A Paradigm-Based Solution to the Riddle of Induction

Mark Changizi[*‡]
Department of Computer Science
National University of Ireland
Cork, Ireland
changizi@cs.ucc.ie

Timothy P. Barber
Keynetic Systems LLC
11931 Chalon Lane
San Diego, CA 92128
tbarber@san.rr.com

Contents

1 Introduction

The goal of a theory of induction as we address it is to explain how our inductive methods and conclusions are justified. In this paper we present our own theory of the justification of induction, which we have labeled the Paradigm Theory of Induction. It is different than many previous theories which aim to provide "just the right" principles to justify this or that inductive method. We believe there can be no such single justification of induction; a theory of the justification of induction needs to be versatile. Our Paradigm Theory explicitly makes our inductive methods dependent on the way we "carve up the world," or a little more precisely, the properties of hypotheses we "acknowledge"; we call this a conceptual framework, or paradigm.[1] Acknowledge certain properties of hypotheses... acquire a unique prior probability distribution and thus a precise inductive method via conditionalizing with Bayes' Theorem. From the fact that it *is* the case that you acknowledge such and such properties it follows that you *ought* to have such and such prior degrees of belief (our interpretation of probability throughout this paper is a logical one). Of course, one cannot move from 'is' to 'ought' for free; we present and motivate several benign, compelling principles of rationality that are responsible for this.

A major thesis of the paper is that conceptual frameworks can replace Personalistic Bayesianism's subjective degrees of belief as the primitive objects in a theory of induction. While a Personalistic Bayesian says, "We ought to engage in inductive method x because prior to acquiring any evidence we

---

[1]The idea that induction might depend on one's conceptual framework is not new. For example, J. C. Harsanyi ([13], p. 363) is sympathetic to a dependency on conceptual frameworks for simplicity-favoring in induction. W. C. Salmon [41] argues for a Kuhnian paradigmatic role for prior probabilities. J. Earman ([10], p. 187) devotes a chapter to Kuhnian issues including paradigms. Di Maio ([27], especially pp. 148-149) can be interpreted as arguing for a sort of conceptual framework outlook on inductive logic. DeVito [9] suggests this with respect to the choice of models in curve-fitting. Also, P. Gärdenfors [12] develops a conceptual framework approach to address Goodman's riddle, and he attributes a conceptual framework approach to Quine [35], Carnap [7] and Stalnaker [45].

believe in hypothesis h_1 to degree .2, h_2 to degree .1, etc.," our theory enables us to say things like, "we ought to engage in inductive method x because we acknowledge such and such properties of hypotheses, i.e., because we possess a certain conceptual framework." Both theories can utilize the standard Bayesian asymptotic convergence results to explain long term intersubjective agreement, but our theory helps to explain the intersubjective agreement of our short term inductive behavior: it is because we tend to share the same conceptual framework.

Our theory is not about how (and whether) we are justified in believing that our inductive methods are reliable, or likely to lead us close to the truth; we do not believe there is a solution to this problem. Our theory is also not about how, psychologically, we come to engage in the inductive methods we do. This is an empirical matter for psychologists. Our theory claims to provide reasons for why we *ought* to engage in certain inductive methods. It is in this sense that the Paradigm Theory aims for a solution to the riddle of induction.

The Paradigm Theory can be interpreted as a generalization of Carnap's well known m^*-Logical Theory, and is powerful enough to accommodate other theories including the Principle of Indifference and Hintikka's $\alpha = 0$-Logical Theory. The Paradigm Theory may also be thought of as a quantitative realization of Leibniz's Principle of Sufficient Reason (interpreted non-metaphysically). We apply our theory to a number of standard inductive behaviors in need of justification: Section 5 shows how several varieties of enumerative induction are justified, Section 6 gives a justification for simplicity-favoring, Section 7 takes up curve-fitting, and Section 8 gives two solutions to Bertrand's Paradox. The Paradigm Theory must be introduced before entering these sections aimed at application: Section 2 discusses the hypothesis set choice and parametrization choice, Section 3 develops some of the basic notions underpinning the Paradigm Theory, and Section 4 presents the theory itself.

2   Hypothesis Set and Parametrization

A major criticism of the Classical Theory and its Principle of Indifference (which says that if there is no known reason to prefer one alternative over another, they should receive equal probability) also applies to the Paradigm

Theory: it is that different conclusions are obtained depending on the choice and parametrization of the hypothesis set. For sets of size of the continuum, priors are not invariant under arbitrary reparametrizations, and so the parameter choice can affect the probabilities. Even for finite hypothesis sets the choice of what counts as a hypothesis can affect the probabilities (e.g., why not take (h_1 or h_2) to be primitive instead of h_1 and h_2 separately?). Indeed, generally, with respect to theories of induction with priors there are two difficult issues: (i) the choice and parametrization of the hypothesis set, and once this is answered, (ii) the choice of prior probabilities. For any theory with priors an answer to (i) can affect the answers to (ii). The charge of contradictory applications of the Principle of Indifference extends, therefore, to any Bayesian theory. In fact, nearly *every* theory with probabilities defined over a hypothesis set (or sample space) is subject to this problem.

2.1   Hypothesis Set Choice

For example, the Logical Theory of Carnap is susceptible to the difficulty of hypothesis set choice through its choice of language L, which must come before the choice of a measure function m which is the determiner of the prior probabilities. Even frequency theories have the analogous difficulty in the choice of reference class, which must come before one can possibly infer objective probabilities. Where, then, lies the problem of hypothesis set choice for the Classical Theory? It seems that there is a problem, but it is a problem for most theories of induction, and so is not a problem for the Classical Theory, per se. The meat of our Paradigm Theory, too, does not address hypothesis set choice and parametrization; however, as we will see, we will utilize the Invariance Theory (to be discussed below) as an initial piece of the Paradigm Theory aimed at helping with the choice of parametrization.

Let us mention first some of the things that have been done to address hypothesis set choice. Keynes requires that the hypothesis set be composed of "*indivisible* alternatives of the [same] form..." ([21], p. 60) before the Principle of Indifference may be applied, and one may find criticisms of this in Kneale ([22], pp. 148-149). On the Logical Theory's choice of language Carnap at one point argues for a *requirement of completeness* where the language "must be sufficient for expressing every qualitative attribute of the individuals in the universe of L" ([5], p. 74), and one can find some criticisms of this in Nagel ([30], p. 792) and van Fraassen ([48], pp. 126-129). Frequency

theorist Salmon ([40], p. 91) believes the reference class should be chosen to be the broadest homogeneous one (and Reichenbach [38] apparently prefers the narrowest reference class).

In each of these attempted solutions to the problem of the choice of hypothesis set (or language, or reference class) a choice of, in some sense, the "most general" hypothesis set is urged. At best these suggestions are informal guides to choosing a hypothesis set; they do not provide rigorous rules which uniquely determine hypothesis sets. And there does not seem to be any theory of hypothesis set choice that is any better than these. This does not mean that hypothesis sets are chosen arbitrarily, however—the set of dogs also seems unsusceptible to mathematical rigorization but we know them when we see them. It is very often obvious what the hypothesis set should be even when what the prior should be is far from obvious. The problem of hypothesis set choice nevertheless remains for nearly every theory of induction, and for the Paradigm Theory we just assume that a hypothesis set is given to us.

2.2 Parameter Choice

On the choice of parametrization of the hypothesis set there have been some successes through what we call the Invariance Theory. The problem of multiple parametrizations is that a uniform probability distribution with respect to one parameter is non-uniform with respect to another, and generally, any probability distribution depends on the parametrization. Which parameter is the "correct" one? The Invariance Theory's answer is this: if you acknowledge certain symmetries (i.e., the transformation group that leads to them), then there *might* be a unique probability distribution that respects these symmetries (i.e., that is invariant under the transformations). If there is, you should choose that prior.

Consider the "cube factory" from van Fraassen ([49], pp. 303-10) as an example. Given that every cube made by a factory has a length from 1 to 3 meters, what is the probability that any given cube has a length from 1 to 2 meters? Our answer should not, it seems, depend on whether the question is asked instead in either of the following equivalent fashions: "Given that every cube has an area of side from 1 to 9 square meters, what is the probability that any given cube has area of side from 1 to 4 square meters?" or "Given that every cube has a volume from 1 to 27 cubic meters, what is the probability

that any given cube has a volume from 1 to 8 cubic meters?" Up to a scalar multiple, there is just one prior that gives the same answer no matter the unit of measurement; it is the (unnormalized) prior with probability density function $\mathcal{P}(x) = 1/x$. Under this prior the answers to the three questions are the same.[2]

The Invariance Theory seems to determine the prior while avoiding the problem of the choice of parameter altogether. This is not so. The method actually determines an invariant parametrization, and the determined prior is the uniform one with respect to that parameter. The Invariance Theory, then, can be of help in determining the parameter of the hypothesis set. The difficulty, though, is that the theory seems to determine the prior at the same time. There were supposed to be *two* orthogonal problems: (i) choice of hypothesis set and parameter, and (ii) choice of prior. There are indeed two distinct problems, and there are two aspects of the Invariance Theory that for our purposes may be teased apart, one that answers the parameter part of (i), and the other that answers (ii). It is important for us to separate these aspects of the Invariance Theory, for we are claiming that the Invariance Theory often possesses the power to determine the hypothesis set parametrization, but we want to leave the choice of prior to our Paradigm Theory which allows more versatile answers to (ii).

The first aspect of the Invariance Theory consists of the determination of parameter. This amounts to a determination of measure, which tells us how to quantify the respective sizes of regions of the hypothesis set. *But this measure should not be conflated with a probability measure.* For example, in the cube factory example the Invariance Theory determined the measure of any region $[a, b]$ to be $\log b - \log a$, but this need not be a probability

---

[2]The mathematics is as follows.

$$(\int_1^2 \frac{1}{x} dx)/(\int_1^3 \frac{1}{x} dx) = \frac{\log 2}{\log 3},$$

$$(\int_1^4 \frac{1}{x} dx)/(\int_1^9 \frac{1}{x} dx) = \frac{\log 4}{\log 9} = \frac{\log 2^2}{\log 3^2} = \frac{2 \log 2}{2 \log 3} = \frac{\log 2}{\log 3},$$

and similarly,

$$(\int_1^8 \frac{1}{x} dx)/(\int_1^{27} \frac{1}{x} dx) = \frac{3 \log 2}{3 \log 3} = \frac{\log 2}{\log 3}.$$

measure, and the density function $\mathcal{P}(x) = 1/x$ need not be a probability density function. As it happens, the second aspect of the Invariance Theory is that it *requires* that the invariant measure be the probability measure, but we can have the first aspect without this second; we can have the hypothesis set parametrization without the determination of the prior.

Let the *Underlying Measure Invariance Theory* refer to the theory that is just like the Invariance Theory except that the resulting measure is not interpreted as the probability measure, but instead as the "underlying measure." By "underlying measure" consider, for example, finite sets where the underlying measure is cardinality, which should not be confused with the probability measure; we all agree that for finite sets cardinality is the underlying measure, but there may be no agreement on how the probabilities should be assigned. Also, suppose a cartesian map of the enemy's territory is under consideration and we are attempting to determine those regions that with high probability possess hidden missile silos. We may agree that the appropriate underlying measure is the Euclidean one, but this in no way settles the choice of probability distribution.

The Underlying Measure Invariance Theory (sometimes) determines unique measures (and thus a unique parametrization) from a given transformation group. If we add to this theory a Principle of Indifference of the form, "Unless there is reason to the contrary, equal measures should receive equal probability," then we get a theory that is extensionally equivalent to the Invariance Theory. That is, the Invariance Theory is equivalent to *invariant underlying measure + uniform prior over that underlying measure.* We do not, however, suggest adding the Principle of Indifference to the Underlying Measure Invariance Theory. Instead, it is the Paradigm Theory that will be added to the Underlying Measure Invariance Theory. We see that there exists a principled, if only sometimes successful, approach to the problem of hypothesis set parametrization choice.

2.3   The Paradigm Theory's "First Component"

Our Paradigm Theory's first component is the Underlying Measure Invariance Theory. More specifically, we assume a hypothesis set is given, and for hypothesis sets of the size of the continuum we allow measures to be determined by invariance arguments. Because of the weakness of such arguments—they very often do not determine a unique measure—we allow a parametriza-

tion to simply be given as well, where the underlying measure is the natural one given the parametrization. By this we just mean that the Paradigm Theory is open to the possibility that the parametrization may be determined, if possible, by invariance arguments. The Paradigm Theory proceeds to dictate how the prior is to be determined "on top of" the underlying measure. The Paradigm Theory collapses to the Invariance Theory when paradigms are a certain simple type (called "totally symmetric"); that is, when in certain special cases the Paradigm Theory prescribes assigning a uniform distribution to the underlying measure.

3   Paradigms, Symmetry and Arbitrariness

In order to state and fully appreciate the principles of the Paradigm Theory we present in the next section, it is necessary to say what paradigms are and to bring out the structure that paradigms naturally induce on the hypothesis set.

A guiding intuition in the rational assignment of prior probabilities is the motto that *names should not matter*. This motto is, generally, behind every symmetry argument and motivates two notions formally introduced in this section. 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. We push for the intuitive notion 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 rarer hypotheses. The principles of the Paradigm Theory in the Section 4, founded on different intuitions, respect this notion in that rarer hypotheses receive greater prior probability than less rare hypotheses.

The second concept motivated by the "names should not matter" motto 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. Section 4 describes how the principles of rationality of the Paradigm Theory lead to a prior probability assignment which gives more defensible types of hypotheses greater prior probability.

Before either of these concepts can be discussed, however, we present the formal definition of a conceptual framework, which we label a paradigm.

3.1 Paradigms

We are assuming, without having to explicitly mention it, that hypothesis sets come equipped with a parametrization. Because the Paradigm Theory determines prior *probabilities* over hypothesis sets, to avoid improper priors it is a useful convention to presume that the total "underlying" (as opposed to "probability") measure over any given hypothesis set is finite.

In logic one speaks of the universe and, given some language L, the interpretations of the elements of the language. The universe is just some set, and if the elements of L are k-ary predicate symbols each symbol has as its interpretation some set of k-tuples from the universe. Specifically, if a predicate is unary, then its interpretation is just some subset of the universe; those elements in the set share the property named by the predicate. A universe together with the interpretations of L is called a (first order) *structure* in logic, and one may think of it as a "possible world."[3]

It is from among hypothesis set H that we wish to justify choosing a hypothesis, and, informally, from the point of view of induction H *is* the universe in the formal sense just mentioned. To acquire a "possible world" we just need to specify the interpretations on H. *This is exactly what the formalization of a conceptual framework is in the Paradigm Theory—a "possible world over the hypothesis set"—and the formal notion is called a paradigm.* For simplicity we define paradigms here only for unary predicates. There is nothing, however, preventing a paradigm from possessing relations (or functions and individual constants, for that matter).

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

For example, if H is the set of six outcomes of a roll of a six-sided die, then one possible paradigm is the one that acknowledges being even and being

---

[3]In fact, in defining a structure one need not even state L, for so long as we have the interpretations, we know what the language must be.

odd; i.e., $Q = \{\{1,3,5\},\{2,4,6\}\}$. Or, if H is the set of points in the interior of a unit circle, then one possible paradigm is the one that acknowledges being within distance .5 from the center. For a third example, if H is the set of possible physical probabilities p of a possibly biased coin ($H = [0,1]$), then one possible paradigm is the one that acknowledges the '$p = 0$' and '$p = 1$' hypotheses; i.e., $Q = \{\{0\},\{1\},(0,1],[0,1)\}$. Alternatively, a paradigm can acknowledge all or some of those properties Kuhn puts forth as "...standard criteria for evaluating the adequacy of a theory..." ([23], pp. 321-322): accuracy, consistency, scope, simplicity, and fruitfulness.

A paradigm should be thought of as the set of properties of hypotheses *acknowledged* (by a person or his/her inductive community) to be in the "basic ontology" of hypothesis properties. We do not mean to suggest that one is not able to discriminate, or notice, other properties. It is just that only those properties in the paradigm are sanctioned as "genuine," "metaphysically real" properties.

From the Paradigm Theory's viewpoint it does not matter where the properties come from. We will usually interpret the properties to be subjective, in the conceptual framework sense; the properties may emanate from a person's or scientist's conscious choice as to what properties are to be acknowledged, or they may be those properties he/she is prone through nature or nurture to acknowledge. But the properties *can* be interpreted as objective, and the resulting inductive method determined by the Paradigm Theory is then interpreted as *the* objective inductive method. Objective paradigms can occur in at least two ways. In the statement of an inductive problem it may *say* that certain properties (and only certain properties) exist and are to be acknowledged. Or, certain properties may be "suggested" by the problem statement's *not* mentioning something; for example, in the way that in the Invariance Theory (see Subsection 2.2) the problem statement not mentioning the size of something suggests that the solution must be scale invariant. We will say nothing more about this objective interpretation of paradigms, and we henceforth speak of paradigms in their non-objective Kuhnian paradigmatic and conceptual framework fashion.

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

10

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, even in those cases where the properties in the paradigm are objectively given. Being unbiased, the properties in the paradigm give us no information about the success/truth of any hypothesis, and in this sense the paradigm is not information. Therefore, the Paradigm Theory really is an *a priori* method. Despite paradigms not, in themselves, saying anything about induction, the rationality principles of the Paradigm Theory bring us from paradigms to prior probability distributions.

If one discounts certain hypotheses on the basis of something not measured by the paradigm (e.g., "too complex") or favors some properties over others, then the Paradigm Theory is not applicable. The principles can only be claimed to be rational under these unbiased circumstances. Consider the following example for which the Paradigm Theory does not apply. 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 seems to be $\{\{1, 2, 3\}, \{4\}\}$, and the Paradigm Theory will say that 4 should be preferred. But side 4 should definitely *not* be preferred. The 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). However, if one has no experience with die or similar sorts of objects and acknowledges these properties, then although one will, in our theory, initially acquire degrees of belief favoring 4, one will eventually come to modify one's beliefs and correct for this mistake.

It may seem to be a great weakness of a theory of induction that it is in principle incapable of application to all those real world cases where we possess background knowledge that favor some hypotheses over others. For example, we have background knowledge that side 4 should be disfavored when rolling the four-sided die above, but the Paradigm Theory has no insight on how to assign priors disfavoring side 4 given this knowledge. There are two points that should be made with respect to this. First, as far as we know, there are no theories of induction that determine exact priors in such cases, and we suspect that it is not susceptible to mathematical rigoriza-

tion. Second, many theories, including the Paradigm Theory, explain why we disfavor side 4 by assuming that at some point in the past we did not disfavor side 4 (and perhaps, as in the Paradigm Theory, even favored side 4), and that conditionalizing ever since on the evidence that side 4 comes up less frequently, side 4 has acquired less probability. About such theories, Earman calls them "never-never-land Bayesianism, where an agent begins as a *tabula rasa*, chooses her priors, and forever after changes her probabilities only by conditionalization" ([10], pp. 139-140). Although "never-never-land" is meant derogatorily, we believe that a theory of the justification induction ought to be just that—a never-never-land theory—and the Paradigm Theory certainly falls under this heading.

In reality we do not possess just one conceptual framework. Rather, different conceptual frameworks become salient depending on the context. One might say that our "total" conceptual framework consists of lots of "little" conceptual frameworks, or paradigms, and which "little" conceptual framework is active at any one time depends on the inductive scenario before us. For example, when presented with the set of physical probabilities p of a biased experiment, the "active" conceptual framework in one context may be the one that acknowledges, say, the property of being a law, but in another context the active conceptual framework may be the one that acknowledges nothing at all. And when presented with the set of polynomials instead of the set of physical probabilities, the active conceptual framework may be the one that acknowledges, say, the properties of being a constant, line, parabola, etc.

3.2   Being Symmetric

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 us as undergraduates, 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.

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 re-

specting the underlying measure such that the paradigm is unchanged but the name for h_1 becomes the name for h_2. Formally, for $p : H \to H$, 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 $p : H \to H$ such that $p(Q) = Q$ and $p(h_1) = h_2$. $\square$

In the definition of 'Q-symmetric' each measure-preserving bijection $p : H \to H$ 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 $p_1 : (a, b, c) \to (c, b, a)$ preserves Q, but the renaming $p_2 : (a, b, c) \to (c, a, b)$ 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. We

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 we 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 the Paradigm Theory.

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 $Q = \{\{p(v)\}|v \in V\} \cup \{\{p(v_1), p(v_2)\} |(v_1, v_2) \in E\}\}$. 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.
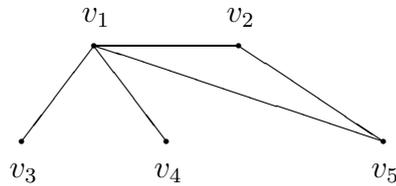
The associated hypothesis set is $H_b = \{v_1, \ldots, v_5\}$ and the associated paradigm is $Q_b = \{\{v_1\}, \ldots, \{v_5\}\} \cup \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_1, v_4\}, \{v_1, v_5\}, \{v_2, v_5\}\}$. Notice that $\{v_1\}$, $\{v_2, v_5\}$, and $\{v_3, v_4\}$ are the Q_b-symmetry types; so only 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.

3.3 Defensibility Hierarchy and Sufficient 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 can *now* be picked for a reason?

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*. □

For instance, for H_d and Q_d above we had:

- $H_d^0 = \{h_0, h_1, h_2, h_3, h_4\}$, $Q_d^0 = \{\{h_0, h_2\}, \{h_1, h_2\}, \{h_3\}, \{h_2, h_3, h_4\}\}$.

- $H_d^1 = \{h_2, h_3, h_4\}$, $Q_d^1 = \{\{h_2\}, \{h_3\}, \{h_2, h_3, h_4\}\}$.

- $H_d^2 = \{h_4\}$, $Q_d^2 = \{\{h_4\}\}$.

- $H_d^3 = \{h_4\}$, $Q_d^3 = \{\{h_4\}\}$.

- 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.[4] 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 mth. 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.

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*. The Paradigm Theory's

---

[4]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).

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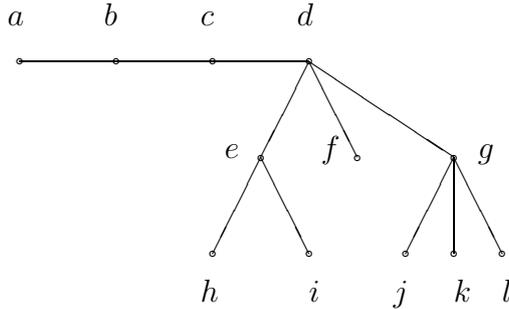
The symmetry types are $\{h, i\}$, $\{j, k, l\}$ and every other vertex is in a singleton symmetry type. The defensibility types are $\{h, i\}$, $\{j, k, l\}$, $\{e, f, g\}$, $\{a, d\}$ and $\{b, c\}$. The defensibility levels are $\Delta^0 = \{h, i, j, k, l\}$, $\Delta^1 = \{e, f, g\}$, and $\Delta^2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

We noted in Subsection 3.2 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,[5] 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, [2], 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 section'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-

---

[5]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, [2], p. 346).

metaphysically only), giving a more finely grained breakdown of "how suffi-cient" 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. The Paradigm Theory, further, quantifies the degree of sufficiency of reason with real numbers in [0,1], as we will soon see.

4 The Paradigm Theory

Section 3 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 in-variant (or "unique") hypotheses—there may be some hypotheses that are "more" invariant, and among *these* there may some that are "even more" in-variant, etc. This led to the defensibility, or invariance, hierarchy. Hypothe-ses that "become symmetric" at some level of the hierarchy are equivalent, and are said to be members of the same equivalence type.

We noted in Section 3 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, rarer hypotheses are to be preferred as it takes less arbitrariness to choose them, (ii) (equivalence) 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 hy-potheses are, *a priori,* more preferred.

These intuitions follow from the three rationality principles concerning prior probabilities we are 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 the 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).

Subsection 4.1 presents the three principles of rationality, Subsection 4.2 gives some preliminary applications, Subsection 4.3 discusses the use of "secondary paradigms" to acquire more detailed prior probability distributions, and Subsection 4.4 sets forth the sort of explanations the Paradigm Theory gives.

4.1   The Principles

The 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. The Paradigm Theory as developed in this paper is only capable of handling cases where there are finitely many symmetry types.[6] We will assume from here on that paradigms induce just finitely many symmetry types.[7]

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

_____

[6]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 ([46], 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.

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

4.1.1 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 the 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 ([19], 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 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 the Paradigm Theory. The 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 our 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.

4.1.2 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, 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 in Subsection 5.2.1), then the Principle of Symmetry just *is* Johnson's Permutability Postulate ([19], pp. 178-189), perhaps more famously known as de Finetti's Finite Exchangeability.

4.1.3 The Basic Theory

Carnap's m^*-based Logical Theory ([5], p. 563) uses versions of the Principles of Type Uniformity and Symmetry (and leads to the inductive method he calls c^*). The 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 the 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), the 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 $c^\dagger$).

For any particular Logical Theory of Carnap's, once a language L is provided, the Logical Theory determines an inductive method. Although this is literally true, it glosses over the fact that there are a multiplicity of logical theories. $c^\dagger$ and c^* are two of them, and Carnap [6] cannot decide a best inductive method from a continuum of inductive methods between these two confirmation functions, parametrized by the familiar λ. The problem is that deciding between these confirmation functions and their corresponding underlying measure functions is just as vexing as is the original problem of determining prior probabilities. Coming to its aid, the Paradigm Theory can be used to determine a unique measure function for the Logical Theory (and thus a unique confirmation function, or inductive method): let the hypothesis set be the set of state-descriptions and let the paradigm consist of those properties of state-descriptions one acknowledges. But the Paradigm Theory is now doing all the work for the Logical Theory, and there is no need to have

the Logical Theory in the first place. In fact, only the first two principles of the Paradigm Theory are needed to accomplish this task.

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 Section 3, 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 section, 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 $h \in H$, $P(h) = \frac{1}{w \cdot c(h)}$.* $\square$

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.* □

The Basic Theory is not the 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 we say things like, "only the Basic Theory is needed to determine such and such probabilities," we mean that the probabilities are not changed upon the application of the third principle (to be stated below) of the Paradigm Theory.

4.1.4   Principle of Defensibility

The third principle of rationality is, as far as we know, not similar to any previous principle in the induction and probability literature. It encapsulates an intuition similar to that used when we developed gradations of invariance in Subsection 3.3. We asked: Among the invariant elements, which are more defensible? Now we 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. We 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.

4.1.5 The Paradigm Theory

With the three principles presented we can state the 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., $c(h, e, Q)$.

The Paradigm Theory is superior to the Basic Theory in the sense that it is able to distinguish higher degrees of defensibility. The Paradigm Theory on H_a/Q_a and H_b/Q_b from Section 3 behaves identically to the Basic Theory. Applying the 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; the 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, the 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 the

Paradigm Theory). Theorem 14 in the Appendix examines some of its properties. Unlike the Basic Theory, the 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, the 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.

Theorem 3 *The following are true about the 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 $h' \neq h$, $P(h') < P(h)$.* $\square$

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

4.2   Simple Applications

By way of example we apply the Basic and Paradigm Theories to some preliminary applications.

4.2.1   Collapsing to the Principle of Indifference

The Paradigm Theory (and the Basic Theory) gives the uniform distribution when the paradigm is empty. This is important because, in other words, the 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 *the 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

28

Symmetry—and not the Principle of Type Uniformity—needs to be used to found the Principle of Indifference as a subcase of the Paradigm Theory.

4.2.2 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 the 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. The 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 we suspect no natural paradigm favors any other.

4.2.3 Leibniz's Triangle

To a second historical example, we noted in Subsection 3.3 the connection of the Paradigm Theory to Leibniz's Principle of Sufficient Reason (interpreted non-metaphysically), and we stated that the 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 the 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, [2], p. 101.)

Here the hypothesis set is plausibly $\{\langle \theta_1, \theta_2, \theta_3 \rangle \mid \theta_1 + \theta_2 + \theta_3 = 180°\}$, 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., $Q = \{\{\langle \theta_1, \theta_2, \theta_3 \rangle, \langle \theta_3, \theta_1, \theta_2 \rangle, \langle \theta_2, \theta_3, \theta_1 \rangle, \langle \theta_3, \theta_2, \theta_1 \rangle, \langle \theta_1, \theta_3, \theta_2 \rangle, \langle \theta_2, \theta_1, \theta_3 \rangle\} \mid \theta_1 + \theta_2 + \theta_3 = 180°\}$. This natural paradigm, regardless of the hypothesis set's underlying measure, results in $\langle 60°, 60°, 60° \rangle$ 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.

4.2.4 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, we 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 the 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.

4.2.5 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, 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.[8] 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.[9] 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. We do not know of a proof that there is no other invariant 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.

The 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.*[10]

4.3 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

---

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

[9]This is a sort of "complexification;" see Section 6.

[10]See H. Putnam ([33], [34], p. 33) for some discussion on underdetermination of interpretation and its effect on theories of reference, and D. Lewis [25] for some commentary and criticism on it.

according to this secondary paradigm? After all, cannot any hypothesis set and paradigm be brought to the 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 the Paradigm Theory.

Here is an example of the sort of thing we 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.

We generalize the Paradigm Theory to allow such secondary paradigms in a moment, but we would first like to further motivate it. There is a sense in which the 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 the Paradigm Theory to break this limit. The 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 the 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 the Paradigm Theory is only as natural and explanatory as is the paradigm (primary, secondary, and so on) used (see Section 4.4). Ad hoc secondary paradigms create ad hoc explanations. Our only use of paradigms beyond primary ones are secondary ones. We use them in Subsubsection 5.3.1 where they are quite explanatory and give the Paradigm Theory the ability to generalize a certain Logical Theory of Hintikka's ($\alpha = 0$). We also note in Chapter 6 their ability to give a non-uniform prior over the simplest hypotheses. If in any particular application of the Paradigm Theory there is no mention of secondary paradigms, then they are presumed not to exist.

4.4 The Paradigm Theory Tactic

In the following sections we use the Paradigm Theory 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 the Paradigm Theory is given (often presented as a theorem). Second, an informal explanatory argument is given. The 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 the Paradigm Theory; and this is a purely mathematical question. Such a demonstration is not enough to count as an explanation 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. We 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 the 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 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."

Our 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.

5 Enumerative Induction

We 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 we 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 1 shows the prior probability assignments for the paradigms used in this section on the hypothesis set H_4.

5.1 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.

5.1.1 $H = H_N$

First we consider no-induction on the hypothesis set H_N, the set of outcome strings for N binary experiments. Table 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 ([50], 5.15-5.154), and it is essentially Carnap's $m^\dagger$ [5] (or $\lambda = \infty$).

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

Table 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 as in Subsection 5.3.1. 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.

Recall that a totally symmetric paradigm is one in which every pair of hypotheses is symmetric. Any totally symmetric paradigm entails the uniform 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. We have let Q_s denote a generic totally symmetric paradigm in Table 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. We 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 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.

5.1.2 $H = [0, 1]$

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 the 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.

5.2   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 [51] for a variety of historical proofs of the rule. In this section we demonstrate how the Paradigm Theory naturally leads to the Rule of Succession when $H = H_N$ and $H = [0, 1]$.

5.2.1 $H = H_N$

The second column of probabilities in Table 1, headed "Q_L," shows the probability assignment on H_4 needed to lead to Laplace's Rule of Succession.[11] 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 ([19], 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

---

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

not matter—distributes the probability uniformly within each complexion. Carnap's Logical Theory with m^* ([5], 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 Subsubsection 4.1.3).

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

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 the Paradigm Theory is identical to that of Johnson, and so Q_L results in Laplace's Rule of Succession.* $\square$

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 the 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,

41

the probability of the next outcome being a 1 is $\frac{r+\lambda/2}{n+\lambda}$. We 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, or $\lambda = 0$, does not, therefore, seem to be justifiable within the 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 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 the 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.* $\square$

The proof is similar to that of Theorem 4 and is omitted. 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" [24]. 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.

5.2.2 $H = [0, 1]$

Now we consider frequency-induction on the hypothesis set $H = [0, 1]$ with the natural uniform underlying measure. We noted in Subsubsection 5.1.2 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 the 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 the 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.* □

If one acknowleges nothing on $[0, 1]$, or more generally one makes no distinctions, the Paradigm Theory collapses to a sort of Principle of Indifference (see Subsection 4.2.1) 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.

5.3 Law-Induction

Frequency-induction allows instance confirmation, the ability to place a probability on the outcome of the very next experiment. C. D. Broad [4] 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 ([5], 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.

H. Jeffreys [18] 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

44

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? The Paradigm Theory determines exact inductive methods capable of giving high posterior probability to laws, and it does so with very natural paradigms.

5.3.1 $H = H_N$

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. The 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 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. We record this in the following theorem whose proof may be found in the Appendix.

Theorem 7 *If there will be N experiments and $1 \le n < N$ 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*

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

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 [14], where a constituent is one possible way the world can be in the following sense: either all things are 0, some things are

45

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 Subsubsection 5.1.1.

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 Q_L from Subsubsection 5.2.1, 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' or '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 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). □

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 [14]. 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 the Paradigm Theory to get *exactly* Hintikka's $\alpha = 0$ assignment, but the only paradigms we 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',σ} 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 we do not believe the 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 the 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 the Paradigm Theory can naturally get to $\alpha = 0$ is with Q_{law_L} (or Q_{const_L}). Ignoring this, the 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. The 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.

5.3.2 $H = [0, 1]$

We have seen in Subsection 5.2 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 den-

sity 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, 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* $\frac{r+1}{n+2}$. $\square$

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.

6 Simplicity-Favoring

Occam's Razor says that one should not postulate unnecessary entities, and this is roughly the sort of simplicity to which we refer (although any notion of simplicity that has the same formal structure as that described below does as well). The 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. The paradigm need only *acknowledge* which hypotheses are simpler than which others.[12] In a sentence, the 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.[13]

Recall from Subsection 3.2 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. 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.* □

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.[14] Informally,

---

[12]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 paper do not allow relations.

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

[14]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

"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).

6.1 Full Complexification Trees

Let Q_{full} be the paradigm represented by the full tree below.
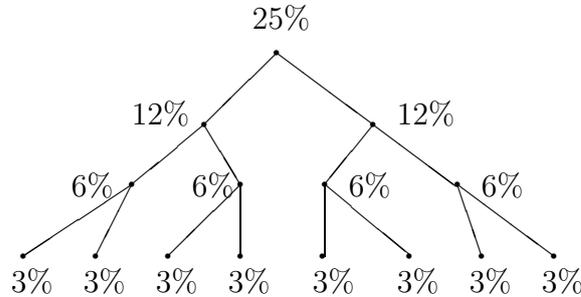
There are four symmetry types (one for each level), so each receives probability $1/4$. The approximate probability for each hypothesis is shown in the figure. Only the Basic Theory is needed here—i.e., the Principles of Type Uniformity and Symmetry—the Principle of Defensibility does not apply. If there 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 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 $P(h) = \frac{1}{m(n+1)b^i}$.* $\square$

---

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.
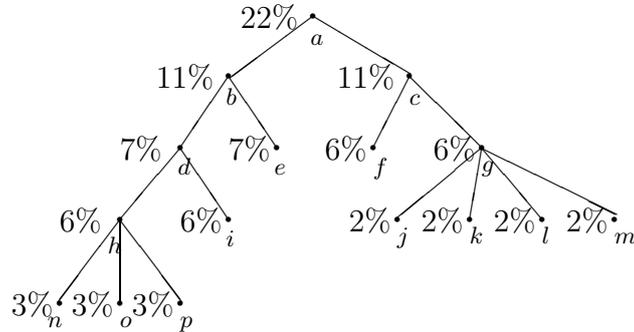
This tells us that the prior probability of a hypothesis drops exponentially the more one complexifies it, i.e., the greater i becomes. For example, consider base 10 numbers as in the hypothesis set $H = \{2; 2.0, \ldots, 2.9; 2.00, \ldots, 2.09; \ldots; 2.90, \ldots 2.99\}$, and suppose the paradigm 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.

6.2 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 the 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.

22% •
a

11% •   11% •
b   c

7% •   7% •   6% •   6% •
d   e   f   g

6% •   6% •   2% • 2% • 2% • 2% •
h   i   j   k   l   m

3% • 3% • 3% •
n   o   p

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 $\Delta^4 = \{a\}$. The 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 (see Subsubsection 4.1.3). 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
4.3, 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 the 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.

6.3 Discussion

We are 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 us address what could be a criticism of our explanation of the justification of simplicity-favoring. Our explanation depends on the resulting graph associated with the paradigm being a tree, with the simpler hypotheses near the root. This occurs because, so we 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. We 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, the 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, we 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 we have given in this section for favoring simpler hypotheses compared to those of other theorists (not all who are aiming at solving the same problem of induction as

us): (i) they are more susceptible to falsification (Popper, [32]), (ii) they are more susceptible to confirmation (Quine, [36]), (iii) they are practically easier to apply (Russell, [39]; Pearson, [31]; Mach, [26]), (iv) they have greater *a priori* likelihood of being true (Jeffreys, [17]), (v) they have been found in the past to be more successful (Reichenbach, [37]), (vi) following the rule 'pick the simplest hypothesis' leads with high probability to true hypotheses (Kemeny, [20]), (vii) they are more informative (Sober, [44]), (viii) they are more stable (Turney, [47]), and (ix) they have higher estimated predictive accuracy (Forster and Sober, [11]). Our 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.

7 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,[15] is typically favored over curves that require $n + 1$ parameters, even when the latter fits the data better than the former. We derive within the 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*. We concentrate only on the hypothesis set of polynomials, and consider only those up to some finite degree. For definiteness we 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

---

[15]Do not confuse this notion of model with that discussed in Subsubsection 4.2.5. There is no relation.

same predictions as M_0. In cases such as this Jeffreys ([17], see also [15], 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). The Paradigm Theory gives a principled defense for Jeffreys' prior probability assignment: if the conceptual framework acknowledges the 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.[16]

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

---

[16]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.

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 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 $\frac{1}{(l+1)sm^k}$.* $\square$

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 $\frac{1}{(l+1)\lfloor sm^k \rfloor}$.

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 kth dimension M_k, and γ depends on the specific information criterion.[17] Once this model is determined, the maximum likely hypothesis

---

[17]See Smith and Spiegelhalter ([43], pp. 218) for many of the information criteria (our

in it is chosen, even though it may well not be the maximum likely hypothesis in the entire hypothesis set. The 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.

We set ourselves the task of finding 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, we 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 is the data). 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

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

γ is their $m/2$) and references to the original papers defending them; see also Aitkin [1].

and then choose the maximum likely hypothesis in that model. We have just proven the following theorem, which we 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 the 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.* $\square$

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 ([43], 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.[18] When $m = e$, the resulting information criterion is precisely Akaike's Information Criterion. This amounts to a sort of answer to Forster and Sobers' ([11], 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 $\gamma = 1$. The

---

[18]Our γ is Smith and Spiegelhalters' $m/2$. Their m is not the same as ours.

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, [43]), where $\gamma = (\log n)/2$ and n is the number of data (Schwarz, [42]).)

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 polynomials 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, [43]). For these "non-extreme" choices of m, curves requiring fewer parameters quickly acquire maximal posterior probability so long as their fit is moderately good.

The 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.

The 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 the 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 the 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 the Paradigm Theory.

8 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 ([3], pp. 4-5). The Classical Theory's use of 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 ([22], 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 ([29], pp. 136-146). Jaynes [16] presents a solution which we discuss more in Subsection 8.1 below. Marinoff [28] catalogues a variety of solutions in a recent article. We approach Bertrand's Paradox in two fashions.

8.1 Generalized Invariance Theory

In our first Paradigm Theory treatment of Bertrand's Paradox we 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, we call it a *prior set*; we denote the elements of this set by ρ rather than h, and denote the set H_ρ.

We wish to determine a prior probability assignment on H_ρ. What "should" the paradigm be? Jaynes [16] 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

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

where R is the radius of the circle. The theory sanctioning this sort of determination of priors we call the *Invariance Theory* (see Subsection 2.2).

We 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, the 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 $T \cap R \cap S = \{\frac{1}{2\pi R r}\}$ and the other symmetry types are infinite, $P(\frac{1}{2\pi R r}) = 1/5$ and every other prior receives negligable prior probability; $1/(2\pi R r)$ 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.[19]

This "trick" of using the Paradigm Theory parasitically on the Invariance Theory can be employed nearly whenever the latter theory determines

---

[19] 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 ([28], pp. 7-8).

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 we 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 the 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 the Paradigm Theory's definition as well. Given that this is so, by Theorem 2 this prior has the greatest prior probability.

The 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 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 we mean to convey is that *the 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 the Paradigm Theory brings more refinement to the Invariance Theory. In the cases where the 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 the Paradigm Theory *if* the invariance proper-

ties are just *acknowledged* and not favored. Also, not only can the 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, the Paradigm Theory allows not only a refinement, but a sort of generalization of the Invariance Theory.

8.2 H is the Sample Space

The second way of naturally approaching Bertrand's Paradox within the 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 section) 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. We choose H_1 as, in our 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

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

where A_i is the area of the ith 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

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

64

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.

9 Conclusion

The intuition underlying the Paradigm Theory is that *rarer 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 rare), which means that there is less "sufficient reason" to choose it. The principles of the Paradigm Theory link with this intuition. The Principle of Type Uniformity and Principle of Symmetry give rarer hypotheses greater prior probability, and the Principle of Defensibility entails that among the rarer hypotheses, those that are rarer should receive greater prior probability. Recall (from Section 4) that these are the *links* of the principles to the "rarer is better" motto—the principles do not actually *say* anything about the rareness of hypotheses, but are motivated for completely different, compelling reasons of their own. Nevertheless, it is a convenient one-liner to say that the Paradigm Theory favors rarer 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 (Subsection 3.3), interpreted nonmetaphysically only.

The favoring of rarer 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 (Subsubsection 4.1.3 and Subsubsection 5.3.1). Besides these achievements, the Paradigm Theory gives explanations for a large variety of inductive phenomena:

- it "correctly" collapses to the Classical Theory's Principle of Indiffer-

ence when no distinctions are made among the hypotheses (Subsubsection 4.2.1),

- it suggests a conceptual framework-based solution to the problem of the under-determination of interpretation for language (Subsubsection 4.2.5),

- it explains why no-inductions are rarely considered rational (Subsection 5.1),

- it explains why frequency-inductions and law-inductions *are* usually considered rational (Subsections 5.2 and 5.3),

- it gives a foundation for Occam's Razor by putting forth the notion that simpler hypotheses are favored because one acknowledges simplicity and, given this, they are less arbitrary (Section 6),

- it accommodates curve-fitting by supposing only that one acknowledges the usual models—constants, lines, parabolas, etc. (Section 7),

- 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 (Subsection 8.1),

- and it accounts for Bertrand's Paradox in another fashion by acknowledging the distance from the center of the circle (Subsection 8.2).

We have no doubt that the Paradigm Theory will find explanatory application to other important inductive phenomena as well.

The Paradigm Theory scores at two levels. The first is illustrated by the gamut of particular explanations like those just enumerated. One should not lose sight of *the point* (i.e., the forest through the trees) of the Paradigm Theory, which is to be a theory of the justification of induction. More accurately, it is meant to be a single theory through which the justification of all inductive conclusions—our *a posteriori* beliefs—may be explained. How, then, *do* we acquire justified *a posteriori* beliefs? The Paradigm Theory of Induction requires one primitive sort of *a priori* object in order to answer the question: the conceptual framework. Once this is in hand, inductive methods are logically determined via the principles of the Paradigm Theory, and

these inductive methods describe the way in which our *a posteriori* beliefs develop in the light of evidence. With conceptual frameworks as the backbone of the Paradigm Theory instead of, say, the subjective degrees of belief of Personalistic Bayesianism, we have been able to explain why our short-term inductive methods, and even our *a priori* probability assignments, achieve so much intersubjective agreement: The high degree of agreement in our inductive methods and conclusions is explained not (only) by the standard in-the-limit results to which any Bayesian may appeal, but by the supposition that we all tend to acknowledge the same properties of hypotheses—i.e., by the supposition that we often possess the same "metaphysics" with respect to hypothesis properties.

10   Appendix

$\delta(h) = \gamma$ (*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 the Paradigm Theory.

Theorem 14 *The following are true concerning the 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. $\square$

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. $\square$

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 the 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. The reader may verify that h and i receive probability $\frac{1}{18}$, but e, f, and g receive probability $\frac{14}{15}\frac{1}{18} < \frac{1}{18}$. $\square$

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 asymetrical, none of the complexions are symmetrical with any others. Thus, each complexion is a symmetry type, and each complexion receives a probability of $1/(2n - 1)$. $\square$

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

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

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 $\frac{n+1}{n+3}$. $\square$

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 6 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.

$\square$
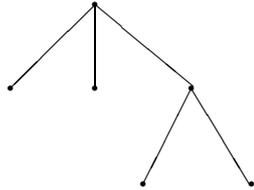
Proof of Theorem 11. There are $n+1$ symmetry types (one for each level), each receiving probability $1/(n+1)$. The symmetry type at depth i has b^i elements. $\square$

References

[1] M. Aitkin. Posterior Bayes factors. *Journal of the Royal Statistical Society, B*, (1):111–142, 1991.

[2] R. Ariew and D. Garber. *G. W. Leibniz. Philosophical Essays*. Hackett Publishing Company, 1989.

[3] J. Bertrand. *Calcul des probabilites*. Gauthier-Villars, 1889.

[4] C. D. Broad. On the relation between induction and probability. *Mind*, 108:389–404, 1918.

[5] R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, 1950.

[6] R. Carnap. *The Continuum of Inductive Methods*. The University of Chicago Press, Chicago, Illinois, 1952.

[7] R. Carnap. A basic system of inductive logic, part II. In R. Carnap and R. C. Jeffrey, editors, *Studies in Inductive Logic and Probability*, pages 7–156. University of California Press, 1971.

[8] R. Carnap. Statistical and inductive probability. In *Readings in the Philosophy of Science*, pages 279–287. Prentice Hall, 1971.

[9] S. DeVito. A gruesome problem. *British Journal for the Philosophy of Science*, 48:391–396, 1997.

[10] J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. The MIT Press, 1992.

[11] M. R. Forster and E. Sober. How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, 45:1–35, 1994.

[12] P. Gärdenfors. Induction, conceptual spaces, and AI. *Philosophy of Science*, 57:78–95, 1990.

[13] J. C. Harsanyi. Bayesian decision theory. Subjective and objective probabilities, and acceptance of empirical hypotheses. *Synthese*, 57:341–365, 1983.

[14] J. Hintikka. A two-dimensional continuum of inductive methods. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 113–132. North-Holland, 1966.

[15] C. Howson. Popper, prior probabilities, and inductive inference. *The British Journal for the Philosophy of Science*, 38:207–224, 1987.

[16] E. T. Jaynes. The well-posed problem. *Foundations of Physics*, pages 477–493, 1973.

[17] H. Jeffreys. *Theory of Probability*. Clarendon Press, 1948.

[18] H. Jeffreys. The present position in probability theory. *The British Journal for the Philosophy of Science*, 5(20):275–289, 1955.

[19] W. E. Johnson. *Logic, Part III: The Logical Foundations of Science*. Cambridge University Press, 1924.

[20] J. G. Kemeny. The use of simplicity in induction. *The Philosophical Review*, 62:391–408, 1953.

[21] J. M. Keynes. *A Treatise on Probability*. Macmillan, 1921.

[22] W. Kneale. *Probability and Induction*. Clarendon Press, 1949.

[23] T. S. Kuhn. Objectivity, value judgments, and theory choice. In T. S. Kuhn, editor, *The Essential Tension*. University of Chicago Press, 1977.

[24] P. S. Laplace. Essai philosophique sur les probabilit'es (1820). English translation: *Philosophical Essays on Probability*.

[25] D. Lewis. Putnam's paradox. *Australasian Journal of Philosophy*, 62(3):221–236, 1984.

[26] E. Mach. *La Connaissance et l'Erreur*.

[27] M. C. Di Maio. Inductive logic: Aims and procedures. *Theoria*, 60(2):129–153, 1994.

[28] L. Marinoff. A resolution of Bertrand's Paradox. *Philosophy of Science*, 61:1–24, 1994.

[29] D. H. Mellor. *The Matter of Chance*. University Press, 1971.

[30] E. Nagel. Carnap's theory of induction. In P. A. Schilpp, editor, *The Philosophy of Rudolf Carnap*. Open Court Press, 1963.

[31] K. Pearson. *The Grammar of Science*. J. M. Dent and Sons, 1892.

[32] K. R. Popper. *The Logic of Scientific Discovery*. Hutchinson and Company, 1959.

[33] H. Putnam. Models and reality. *Journal of Symbolic Logic*, 45:464–482, 1980.

[34] H. Putnam. *Reason, Truth and History*. Cambridge University Press, 1981.

[35] W. V. O. Quine. *Word and Object*. MIT Press, 1960.

[36] W. V. O. Quine. On simple theories in a complex world. *Synthese*, 15(1):103–106, 1963.

[37] H. Reichenbach. *Experience and Prediction.* University of Chicago Press, 1938.

[38] H. Reichenbach. *The Theory of Probability.* University of California Press, 1949.

[39] B. Russell. *Mysticism and Logic.* Longmans; now Allen and Unwin, 1918.

[40] W. C. Salmon. *The Foundations of Scientific Inference.* University of Pittsburgh Press, 1966.

[41] W. C. Salmon. Rationality and objectivity in science, *or* Tom Kuhn meets Tom Bayes. In C. W. Savage, editor, *Scientific Theories*, pages 175–204. University of Minnesota Press, 1990.

[42] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[43] A. F. M. Smith and D. J. Spiegelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, B*, (2):213–220, 1980.

[44] E. Sober. *Simplicity.* Oxford University Press, 1975.

[45] R. Stalnaker. Anti-essentialism. *Midwest Studies in Philosophy*, IV:343–355, 1979.

[46] F. Suppe. *The Semantic Conception of Theories and Scientific Realism.* 1989.

[47] P. Turney. The curve fitting problem: A solution. *British Journal for the Philosophy of Science*, 41:509–530, 1990.

[48] B. C. van Fraassen. Probabilities and the problem of individuation. In S. A. Luckenbach, editor, *Probabilities, Problems, and Paradoxes*, pages 121–138. Dickenson, 1972.

[49] B. C. van Fraassen. *Laws and Symmetry*. Clarendon Press, 1989.

[50] L. Wittgenstein. *Tractatus Logico-philosophicus*. Routledge and Kegan Paul, 1961.

[51] S. L. Zabell. The rule of succession. *Erkenntnis*, 31:283–321, 1989.