 1,826,213 to one that tomorrow

will be the same as the past with respect to the sun rising or not, and since we know it came up today, those are the odds of the sun rising tomorrow.

$H = [0, 1]$

Now we consider frequency-induction on the hypothesis set $H = [0, 1]$ with the natural uniform underlying measure. We noted earlier that $[0, 1]$ "more easily" leads to frequency-induction than H_N; disregarding the order of experiments puts one well on the path toward frequency-induction. We should suspect, then, that it should be easier to acquire frequency-inductions with $[0, 1]$ as the hypothesis set than H_N via Paradigm Theory. In fact, frequency-induction is guaranteed on $[0, 1]$ since paradigms lead to open-minded priors which, in turn, lead to frequency-induction. One cannot be a skeptic about frequency-induction in $[0, 1]$. Frequency-induction on [0,1] is justified because every conceptual framework leads to it.

For Laplace's Rule of Succession, Laplace assigned the uniform prior probability distribution over the underlying measure, from which the Rule follows. Here is the associated result for Paradigm Theory.

Theorem 6 *Any totally symmetric paradigm entails the uniform assignment on* $[0, 1]$. *Therefore, any totally symmetric paradigm results in Laplace's Rule of Succession.* $\triangle$

If one acknowledges nothing on $[0, 1]$, or more generally one makes no distinctions, Paradigm Theory collapses to a sort of Principle of Indifference (see Subsection 3.2.1) and one engages in frequency-induction via Laplace's Rule of Succession. Laplace's Rule of Succession is justified because when presented with hypothesis set $[0, 1]$ we possess a conceptual framework that does not distinguish between any hypotheses.

Law-Induction

Frequency-induction allows instance confirmation, the ability to place a probability on the outcome of the very next experiment. C. D. Broad (1918) challenged whether frequency-induction, Laplace's Rule of Succession in particular, is ever an adequate description of learning. The premises that lead to the Rule of Succession also entail that if there will be N experiments total and one has conducted n so far, all of which are found to be 1 (i.e., $r = n$), then the probability that *all* outcomes will be 1 is $(n + 1)/(N + 1)$. If N is large

compared to n, $(n+1)/(N+1)$ is small; and this is the origin of Broad's complaint. In real situations N, if not infinite, is very large. Yet we regularly acquire high degree of belief in the general law that all outcomes will be 1 with only a handful of experiments (small n). For example, we all conclude that all crows are black on the basis of only a small (say 100) sample of black crows. If, by 'crow,' we mean those alive now, then N is the total number of living crows, which is in the millions. In this case, after seeing 100 black crows, or even thousands, the probability via the Rule of Succession premises of the law 'all crows are black' is miniscule. The probability that all crows are black becomes high only as n approaches N—only after we have examined nearly every crow! Therefore, the premises assumed for the Rule of Succession cannot be adequate to describe some of our inductive methods.

Carnap (1950, pp. 571–572) makes some attempts to argue that instance confirmation is sufficient for science, but it is certain that we (even scientists) do in fact acquire high probability in universal generalizations, and the question is whether (and why) we are justified in doing so.

Jeffreys (1955) takes Broad's charge very seriously. "The answer is obvious. The uniform assessment of initial probability says that before we have any observations there are odds of $N-1$ to 2 against any general law holding. This expresses a violent prejudice against a general law in a large class" (ibid., p. 278). He suggests that the prior probability that a general law holds be a constant > 0, independent of N. This allows learning of general laws. For example, fix a probability of $.1$ that a general law holds, $.05$ for the 'all 0s' law, $.05$ for the 'all 1s' law, the probability uniformly distributed over the rest. After seeing just five black crows the probability of the 'all 0s' law is $.64$, and after seeing ten black crows the probability becomes $.98$; and this is largely independent of the total number of crows N.

The problem with this sort of explanation, which is the sort a Personalistic Bayesian is capable of giving, is that there seems to be no principled reason for why the general laws should receive the probability assignments they do; why not $.06$ each instead of $.05$, or why not $.4$ each? Paradigm Theory determines exact inductive methods capable of giving high posterior probability to laws, and it does so with very natural paradigms.

$H = H_N$

Beginning with H_N as the hypothesis set, suppose one acknowledges only two properties: being a general law and not being a general law. With this

comprising the paradigm Q_{law} the induced symmetry types are the same as the acknowledged properties. Paradigm Theory gives probability $.5$ to a general law holding—$.25$ to 'all 0s', $.25$ to 'all ones'—and $.5$ uniformly distributed to the rest; see the "Q_{law}" column in Table 3.1. Largely independent of the total number of crows, after seeing just one black crow the probability that all crows are black is $.5$. After seeing 5 and 10 black crows the probability becomes $.94$ and $.998$, respectively—near certainty that all crows are black after just a handful of observations. I record this in the following theorem whose proof may be found in the appendix to this chapter.

Theorem 7 *If there will be N experiments and $1 \leq n < N$ have been conducted so far, all which resulted in 1, then the probability that all N experiments will result in 1, with respect to the paradigm Q_{law} on the hypothesis set H_N, is approximately*

$$\frac{2^{n-1}}{1+2^{n-1}}.\triangle$$

One is open to the confirmation of universal generalizations if one acknowledges being a law and acknowledges no other properties. Of course, the theorem holds for any paradigm that induces the same symmetry types as Q_{law}. For example, suppose that a paradigm Q_{const} acknowledges the *constituents*, from Hintikka (1966), where a constituent is one possible way the world can be in the following sense: either all things are 0, some things are 0 and some are 1, or all things are 1. The induced symmetry types are the same as those induced by Q_{law}.

Similar results to Theorem 7 follow from any paradigm that (i) has {'all 0s','all 1s'} as a symmetry type (or each is alone a symmetry type), and (ii) there is some natural number k such that for all N the total number of symmetry types is k. Q_{law} and Q_{const} are special cases of this, with $k = 2$. Each paradigm satisfying (i) and (ii) entails an inductive method that is capable of giving high posterior probability to universal generalizations. This is because the two laws each receive the probability $1/(2k)$ (or $1/k$ if each is invariant) no matter how large is the number of "crows in the world" N.

There is a problem with paradigms satisfying (i) and (ii). Paradigms satisfying (i) and (ii) are not able to engage in frequency-induction when some but not all experiments have resulted in 1. This is because frequency-induction on H_N requires that one distinguish among the $N+1$ complexions, and this grows with N, and so (ii) does not hold. Specifically considering Q_{law} and Q_{const}, the most natural paradigms satisfying (i) and (ii), when some but not

all experiments have resulted in 1 the Q_{law} and Q_{const} assignment does not learn at all. This is because the probabilities are uniformly distributed over the outcome strings between the 'all 0s' and 'all 1s' strings, just like when the paradigm is Q_s from earlier.

To "fix" this problem it is necessary to employ a secondary paradigm. We concentrate only on fixing Q_{law} for the remainder of this subsection, but the same goes for Q_{const} as well. What we need is a secondary paradigm on the set of strings between 'all 0s' and 'all 1s' that distinguishes the complexions, i.e., has them as symmetry types. Let the secondary paradigm be the one acknowledging the complexions and the property of having more 0s than 1s, which is like the earlier Q_L, and let the hypothesis set be $H_N - \{$'all 0s','all 1s'$\}$ instead of H_N. The resulting inductive behavior is like Laplace's Rule of Succession for strings that are neither 'all zeros' nor 'all ones', and similar to that of Q_{law} described in Theorem 7 for the 'all 0s' and 'all 1s' strings. We denote this paradigm and secondary paradigm duo by Q_{law_L}, and one can see the resulting prior probability on H_4 in Table 3.1. The proof of part (a) in the following theorem emanates, through de Finetti's Representation Theorem, from part (a) of Theorem 9; (b) is proved as in Theorem 4.

Theorem 8 *Q_{law_L} assigns prior probabilities to H_N ($n < N$) such that if 1 occurs r times out of n total, then (a) if $r = n > 0$ the probability that all outcomes will be 1 is approximately $\frac{n+1}{n+3}$, and (b) if $0 < r < n$ the probability that the next outcome will be a 1 is $\frac{r+1}{n+2}$ (i.e., the inductive method is like that of Q_L).* $\triangle$

After seeing 5 and 10 black crows, the probability that all crows are black is approximately $.75$ and $.85$, respectively.

How natural is the primary/secondary paradigm pair Q_{law_L}? It acknowledges being a law (or in Q_{const}'s case, acknowledges the constituents), acknowledges the complexions, and acknowledges having more 0s than 1s. But it also believes that the laws (or constituents) are more important (or "more serious" parts of the ontology) than the latter two properties. "Primarily, the members of our paradigm acknowledge laws; we acknowledge whether or not all things are 0, and whether or not all things are 1. Only secondarily do we acknowledge the number of 0s and 1s and whether there is a greater number of 0s than 1s." Having such a conceptual framework would explain why one's inductive behavior allows both frequency-induction and law-induction. Note that if Q_L were to be primary and Q_{law} secondarily applied to each symmetry type induced by Q_L, then the result would be no different than Q_L alone. The same

is true if we take as primary paradigm the union of both these paradigms. Thus, if being a law is to be acknowledged independently of the other two properties at all, it must be via relegating the other two properties to secondary status.

The above results on universal generalization are related to one inductive method in Hintikka's two-dimensional continuum (Hintikka, 1966). Q_{law_L} (and Q_{const_L}) corresponds closely to Hintikka's logical theory with $\alpha = 0$ (ibid., p. 128), except that Hintikka (primarily) assigns probability $1/3$ to each constituent: $1/3$ to 'all 0s', $1/3$ to 'all 1s', and $1/3$ to the set of strings in between. In Q_{law} (and Q_{const}) 'all 0s' and 'all 1s' are members of the same symmetry type, and so the probabilities were, respectively, $1/4$, $1/4$, $1/2$. Then (secondarily) Hintikka divides the probability of a constituent evenly among the structure-descriptions, which are analogous to our complexions. Finally, the probability of a structure-description is evenly divided among the state-descriptions, which are analogous to our outcome strings. Q_{law_L}, then, acknowledges the same properties as does Hintikka's "$\alpha = 0$"-logical theory, and in the same order.

It *is* possible for Paradigm Theory to get *exactly* Hintikka's $\alpha = 0$ assignment, but the only paradigms I have found that can do this are artificial. For example, a paradigm that does the job is the one that acknowledges 'all 0s' and the pairs $\{$'all 1s', $\sigma\}$ such that σ is a non-law string. 'all 0s' and 'all 1s' are now separate symmetry types, and the non-law strings in between comprise the third. Each thus receives prior probability $1/3$ as in Hintikka's "$\alpha = 0$"-Logical Theory. This paradigm is indeed artificial, and I do not believe Paradigm Theory can give any natural justification for the $\alpha = 0$ inductive method.

With Q_{law_L} in hand we can appreciate more fully something Paradigm Theory can accomplish with secondary paradigms: a principled defense and natural generalization of Hintikka's "$\alpha = 0$"-logical theory. Well, not exactly, since as just mentioned the nearest Paradigm Theory can naturally get to $\alpha = 0$ is with Q_{law_L} (or Q_{const_L}). Ignoring this, Paradigm Theory gives us a principled reason for why one should engage in law-induction of the $\alpha = 0$ sort: because one holds Q_{law} (or Q_{const}) as the conceptual framework, and Q_L secondarily. Paradigm Theory also allows different notions of what it is to be a law, and allows different properties to replace that of being a law. The $\alpha = 0$ tactic can be applied now in any way one pleases.

$H = [0, 1]$

We have seen in Subsection 3.2.2 that $[0, 1]$ as the hypothesis set makes frequency-

induction easier to obtain than when the hypothesis set is H_N. Informally, one must expend energy when given H_N so as to treat the complexions as the primitive objects upon which probabilities are assigned, whereas this work is already done when given $[0, 1]$ instead. To do this job on H_N for law-induction we required secondary paradigms in order to have frequency-induction as well, but it should be no surprise that on $[0, 1]$ having both comes more easily.

As in the previous subsubsection we begin with the paradigm that acknowledges being a law and not. We call it by the same name, Q_{law}, although this is strictly a different paradigm than the old one since it is now over a different hypothesis set. There are two symmetry types, $\{0, 1\}$ and $(0, 1)$. Thus, $p = 0$ and $p = 1$ each receives probability $.25$, and the remaining $.5$ is spread uniformly over $(0, 1)$. This is a universal-generalization (UG) open-minded prior probability distribution, where not only is the prior probability density always positive, but the $p = 0$ and $p = 1$ hypotheses are given positive probability; this entails an inductive method capable of learning laws. It is also open-minded, and so is an example of frequency-induction as well; we do not need secondary paradigms here to get this. In fact, because the prior is uniform between the two endpoints the inductive behavior follows Laplace's Rule of Succession when the evidence consists of some 0s and some 1s. The following theorem records this; the proof of (a) is in the appendix to this chapter, and (b) is derived directly from Laplace's derivation of the Rule of Succession.

Theorem 9 *Q_{law} on $[0, 1]$ entails the prior probability distribution such that if 1 occurs r times out of n total, then (a) if $r = n > 0$ the probability that $p = 1$ is $\frac{n+1}{n+3}$, and (b) if $0 < r < n$ the probability that the next outcome will be a 1 is $\frac{r+1}{n+2}$.* $\triangle$

If one holds $[0, 1]$ as the hypothesis set and acknowledges being a law and nothing else, one is both able to give high probability to laws and converge to the relative frequency. Turned around, we should engage in law- and frequency-induction (of the sort of the previous theorem) because our conceptual framework acknowledges the property of being a law. One need make no primitive assumption concerning personal probabilities as in Personalistic Bayesianism, one need only the extremely simple and natural Q_{law}.

Similar results to Theorem 9 can be stated for any paradigm such that the two laws appear in symmetry types that are finite (the laws are distinguished, at least weakly). For any such paradigm the two laws are learnable because they acquire positive prior probability, and frequency-induction proceeds (asymptotically, at least) because the prior is open-minded. In an informal sense, "any"

natural paradigm acknowledging the laws results in both law- and frequency-induction.

3.2.3 Simplicity-Favoring

Occam's Razor says that one should not postulate unnecessary entities, and this is roughly the sort of simplicity to which I refer (although any notion of simplicity that has the same formal structure as that described below does as well). Paradigm Theory is able to provide a novel justification for simplicity: *when the paradigm acknowledges simplicity, it is "usually" the case that simpler hypotheses are less arbitrary and therefore receive higher prior probability.* This explanation for the preferability of simpler hypotheses does *not* assume that we must favor simpler hypotheses in the paradigm (something the paradigm does not have the power to do anyway). The paradigm need only *acknowledge* which hypotheses are simpler than which others.[9] In a sentence, Paradigm Theory gives us the following explanation for why simpler hypotheses are preferred: *simpler hypotheses are less arbitrary.*

For any hypothesis there are usually multiple ways in which it may be "complexified"—i.e., unnecessary entities added—to obtain new hypotheses. Each complexification itself may usually be complexified in multiple ways, and so may each of its complexifications, and so on. A *complexification tree* is induced by this complexification structure, starting from a given hypothesis as the root, its complexifications as the children, their complexifications as the grandchildren, etc.[10]

Recall from Subsection 3.1.2 that certain paradigms are representable as graphs. Consider the following two special cases of trees whose associated paradigms result in the root being the lone maximally defensible element; the proof is found in the appendix to this chapter. A tree is *full* if every leaf is at the same depth in the tree.

Theorem 10 *The paradigm associated with any full tree or finite-depth binary tree places the root as the lone maximally defensible element. But not every paradigm associated with a tree does so, and these two cases do not exhaust the trees that do so.* △

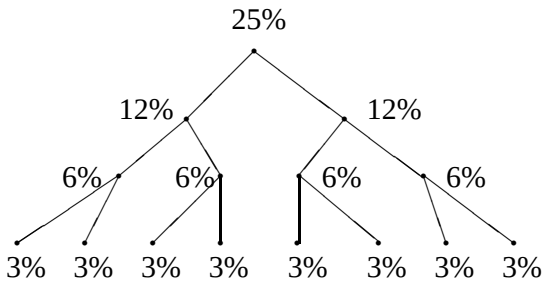
[9] In fact, it suffices to acknowledge the two-element subsets for which one element is simpler than the other; after all, paradigms as defined for the purposes of this chapter do not allow relations.

[10] I am ignoring the possibility that two hypotheses may "complexify" to the same hypothesis, in which case the structure is not a tree.

If a hypothesis set H consists of h, all of its complexifications and all of their complexifications and so on, and the paradigm on H is the complexification tree with root h—i.e., the paradigm acknowledges the pairs of hypotheses for which one is a complexification of the other—then the paradigm puts h alone at the top of the hierarchy if the tree is full or finite binary.[11] Informally, "most" natural notions of hypothesis and complexification imply complexification trees that are full. Such paradigms naturally accommodate Occam's Razor; acknowledging simplicity results in setting the lone most defensible element to what Occam's Razor chooses for many natural (at least finite binary and full) complexification trees. The hypothesis that posits the least unnecessary entities is, in these cases, the lone most defensible hypothesis, and thus acquires the greatest prior probability (via Theorem 3).

Full Complexification Trees

Let Q_{full} be the paradigm represented by the full tree below.



There are four symmetry types (one for each level), so each receives probability $1/4$. The approximate probability for each hypothesis is shown in the figure. Only the Basic Theory is needed here—i.e., the Principles of Type Uniformity and Symmetry—the Principle of Defensibility does not apply. If there are m such trees, the m roots each receive probability $\frac{1}{4m}$, the $2m$ children each receive $\frac{1}{8m}$, the $4m$ grandchildren each receive $\frac{1}{16m}$, and the $8m$ leaves

[11] We are assuming that the paradigm acknowledges only those pairs of hypotheses such that one is an "immediate" complexification of the other, i.e., there being no intermediate complexification in between. Without this assumption the complexification trees would not be trees at all, and the resulting graphs would be difficult to illustrate. However, the results in this section do not depend on this. If the paradigm acknowledges every pair such that one is simpler than the other, then all of the analogous observations are still true.

each receive $\frac{1}{32m}$. The following theorem generalizes this example. Recall that the depth of the root of a tree is zero.

Theorem 11 *Suppose the paradigm's associated graph consists of m full b-ary ($b \geq 2$) trees of depth n, and that hypothesis h is at depth i in one of them. Then $P(h) = \frac{1}{m(n+1)b^i}$.* $\triangle$

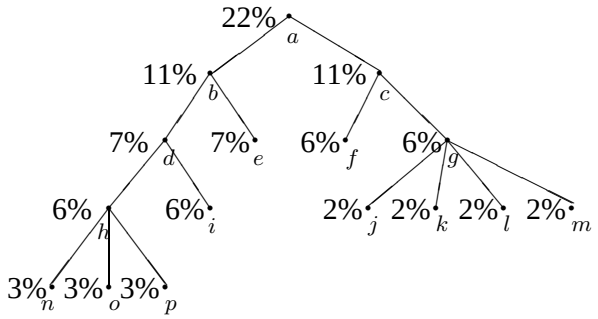
This tells us that the prior probability of a hypothesis drops exponentially the more one complexifies it, i.e., the greater i becomes. For example, consider base 10 numbers as in the hypothesis set $H = \{2; 2.0, \ldots, 2.9; 2.00, \ldots, 2.09; \ldots; 2.90, \ldots 2.99\}$, and suppose the paradigm is the one corresponding to the complexification tree. Here we have a 10-ary tree of depth two; 2 is the root, the two-significant-digit hypotheses are at depth one, and the three-significant-digit hypotheses are at depth two. $P(2) = 1/3$, the probability of a hypothesis at depth one is $1/30$, and the probability of a hypothesis at depth two is $1/300$.

When there are multiple trees, the roots may be interpreted as the "serious" hypotheses, and the complexifications the "ridiculous" ones. Theorem 11 tells us that when one acknowledges simplicity and the resulting paradigm is represented by multiple b-ary trees of identical depth, one favors the serious hypotheses over all others. This is a pleasing explanation for why prior probabilities tend to accrue to the simplest hypotheses, but it results in each of these hypotheses being equally probable. A conceptual framework may be more complicated, acknowledging properties capable of distinguishing between the different complexification trees. In particular, a secondary paradigm may be applied to the set of roots, with the understanding that the properties in the secondary paradigm are acknowledged secondarily to simplicity.

Asymmetrical Complexification Trees

We saw in Theorem 10 that any—even an asymmetrical—finite binary tree results in the root being the lone most defensible element. The Principle of Defensibility tends to apply non-trivially when trees are asymmetrical, unlike when trees are full where it makes no difference. The next example shows an asymmetrical tree where Paradigm Theory "outperforms" the Basic Theory. To demonstrate the last sentence of Theorem 10, that 'full' and 'finite binary' do not exhaust the trees that result in a lone most defensible root, we have chosen the tree to be non-binary. We leave it as an exercise to find the probabilities for a similar asymmetrical binary tree.

Let $H_{asymm} = \{a, \ldots, p\}$, and Q_{asymm} be as pictured.



With semicolons between the equivalence types, the invariance levels are $\Delta^0 = \{j, k, l, m; n, o, p\}$, $\Delta^1 = \{f, g; h, i\}$, $\Delta^2 = \{d, e\}$, $\Delta^3 = \{b, c\}$ and $\Delta^4 = \{a\}$. Paradigm Theory assigns the probabilities as follows: $P(n) = P(o) = P(p) = 7/231 \approx 3\%$. $P(j) = \ldots = P(m) = 1/44 \approx 2\%$. $P(f) = \ldots = P(i) = 9/154 \approx 6\%$. $P(d) = P(e) = 45/616 \approx 7\%$. $P(b) = P(c) = 135/1232 \approx 11\%$. $P(a) = 135/616 \approx 22\%$. Notice how the Principle of Defensibility is critical to achieve this assignment. The Basic Theory alone agrees with this assignment on the leaves, but on the others it assigns each a probability of $1/11 \approx 9\%$ instead. The Basic Theory does not notice the structure of the invariant hypotheses and so gives them each the same probability.

This example brings out the importance of the Principle of Defensibility. The Basic Theory can be viewed as a natural generalization of Carnap's m^*-logical theory. Except for cases where the tree is full, the Basic Theory is inadequate, ignoring all the structure that we *know* is there. The Basic Theory's weakness is, as discussed in Subsection 3.1.3, that it is capable of seeing only two degrees of detail. The Principle of Defensibility simply says that among the invariant hypotheses there are, from the point of view of the paradigm already before you (i.e., no secondary paradigm is needed), those that are more and less defensible—notice this. It is this principle that allows Paradigm Theory to break the bonds of a simple generalization of Carnap's m^*-logical theory and secure a full explanation and justification for simplicity-favoring.

Discussion

I am in no way elucidating the difficult question of *What is simplicity?* or *What counts as fewer entities?*; if 'grue' is considered simpler than 'green', then it may well end up with greater prior probability. In this subsection we have discussed why simpler hypotheses, supposing we agree on what this means, should be favored. When one acknowledges—*not favors*—simplicity in the paradigm and the paradigm can be represented as a (full or finite binary, among others) tree, the simpler hypotheses receive higher prior probability. This occurs not because they are simpler, but because they are less arbitrary.

Let me address what could be a criticism of this explanation of the justification of simplicity-favoring. This explanation depends on the resulting graph associated with the paradigm being a tree, with the simpler hypotheses near the root. This occurs because, so I asserted, there are usually multiple ways of complexifying any given hypothesis; and these complexifications are a hypothesis' daughters in the tree. What if this is not true? For example, what if one is presented with a hypothesis set consisting of one simple hypothesis and just one of its complexifications? Acknowledging simplicity here does not entail simplicity-favoring; each hypothesis is equally probable. I claim that holding such a hypothesis set is uncommon and unnatural. Most of the time, if we consider a complexification of a hypothesis and notice that it is a complexification, then we also realize that there are other complexifications as well. Choosing to leave the others out of the hypothesis set and allowing only the one to remain is ad hoc. Worse than this example, suppose for each hypothesis there are multiple *simplifications* rather than multiple complexifications for each hypothesis? If this is so, Paradigm Theory ends up favoring more complex hypotheses instead. While certainly one can concoct hypothesis sets where acknowledging simplicity results in a simplification tree instead of a complexification tree, I do not believe there to be very many (if any) natural examples. And if such a hypothesis set *is* presented to one acknowledging simplicity, the most complex hypothesis is indeed the most favorable. These observations are not unexpected: unusual conceptual frameworks may well entail unusual inductive behavior.

For the sake of contrast it is helpful to look at the reasons I have given in this section for favoring simpler hypotheses compared to those of other theorists: (i) they are more susceptible to falsification (Popper, 1959), (ii) they are more susceptible to confirmation (Quine, 1963), (iii) they are practically easier to apply (Russell, 1918; Pearson, 1992; Mach, 1976), (iv) they have greater *a*

priori likelihood of being true (Jeffreys, 1948), (v) they have been found in the past to be more successful (Reichenbach, 1938), (vi) following the rule 'pick the simplest hypothesis' leads with high probability to true hypotheses (Kemeny, 1953), (vii) they are more informative (Sober, 1975), (viii) they are more stable (Turney, 1990), and (ix) they have higher estimated predictive accuracy (Forster and Sober (1994)). Paradigm Theory's reason for favoring simpler hypotheses is that we acknowledge simplicity and, since for each hypothesis there tends to be multiple complexifications (and not multiple simplifications), simpler hypotheses are less arbitrary.

3.2.4 Curve-Fitting

In curve-fitting the problem is to determine the best curve given the data points. The phenomenon that needs to be explained is that a curve that is a member of an n parameter family, or model,[12] is typically favored over curves that require $n + 1$ parameters, even when the latter fits the data better than the former. I derive within Paradigm Theory a class of information criteria dictating the degree to which a simpler curve (say, a linear one) is favored over a more complex one.

In curve-fitting generally, the data are presumed to be inaccurate, and no hypothesis can be excluded *a priori*. I concentrate only on the hypothesis set of polynomials, and consider only those up to some finite degree. For definiteness I presume that each dimension is bounded to a finite range, and that the underlying measure is uniform in each $M' - M$ (where M' is a model with one more dimension than M). The first of these conditions on the hypothesis set is perhaps the only questionable one. The parameter bounds may be set arbitrarily high, however; so high that it is difficult to complain that the bound is too restrictive.

Suppose we have models M_0 and M_1, M_0 with parameter a_0 and M_1 with parameters a_0 and a_1, where the parameters range over the reals within some bound and the models are such that for some value of a_1, M_1 makes the same predictions as M_0. In cases such as this Jeffreys (1948) (see also Howson, 1987, pp. 210–211] proposes that M_0 and $M_1 - M_0$ each receive prior probability $1/2$. We shall denote M_0 and $M_1 - M_0$ as, respectively, S_0 and S_1 ("S" for symmetry type). Paradigm Theory gives a principled defense for Jeffreys' prior probability assignment: if the conceptual framework acknowledges the

[12] Do not confuse this notion of model with that discussed in Subsection 3.2.1. There is no relation.

two models, then there are two symmetry types—M_0 and $M_1 - M_0$—each receiving prior probability $1/2$ via the Principle of Type Uniformity, and the probability density is uniform over each symmetry type via the Principle of Symmetry.[13]

How the prior probability *density* compares in S_0 and S_1 depends on the choice of underlying measure. Let us first suppose that the measure is the Euclidean one, where length is always smaller than area, area always smaller than volume, etc. Because M_0 is one dimension smaller than M_1, the prior probability density on S_0 is infinitely greater than that on S_1. Thus, any specific curve in S_1 receives prior probability density that is vanishingly small compared to the prior probability of a curve in S_0. More generally, consider M_0, M_1, ..., M_l, where each model is the superset of the previous one resulting from adding one parameter, ranging over the reals within some bound, to allow polynomials of one higher degree. Each subset M_0 and $M_{k+1} - M_k$ for $0 \leq k < l$ is a symmetry type—denoted, respectively, by S_0 and S_{k+1} for $0 \leq k < l$—and receives prior probability $1/(l+1)$. With the Euclidean underlying measure, the probability density over the symmetry types decreases infinitely as the number of extra parameters is increased. Generally, then, curves that are members of models with fewer parameters are *a priori* favored because we possess a conceptual framework that acknowledges the models (and nothing else).

The Euclidean underlying measure is very strong, resulting in simpler curves having greater posterior probability density *no matter the data*. Since each curve in S_0 has a prior probability density that is infinitely greater than each in S_k for $k > 0$, this effectively means that one restricts oneself to the polynomials of least degree. Perhaps less radical underlying measures should be used, ones that agree that higher degrees have greater underlying measure (intuitively, more polynomials), but not *infinitely* greater (intuitively, not infinitely more polynomials). Suppose, instead, that the underlying measure is s in S_0, and m times greater in each successive degree of greater dimension; i.e., S_k has as underlying measure sm^k for some positive real number m. One may find it convenient to act as if the hypothesis set is finite, and that there are (the truncation of) sm^k curves in S_k. Then one can say that a curve in S_k has prior *probability* equal to so and so, rather than probability *density* equal to so and so. At any rate, the important supposition behind the discussion below is that the underlying measure is m times greater as the degree is increased, not whether the hypothesis set is finite or not. Under these conditions, individual curves

[13]Because M_0 is a subset of M_1, elements inside M_0 cannot be interchanged with those outside without affecting the paradigm. This is true regardless of the measure of the two regions.

have probability as stated in the following theorem.

Theorem 12 *Let the hypothesis set $H_{l,s,m}$ be as just described above, i.e., the set of polynomials such that (i) each has degree less than or equal to l, and (ii) $M_k - M_{k-1}$ has a uniform underlying measure equal to sm^k within some finite range. Let Q_{model} be the paradigm that acknowledges the models over $H_{l,s,m}$. If curve h is in S_k for some $0 \leq k \leq l$, then its prior probability density is $\frac{1}{(l+1)sm^k}$.* $\triangle$

The symmetry types S_k each receive prior probability $1/(l+1)$ by the Principle of Type Uniformity. A hypothesis in S_k must share its probability with a measure of sm^k hypotheses, and by the Principle of Symmetry the theorem follows. If one imagines that $H_{l,s,m}$ is finite, then a curve h in S_k receives prior probability equal to $\frac{1}{(l+1)\lfloor sm^k \rfloor}$ (where $\lfloor x \rfloor$ stands for the truncation of x).

One can see from the m^{-k} term that curves requiring a greater number of parameters receive exponentially lower prior probability density. Acknowledge the natural models... exponentially favor polynomials of lower degree. This observation holds regardless of the value of l and s. As for m, larger values mean that curves requiring a greater number of parameters are more disfavored.

There are a class of curve-fitting techniques called "information criteria" which prescribe picking the model that has the largest value for $\log L_k - \gamma k$, where k is the number of parameters of the model, $\log$ is the natural logarithm, L_k is the likelihood ($P(e|h)$) of the maximum likely hypothesis in the model of k^{th} dimension M_k, and γ depends on the specific information criterion. [See Smith and Spiegelhalter (1980, pp. 218) for many of the information criteria (my γ is their $m/2$) and references to the original papers defending them; see also Aitkin (1991).] Once this model is determined, the maximum likely hypothesis in it is chosen, even though it may well not be the maximum likely hypothesis in the entire hypothesis set. Paradigm Theory natural leads to a class of information criteria emanating from the supposition that the paradigm is Q_{model} and the underlying measure of S_{k+1} is m times greater than that of S_k.

Our task is now to find the curve, or hypothesis, with the greatest posterior probability density given that models M_0 through M_l are acknowledged in the paradigm (i.e., Q_{model} is the paradigm). For simplicity, I will for the moment treat $H_{l,s,m}$ as if it is finite, with (the truncation of) sm^k curves in S_k. We want to find h such that it maximizes, via Bayes' Theorem, $P(e|h)P(h)/P(e)$ (e represents the data); that is, we wish to find h with maximum posterior probability. It suffices to maximize the natural logarithm of the posterior probability,

or

$$\log P(e|h) + \log P(h) - \log P(e).$$

$P(e)$ is the same for every hypothesis, and we may ignore it. Theorem 12 informs us of the $P(h)$ term, which is the prior probability of h given Q_{model}, and we have

$$\log P(e|h) + \log[\frac{1}{(l+1)sm^k}]$$

if h is in S_k. This manipulates easily to

$$\log P(e|h) - (\log m)k - \log(l+1) - \log s.$$

l and s are the same for each hypothesis, and so they may also be ignored. This allows l, the maximum degree of polynomials allowed in the hypothesis set, to be set arbitrarily high. When the hypothesis set is treated as finite, s can be set arbitrarily high, thereby allowing the set to approximate an infinite one. Thus, the hypothesis with the maximal posterior probability is the one that maximizes

$$\log P(e|h) - (\log m)k.$$

This may be restated in the information criterion form by saying that one should choose the model that has the largest value for

$$\log L_k - (\log m)k,$$

and then choose the maximum likely hypothesis in that model. I have just proven the following theorem, which I state for records sake, and retranslate into its corresponding infinite hypothesis set form.

Theorem 13 *Let the hypothesis set be $H_{l,s,m}$ and the paradigm be Q_{model}; let the prior probability distribution be determined by Paradigm Theory. The hypothesis with the greatest posterior probability density is determined by choosing the model with the largest value for* $\log L_k - (\log m)k$ *and then picking the maximum likely hypothesis in that model.* $\triangle$

Notice that $\log m$ is filling the role of the γ in the information criteria equation. As m increases, goodness of fit is sacrificed more to the simplicity of the curves requiring fewer parameters since the number of parameters k gets weighed more heavily.

Consider some particular values of m. $m < 1$ means that the underlying measure of S_{k+1} is *less* than that of S_k; that there are, informally, fewer polynomials of the next higher degree. This is very unnatural, and the corresponding information criterion unnaturally favors more complex curves over simpler ones. $m = 1$ implies that moving to higher dimensions does not increase the underlying measure at all. In this case, the second term in the information criterion equation becomes zero, collapsing to the Maximum Likelihood Principle. When moving up in degree and dimension, it is only natural to suppose that there are, informally, more polynomials of that degree. With this in mind, it seems plausible that one chooses $m > 1$. $m = 2$ implies that moving to the next higher dimension doubles the underlying measure, which intuitively means that the number of hypotheses in S_{k+1} is twice as much as in S_k. The value of γ for $m = 2$ is $\gamma = \log m \approx .69$. Smith and Spiegelhalter (1980, pp. 219) observe that when $\gamma < .5$ more complex models still tend to be favored, and this does not fit our curve-fitting behavior and intuition; it is pleasing that one of the first natural values of m behaves well. (My γ is Smith and Spiegelhalters' $m/2$. Their m is not the same as mine.) When $m = e$, the resulting information criterion is precisely Akaike's Information Criterion. This amounts to a sort of answer to Forster and Sobers' (1994, p. 25) charge, "But we do not see how a Bayesian can justify assigning *priors* in accordance with this scheme," where by this they mean that they do not see how a prior probability distribution can be given over the curves such that the resulting information criterion has $\gamma = 1$. Paradigm Theory's answer is that if one acknowledges the natural models, and one assigns underlying measures to degrees in such a way that the next higher degree has e times the underlying measure of the lower degree, then one curve-fits according to Akaike's Information Criterion. When $m = 3$, $\gamma \approx 1.10$, and the resulting inductive method favors simpler curves just slightly more than in Akaike's. Finally, as $m \to \infty$, the underlying measure on $M_1 - M_0$ becomes larger and larger compared to that of M_0, and all curves requiring more than the least allowable number of dimensions acquire vanishingly small prior probability density; i.e., it approaches the situation in Jeffreys' prior discussed above. (There is also a type of Bayesian Information Criterion, called a "global" one (Smith and Spiegelhalter, 1980), where $\gamma = (\log n)/2$ and n is the number of data (Schwarz, 1978).)

The question that needs to be answered when choosing a value for m is, "How many times larger is the underlying measure of the next higher degree?," or intuitively, "How many times more polynomials of the next higher degree are to be considered?" Values for m below 2 seem to postulate too few polyno-

mials of higher degree, and values above, say, 10 seem to postulate too many. The corresponding range for γ is $.69$ to 2.30, which is roughly the range of values for γ emanating from the information criteria (Smith and Spiegelhalter, 1980). For these "non-extreme" choices of m, curves requiring fewer parameters quickly acquire maximal posterior probability so long as their fit is moderately good.

Paradigm Theory's explanation for curve-fitting comes down to the following: We favor (and ought to favor) lines over parabolas because we acknowledge lines and parabolas. The reasonable supposition that the hypothesis set includes more curves of degree $k+1$ than k is also required for this explanation.

Paradigm Theory's class of information criteria avoids at least one difficulty with the Bayesian Information Criteria. The Personalistic Bayesian does not seem to have a principled reason for supposing that the prior probabilities of M_0, $M_1 - M_0$, etc., are equal (or are any particular values). Why not give M_0 much more or less prior probability than the others? Or perhaps just a little more or less? In Paradigm Theory the models induce M_0, $M_1 - M_0$, etc., as the symmetry types, and the Principle of Type Uniformity sets the priors of each equal.

Another advantage to Paradigm Theory approach is that the dependence on the models is explicitly built in through the paradigm. *Any* choice of subsets is an allowable model choice for Paradigm Theory.

3.2.5 Bertrand's Paradox

Suppose a long straw is thrown randomly onto the ground where a circle is drawn. Given that the straw intersects the circle, what is the probability that the resulting chord is longer than the side of an inscribed equilateral triangle (call this event B). This is Bertrand's question (Bertrand, 1889, pp. 4–5). The Principle of Indifference leads to very different answers depending on how one defines the hypothesis set H.

H_0 If the hypothesis set is the set of distances between the center of the chord and the center of the circle, then the uniform distribution gives $P(B) = 1/2$.

H_1 If the hypothesis set is the set of positions of the center of the chord, then the uniform distribution gives $P(B) = 1/4$.

H_2 If the hypothesis set is the set of points where the chord intersects the circle, then the uniform distribution gives $P(B) = 1/3$.

Kneale (1949, pp. 184–188) argues that the solution presents itself once the actual physical method of determining the chord is stated, and a critique can be found in Mellor (1971, pp. 136–146). Jaynes (1973) presents a solution which I discuss more below. Marinoff (1994) catalogues a variety of solutions in a recent article. I approach Bertrand's Paradox in two fashions.

Generalized Invariance Theory

In the first Paradigm Theory treatment of Bertrand's Paradox I take the hypothesis set to be the set of all possible prior probability distributions over the points in the interior of the circle—each prior probability distribution just *is* a hypothesis. To alleviate confusion, when a hypothesis set is a set of prior probability distributions over some other hypothesis set, I call it a *prior set*; I denote the elements of this set by ρ rather than h, and denote the set H_ρ.

I wish to determine a prior probability assignment on H_ρ. What "should" the paradigm be? Jaynes (1973) argues that the problem statement can often hold information that can be used to determine a unique distribution. In the case of Bertrand's Problem, Jaynes argues that because the statement of the problem does not mention the angle, size, or position of the circle, the solution must be invariant under rotations, scale transformations, and translations. Jaynes shows that there is only one such solution (in fact, translational invariance alone determines the solution), and it corresponds to the H_0 case above, with $P(B) = 1/2$: the probability density in polar coordinates is

$$\mathcal{P}(r,\theta) = \frac{1}{2\pi R r},\ \ 0 \leq r \leq R,\ \ 0 \leq \theta \leq 2\pi$$

where R is the radius of the circle. The theory sanctioning this sort of determination of priors I call the *Invariance Theory* (see Changizi and Barber, 1998).

I will interpret the information contained in the problem statement more weakly. Instead of picking the prior distribution that has the properties of rotation, scale, and translational invariance as Jaynes prescribes, suppose one merely *acknowledges* the invariance properties. That is, the paradigm is comprised of the subsets of prior probability distributions that are rotation, scale, and translation invariant, respectively. For every non-empty logical combination of the three properties besides their mutual intersection there are continuum many hypotheses. Supposing that each subset of the prior set corresponding to a logical combination of the three properties has a different measure, Paradigm Theory induces five symmetry types: $T \cap R \cap S$, $\neg T \cap R \cap S$, $R \cap \neg S$, $\neg R \cap S$ and $\neg R \cap \neg S$ (three logical combinations are empty), where T, R and S

denote the set of translation-, rotation- and scale-invariant priors, respectively. Each receives prior probability $1/5$, and since $T \cap R \cap S = \{\frac{1}{2\pi Rr}\}$ and the other symmetry types are infinite, $P(\frac{1}{2\pi Rr}) = 1/5$ and every other prior receives negligible prior probability; $1/(2\pi Rr)$ is the clear choice. In as much as the properties of this paradigm are objective, being implicitly suggested by the problem, this solution is objective.[14]

This "trick" of using Paradigm Theory parasitically on the Invariance Theory can be employed nearly whenever the latter theory determines a unique invariant distribution; and in all but some contrived cases the unique distribution is maximally probable. Some contrived cases may have it that, say, in the prior set ρ_1 is the unique prior that is scale and rotation invariant (where I suppose now that these are the only two properties in the paradigm), but that there is exactly one other prior ρ_2 that is neither scale nor rotation invariant (and there are infinitely many priors for the other two logical combinations). Here there are at most four symmetry types, $\{\rho_1\}$, $\{\rho_2\}$, $R \cap \neg S$ and $\neg R \cap S$. Each of these two priors receives prior probability $1/4$, and so ρ_1 is no longer the maximally probable prior.

Now, as a matter of fact, the invariance properties people tend to be interested in, along with the prior sets that are typically considered, have it that there are infinitely many priors that are not invariant under any of the invariance properties. And, if the Invariance Theory manages to uniquely determine a prior, there are almost always going to be multiple priors falling in every logical combination of the invariance properties except their mutual intersection. If this is true, then Paradigm Theory's induced symmetry types have the unique prior as the only prior alone in a symmetry type, i.e., it is the only *invariant* prior in Paradigm Theory's definition as well. Given that this is so, by Theorem 2 this prior has the greatest prior probability.

Paradigm Theory need not, as in the treatment of Bertrand's Problem above, give infinitely higher prior probability to the unique invariant prior than the others, however. Suppose, for example, that the Invariance Theory "works" in that there is exactly one prior ρ_0 that is both scale and rotation invariant, but that there are exactly two priors ρ_1 and ρ_2 that are scale invariant and not rotation invariant, exactly three priors ρ_3, ρ_4 and ρ_5 that are rotation and not scale invariant, and infinitely many priors that are neither (again, where only rotation and scale invariance are the properties in the paradigm). There are now four

[14] And the solution seems to be correct, supposing the frequentist decides such things, for Jaynes claims to have repeated the experiment and verified that $P(B) \approx 1/2$, although see Marinoff's comments on this (Marinoff, 1994, pp. 7–8).

symmetry types, each receiving prior probability $1/4$. The probability of the unique invariant prior is $1/4$, that of each of the pair is $1/8$, and that of each of the triplet is $1/12$. The point I mean to convey is that *Paradigm Theory not only agrees with the Invariance Theory on a very wide variety of cases, but it tells us the degree to which the Invariance Theory determines any particular prior.* In this sense Paradigm Theory brings more refinement to the Invariance Theory. In the cases where Paradigm Theory does not agree with the Invariance Theory, as in the "contrived" example above, there is a principled reason for coming down on the side of Paradigm Theory *if* the invariance properties are just *acknowledged* and not favored. Also, not only can Paradigm Theory be applied when the Invariance Theory works, it can be applied when the Invariance Theory fails to determine a unique prior; in this sense, Paradigm Theory allows not only a refinement, but a sort of generalization of the Invariance Theory.

H is the Sample Space

The second way of naturally approaching Bertrand's Paradox within Paradigm Theory takes the hypothesis set to be the set of possible outcomes of a straw toss. In determining the hypothesis set more precisely, one informal guide is that one choose the "most general" hypothesis set. This policy immediately excludes H_0 (see the beginning of this subsection) since it does not uniquely identify each chord in the circle. H_1 and H_2 are each maximally general and are just different parametrizations of the same set. I choose H_1 as, in my opinion, the more natural parametrization, with the underlying measure being the obvious Euclidean area.

What "should" the paradigm be? The problem has a clear rotational symmetry and it would seem very natural to acknowledge the distance between the center of the chord and the center of the circle; this set of distances just *is* H_0 and we will be "packing in" H_0 into the paradigm. Rather than acknowledging *all* of the distances, suppose that one acknowledges n of them (n equally spaced concentric rings within the circle); we will see what the probability distribution looks like as n approaches infinity. Each ring has a different area, and so each is its own symmetry type. Therefore each has a probability of $1/n$. The probability density is

$$\mathcal{P}(r,\theta) = \frac{1/n}{A_i} = \frac{1/n}{(2i-1)\pi R^2/n^2} = \frac{n}{(2i-1)\pi R^2},\ r \in [iR/n, (i+1)R/n],$$

where, $i = 1, \ldots, n$, A_i is the area of the i^{th} concentric ring from the center.

As n gets large, $iR/n \approx r$, so $i \approx rn/R$. Thus

$$\mathcal{P}(r,\theta) = \frac{n}{(2rn/R - 1)\pi R^2} = \frac{n}{(rn - R/2)2\pi R}$$

and since n is large, $rn - R/2 \approx rn$, giving

$$\mathcal{P}(r,\theta) = \frac{1}{2\pi Rr}$$

which is exactly what Jaynes concludes. Acknowledge how far chords are from the center of the circle and accept one of the more natural parametrizations... get the "right" prior probability density function.

If instead of acknowledging the distance from the center of the circle one acknowledges the property of being *within* a certain radius, then the sets in the paradigm are nested and the resulting symmetry types are the same as before, regardless of the underlying measure.

3.3 "Solution" to riddle and theory of innateness

The intuition underlying Paradigm Theory is that *more unique is better*, or *arbitrariness is bad*, and this is related to the idea that *names should not matter*, which is just a notion of symmetry. The more ways there are to change a hypothesis' name without changing the structure of the inductive scenario (i.e., without changing the paradigm), the more hypotheses there are that are just like that hypothesis (i.e., it is less unique), which means that there is less "sufficient reason" to choose it. The principles of Paradigm Theory link with this intuition. The Principle of Type Uniformity and Principle of Symmetry give more unique hypotheses greater prior probability, and the Principle of Defensibility entails that among the more unique hypotheses, those that are more unique should receive greater prior probability. Recall (from Subsection 3.1.3) that these are the *links* of the principles to the "more unique is better" motto—the principles do not actually *say* anything about the uniqueness of hypotheses, but are motivated for completely different, compelling reasons of their own. Nevertheless, it is a convenient one-liner to say that Paradigm Theory favors more unique hypotheses, and not just qualitatively, but in a precise quantitative fashion. In this sense the theory is a quantitative formalization of Leibniz's Principle of Sufficient Reason, interpreted nonmetaphysically only.

The favoring of more unique hypotheses, despite its crudeness, is surprisingly powerful, for it is a natural, radical generalization of both Carnap's

m^*-logical theory and (through the use of secondary paradigms) Hintikka's "$\alpha = 0$"-logical theory, arguably the most natural and pleasing inductive methods from each continuum. Besides these achievements, Paradigm Theory gives explanations for a large variety of inductive phenomena:

- it "correctly" collapses to the Classical Theory's Principle of Indifference when no distinctions are made among the hypotheses,
- it suggests a conceptual framework-based solution to the problem of the underdetermination of interpretation for language,
- it explains why no-inductions are rarely considered rational,
- it explains why frequency-inductions and law-inductions *are* usually considered rational,
- it gives a foundation for Occam's Razor by putting forth the notion that simpler hypotheses are favored because one acknowledges simplicity, and simpler hypotheses are (usually) less arbitrary,
- it accommodates curve-fitting by supposing only that one acknowledges the usual models—constants, lines, parabolas, etc.,
- it allows a sort of generalization of the Invariance Theory for choosing unique prior probability distributions, and this is used to solve Bertrand's Paradox,
- and it accounts for Bertrand's Paradox in another fashion by acknowledging the distance from the center of the circle.

In the first section of this chapter I laid out the goals of a theory of logical induction, and the related goals of a theory of innateness. How does Paradigm Theory fare in regard to these goals?

How Paradigm Theory "solves" the riddle of induction

Let us briefly recall our basic aim for a logical theory of induction. Ultimately, we would like to reduce all oughts in induction and inference—you *ought* to choose the simplest hypothesis, you *should* believe the next fish caught will be a bass, it is *wrong* to draw a parabola through three collinear points, and so on—to just a small handful of basic, axiomatic, or primitive oughts. The hope is that all oughts we find in induction can be derived from these primitive oughts. We would then know, given just a set of hypotheses and the evidence, exactly what degrees of confidence we should have in each hypothesis. If we had this, we would have a solution to the riddle of induction.

Alas, as discussed at the start of this chapter, this is impossible; there is no solution to the riddle of induction. There are, instead, multiple inductive methods, and although some may well be irrational or wrong, it is not the case that there is a single right inductive method. This was because any inductive method makes what is, in effect, an assumption about the world, an assumption which is left hanging without defense or justification for why *it* should be believed.

If we are to have a theory of logical induction, we must lower the bar. We would *still*, however, like the theory to consist of a small handful of primitive oughts. But we are going to have to resign ourselves to the persistence of a leftover variable of some kind, such that different settings of the variable lead to different inductive methods. A theory of logical induction would, at best, allow statements of the form

> If the variable is X, then the primitive oughts entail that one should proceed with inductive method M.

But it would defeat the whole purpose of our theory if this variable stood for variable *a priori* beliefs about the world, because the theory would then only be able to say that if you started out believing X, then after seeing the evidence you should believe Y. We want to know why you should have started out believing X in the first place. How did you get *those* beliefs in ignorance about the world?

And this was the problem with the Bayesian approach. The Bayesian approach was good in that it declares a primitive ought: one should use Bayes' Theorem to update probabilities in light of the evidence. And to this extent, Paradigm Theory also utilizes Bayes' Theorem. But the Bayesian approach leaves prior probabilities left over as a free-for-the-picking variable, and priors are just claims about the world.

With this in mind, we required that the variable in any successful theory of logical induction not stand for beliefs or claims about the world. Because any choice of the variable leads, via the primitive principles of ought, to an inductive method, any choice of variable ends up entailing a claim about the world. But that must be distinguished from the variable itself being a claim about the world. We required that the variable have some meaningful, non-inductive interpretation, so that it would be meaningful to say that an inductive agent entered the world with a setting of the variable but nevertheless without any *a priori* beliefs about the world. We would then say that the agent, being rational, should follow the primitive principles of ought and thereby end up

with what are claims about the world. But the claims about the world were not inherent to the variable, they only come from joining the variable with the principles of ought.

In this chapter I introduced a kind of variable called a "paradigm," which is central to Paradigm Theory. Paradigms are not about the world. Instead, they are *conceptualizations* of the world, and more exactly, conceptualizations of the space of hypotheses. Paradigms say which hypotheses are deemed to be similar to one another, and which are not. More precisely, a paradigm is the set of properties of hypotheses the inductive agent acknowledges. The set of properties of hypotheses acknowledged does not comprise a claim about the world, nor does it possess any 'ought's. It is just a way of looking at the set of hypotheses, and no more than that. Paradigms, then, are non-inductive and have a meaningful interpretation. This is the kind of variable we wanted in a theory of logical induction.

But we also needed principles capable of taking us from the variable—the paradigm in Paradigm Theory—to an inductive method. Paradigm Theory achieves this via three primitive principles of ought, along with the Bayesian principle of evidence (Bayes' Theorem). The three principles concern non-arbitrariness in the assignment of prior probabilities, and given a paradigm the principles entail a unique prior probability distribution. The Bayesian principle of evidence finishes the job by stating how one ought to modify prior probabilities to posterior probabilities as evidence accumulates. In sum, Paradigm Theory allows statements like this:

> If, before knowing anything about the world, you conceptualize the space of hypotheses in a fashion described by paradigm Q, then via the three primitive principles of prior probability determination you should have certain prior probabilities $P_Q(h)$ on those hypotheses. And, furthermore, when evidence is brought to bear on the logical probabilities of the hypotheses, one should obtain posterior probabilities by using Bayes' Theorem.

The most important thing to notice about this is that the statement begins with the inductive agent *not* making any claim about the world. The statement does not simply say that if you have certain beliefs you ought to have certain others. It requires only that the completely-ignorant-about-the-world inductive agent enter the world with a way of looking at it. Without any preconceptions about the world (although he has preconceptions about the properties of hypotheses), the theory nevertheless tells the agent how he ought to proceed with induction. The theory thereby reduces all inductive oughts to a few primitive principles of

ought, and these primitive oughts are the *only* inductive primitives one needs for a theory of induction. *At base, to justifiably follow an inductive method is to have a paradigm and to follow certain abstract principles of non-arbitrariness and principles of evidence.*

Some readers might say that this is all well and good, but does it really get us anywhere? We are still stuck with paradigms, and there is no way to justify why an inductive agent has the paradigm he has. We have simply pushed the indeterminacy of inductive methods downward, to prior probabilities, and then further downward to paradigms. We still have not answered the question of which inductive method we should use, because we have not given any reason to pick any one paradigm over another. That is, suppose that—poof—a rational, intelligent agent suddenly enters a universe. We still do not know what he should do in regards to learning, and so Paradigm Theory is useless for him.

The response to this kind of criticism is that it is essentially taking Paradigm Theory to task for not being a solution to the riddle of induction. To see this, note that the criticism can be restated as, "If Paradigm Theory is so great, why isn't it telling us what one should believe given just the hypothesis set and the evidence?" But this is just to ask why Paradigm Theory does not solve the riddle of induction. The answer, of course, is that there is no solution to the riddle of induction; i.e., there is no single way that one ought to take a set of hypotheses and evidence and output posterior probabilities in the hypotheses. This kind of criticism has forgotten to lower the bar on what we should be looking for in a theory of logical probability. At best, we can only expect of a theory of logical probability that it reduce inductive oughts to a small number of primitive ones, and to some variable that is not about the world. We cannot expect to have no variable left over.

It should be recognized that it was not *prima facie* obvious, to me at least, that it would even be possible to obtain this lowered-bar theory of logical induction. *Prima facie*, it seemed possible that there would be no way, even in principle, to reduce inductive oughts to a few primitive oughts and some meaningful, non-inductive variable. Paradigm Theory is an existence proof: a lowered-bar theory of logical probability exists. I have presented no argument that no other theory of logical probability could not also satisfy these requirements I have imposed; there probably exist other such theories, perhaps others with superiorities over Paradigm Theory.

How Paradigm Theory serves as a theory of innateness

Paradigm Theory provides a kind of best-we-can-hope-for solution to the riddle of induction. But I had also stated at the start of this chapter that we were simultaneously looking for a theory that would serve as a theory of innateness, and I had put forth requirements we demanded of such a theory. The requirements were that we be able to model rational intelligent agents as following certain fixed learning principles, and that any innate differences in their resultant inductive method would be due to some setting of a variable with a weak, non-inductive, meaningful interpretation. Under the working assumption that the brain is rational, the theory would apply to the brain as well. The theory of innateness would provide a way of economically explaining why different agents—or different kinds of brain—innately engage in different inductive methods. We would not have to commit ourselves to a belief that the principles of learning may be innately chosen willy nilly; there is a single set of learning principles that anyone ought to follow. We would also not be committed to a view that brains enter the world with *a priori* beliefs about it, a view that seems a little preposterous. Instead, brains would only have to innately be equipped with some other kind of difference, although what that difference might be will depend on the kind of theory of innateness that is developed.

Recall that the Bayesian framework is a nice step forward in this regard, and has accordingly been taken up in psychology, neuroscience, computer and the decision sciences to study learning and interactions with an uncertain world. All innate differences in inductive methods will be due *not* to innate differences in how evidence is to be used to modify the degrees of confidence in hypotheses. All innate differences stem from innate differences in prior probabilities, and here lies the problem with the Bayesian framework as a theory of innateness: prior probabilities are *a priori* beliefs about the world, and thus they are not non-inductive, as we require for a theory of innateness.

The Bayesian framework should not, however, be abandoned: it gets the evidence principle right. What we would like is to dig deeper into prior probabilities and find principles of prior probability determination that any agent should follow, so that from some non-inductive innate variable comes prior probabilities via these principles. And this is where Paradigm Theory enters. Objects called “paradigms” were introduced which were interpreted as conceptual frameworks, or ways of conceptualizing the space of hypotheses. Paradigms were not about the world. Paradigm Theory introduced principles of prior probability determination saying how, given a paradigm, one ought to

assign prior probabilities. Paradigm Theory, then, appears to satisfy the requirements of a theory of innateness.

But, someone might criticize, we are still left with innate paradigms, or innate ways of conceptualizing the set of hypotheses, or innate ways of lumping some hypotheses together as similar or of the same type. Is this any better than innate prior probabilities? Perhaps it is strange to hypothesize that brains have *a priori* beliefs about the world, but is it not also strange to hypothesize that brains have *a priori* ways of carving up the space of hypotheses?

As a response, let me first admit that it *is*, *prima facie*, a bit strange. However, one has to recognize that if a brain engages in an inductive method, then it *must* have entered the world with *some* innate structure that is sufficient to entail the inductive method. Such innate "structure" will either be learning algorithms of some kind unique to that kind of brain, or perhaps the Bayesian evidence principle along with prior probabilities unique to that kind of brain, or perhaps the Bayesian evidence principle and principles of prior probability determination along with a paradigm unique to that kind of brain, etc. It may seem *prima facie* odd to believe that *any* of these kinds of "structures" could be innate. One reason for this first reaction to innate structures is that there is, I believe, a tendency to revert to thinking of brains as blank slate, universal learning machines: brains enter the world completely unstructured, and shape themselves by employing universal learning algorithms to figure out the world. But as we have discussed, there is no universal learning algorithm, and so there cannot be brains that enter the world without innate learning-oriented structures. We are, then, stuck with innate learning-oriented structure, no matter how strange that might seem. Thus, the fact that innate paradigms strike us as strange is not, alone, an argument that Paradigm Theory is supposing something outlandish.

But, one may ask, are paradigms any less outlandish than prior probabilities? What have we gained by moving from innate structure in the form of prior probabilities to innate structure in the form of paradigms? We have gained in two ways. First, we have isolated further principles of rationality that inductive agents ought to follow; namely, the non-arbitrariness principles of prior probability determination (the Principles of Type Uniformity, Symmetry and Defensibility). Second, paradigms are a much weaker innate structure, being only about the kinds of hypotheses there are, rather than about the degree of confidence in the hypotheses.

Note that Paradigm Theory as a theory of innateness is not necessarily committed to actual innate paradigms in the head, whatever that might mean. It is

commonplace for researchers to hypothesize that different kinds of organisms have what are, in effect, different innate prior probabilities, but such researchers do not commit themselves to any view of what mechanisms may instantiate this. Prior probabilities are primarily a theoretical construct, and allow us to understand brains and learning agents within the Bayesian framework. Similarly, Paradigm Theory is not committed to any particular mechanism for implementing innate paradigms. Rather, paradigms are a theoretical construct, allowing us to describe and explain the behaviors of inductive agents and brains in an economical fashion.

Paradigm Theory is *a* theory of innateness satisfying the requirements we set forth, but there is no reason to believe there are not others also satisfying the requirements, perhaps better theories in many ways.

Appendix to chapter: Some proofs

This section consists of some proofs referred to in this chapter.

Here are some definitions. $\delta(h) = \gamma$ (*h is γ-Q-invariant in H*) if and only if $h \in \Delta^\gamma$. $\delta(h)$ is the ordinal number indicating the invariance level of h. Say that t is a *Q^γ-symmetry type* in H if and only if t is a $Q \sqcap H^\gamma$-symmetry type in H^γ. Let κ_n be the cardinality of H^n (which is also the number of singleton Q^{n-1}-symmetry types), let s_n be the number of non-singleton Q^n-symmetry types, and let $e(h)$ be the cardinality of the Q-equivalence type of h. Notice that $\kappa_{n+1} = card(I(Q^n, H^n))$ ('$card(A)$' denotes the cardinality of set A). We denote $\frac{\kappa_{i+1}}{s_i+\kappa_{i+1}}$ by r_i and call it the *singleton symmetry type ratio at level i*. The following theorem states some of the basic properties of Paradigm Theory.

Theorem 14 *The following are true concerning Paradigm Theory.*

1. $P(H^{n+1}) = r_n P(H^n)$ $(P(H^0) = 1)$.
2. $P(H^{n+1}) = r_0 r_1 \cdots r_n$.
3. $P(\Delta^n) = (1 - r_n)P(H^n)$.
4. $P(h) = \frac{r_{\delta(h)}}{e(h)\kappa_{\delta(h)+1}} P(H^{\delta(h)})$.
5. $P(h) = \frac{r_0 \cdots r_{\delta(h)}}{e(h)\kappa_{\delta(h)+1}}$.

Proof. Proving 1, there are $s_n + \kappa_{n+1}$ Q^n-symmetry types, and κ_{n+1} of them are singletons which "move up" to the $n+1^{\text{th}}$ level. Since each Q^n-symmetry type gets the same probability, H^{n+1} gets the fraction

$$\frac{\kappa_{n+1}}{s_n + \kappa_{n+1}}$$

of the probability of H^n. 2 is proved by solving the recurrence in 1. 3 follows from 1 by recalling that $P(\Delta^n) = P(H^n) - P(H^{n+1})$. To prove 4, notice that the probability of a hypothesis h is

$$\frac{P(\Delta^{\delta(h)})}{s_{\delta(h)} e(h)}.$$

Substituting $P(\Delta^{\delta(h)})$ with the formula for it from 3 and some algebraic manipulation gives the result. Finally, 5 follows from 2 and 4. $\triangle$

Proof of Theorem 3. To prove 1, it suffices to show that for all i, $\frac{P(\Delta^i)}{s_i} \leq \frac{P(\Delta^{i+1})}{s_{i+1}}$. By Theorem 14,

$$P(\Delta^i) = \frac{s_i}{s_i + \kappa_{i+1}} P(H^i)$$

and

$$P(\Delta^{i+1}) = \frac{s_{i+1}}{s_{i+1} + \kappa_{i+2}} P(H^{i+1}) = \frac{s_{i+1}}{s_{i+1} + \kappa_{i+2}} \frac{\kappa_{i+1}}{s_i + \kappa_{i+1}} P(H^i)$$

By substitution we get

$$\frac{P(\Delta^{i+1})}{s_{i+1}} = \frac{P(\Delta^i)}{s_i} \frac{\kappa_{i+1}}{s_{i+1} + \kappa_{i+2}}$$

It therefore suffices to show that

$$1 \leq \frac{\kappa_{i+1}}{s_{i+1} + \kappa_{i+2}},$$

and this is true because the denominator is the total number of Q^{i+1} symmetry types, which must be less than or equal to the numerator, which is the total number of hypotheses in H^{i+1}. 2 follows easily from 1. $\triangle$

It is not the case that less defensible equivalence types always have less probability. It is also not the case that more defensible hypotheses never have lower probability than less defensible hypotheses. A more defensible hypothesis h_1 can have less probability than a less defensible hypothesis h_2 if the equivalence type of h_1 is large enough compared to the equivalence type of h_2. The following theorem states these facts.

Theorem 15 *The following are true about Paradigm Theory.*

1. *There are equivalence types d_1 less defensible than d_2 such that $P(d_1) = P(d_2)$.*
2. *There are hypotheses h_1 not more defensible than h_2 such that $P(h_1) \not\leq P(h_2)$.*

Proof. To prove 1, consider a paradigm represented by a two-leaf binary tree. The root comprises one equivalence type, and the pair of leaves is the other. Each equivalence type is also a symmetry type here, and so each gets probability $1/2$.

Proving 2, consider the tree on H_f from Section 3.1.2. The reader may verify that h and i receive probability $\frac{1}{18}$, but e, f, and g receive probability $\frac{14}{15}\frac{1}{18} < \frac{1}{18}$. $\triangle$

Proof of Theorem 4. When n is even there are $2n$ complexions and Laplace's method gives each a probability of $1/2n$. For each complexion there is a symmetrical one with respect to Q_L with which it may be permuted (without changing Q_L), so there are n symmetry types, each receiving via Q_L a probability of $1/n$. Each symmetry type contains exactly two complexions of equal size, and so each complexion gets a probability assigned of $1/2n$. (The non-complexion set in Q_L does not come into play when n is even.)

When n is odd there are $2n - 1$ complexions and Laplace's method gives each a probability of $1/(2n-1)$. Now there are an odd number of complexions, and the "middle" one is not symmetrical with any other complexion. Furthermore, because Q_L contains the set of all sequences with more 0s than 1s, and this set is asymmetrical, none of the complexions are symmetrical with any others. Thus, each complexion is a symmetry type, and each complexion receives a probability of $1/(2n - 1)$. $\triangle$

Proof of Theorem 7. There are $2^N - 2$ sequences that are not predicted by the 'all 1s' or 'all 0s' laws, and these must share the $.5$ prior probability assignment.

There are $2^{N-n}-1$ sequences of length N with the first n experiments resulting in 1 but not all the remaining $N-n$ experiments resulting in 1; the total prior probability assigned to these strings is therefore

$$q = \frac{1}{2}\frac{2^{N-n}-1}{2^N-2}.$$

The probability that after seeing n 1s there will be a counterexample is

$$\frac{q}{.25+q}.$$

With some algebra, the probability that after seeing n 1s the remaining will all be 1 is

$$\frac{1}{2}\frac{2^N-2}{2^N(2^{-n}+2^{-1})-2},$$

which, for any moderately sized N becomes, with some algebra, approximately

$$\frac{2^{n-1}}{1+2^{n-1}}.\triangle$$

Proof of (a) in Theorem 9. We want the probability that $p=1$ given that we have seen n 1s and no 0s ($n>0$); i.e., $P(p=1|1^n)$, where 1^n denotes the string with n 1s. By Bayes' Theorem

$$P(p=1|1^n) =$$

$$\frac{P(p=1)P(1^n|p=1)}{P(p=1)P(1^n|p=1)+P(p\in(0,1))P(1^n|p\in(0,1))+P(p=0)P(1^n|p=0)}.$$

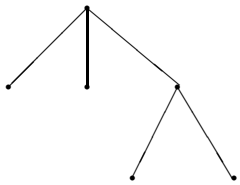
The only term that is not immediately obvious is $P(1^n|p\in(0,1))$, which is $\int_0^1 p^n dp = 1/(n+1)$. Thus we have

$$\frac{.25(1)}{.25(1)+.5\frac{1}{n+1}+.25(0)},$$

and with a little manipulation this becomes $\frac{n+1}{n+3}$. $\triangle$

Proof Sketch of Theorem 10. 1 is simple and 2 is proved by induction on the depth of the binary tree. 1 and 2 do not exhaust the types of trees that result in the root being the lone maximally defensible element; see the H_{asymm}/Q_{asymm} example in Section 3.2.3 for a non-binary non-full tree that puts the root alone at the top. Informally, most trees put the root at the top. We have made no

attempt to characterize the class of trees that put the root at the top. See the following tree for 3.



△

Proof of Theorem 11. There are $n+1$ symmetry types (one for each level), each receiving probability $1/(n+1)$. The symmetry type at depth i has $\ddot{\theta}$ elements. △