A Game Theoretic Argument For Ockham's Razor

Conor Mayo-Wilson

September 8, 2009

Contents

Acknowledgments

As is obvious from the number of citations to his work and the style of the central argument in what follows, I owe an enormous intellectual debt to Kevin Kelly. Kevin spent endless hours discussing the concept of simplicity with me; he provided extensive written criticisms on earlier drafts of this thesis, and perhaps most importantly, he gave me personal encouragement at every step of the way. I cannot thank Kevin enough for the immense time and energy that he has devoted to making this project come to fruition.

David Danks has also been an enormous asset to me in writing this thesis. His attention to detail never ceases to amaze me, and the quickness and thoroughness with which he provided comments on earlier drafts of this thesis were absolutely essential in the evolution of my writing and arguments. To put it bluntly, much of what is written below would be lacking in motivation, organization, and cogency were it not for David's comments and suggestive questions on earlier drafts. Moreover, David pushed me to provide informal explanations of technical results so as to sharpen the philosophical interpretations of the theorems in the later chapters.

I also owe a great deal of thanks to a third member of the Carnegie Mellon Philosophy department who was not an official member of my committee: Teddy Seidenfeld. Teddy likely agrees with very few of the philosophical conclusions below, but were it not for his questions and constant probing, I would not have written this thesis at all. His influence on my thinking should, I hope, be apparent in the discussions of decision theory and game theory throughout the thesis. I owe thanks, therefore, for the time he has set aside to discuss any number of issues concerning foundations of game theory, decision theory, and finite-additivity, among other topics.

Finally, my family and friends have been immensely supportive during the time in which I wrote this thesis. Few of them actually care about the subject matter of this paper, but they cared deeply about my (sometimes questionable) emotional and physical well-being as I wrote it. I can't mention everyone, but special thanks are due to my parents, my brother, my roommates, Shawn and Lena, Stephanie, Steve, and Ruth.

Introduction

Beginning in Ancient Greece with Aristotle and Ptolemy,[1] continuing in the Renaissance and Enlightenment in the works of Copernicus, Galileo, and Newton,[2] and persisting in the 20th century in the writings of Feynmann and Einstein,[3] scientists have consistently appealed to the principle that, all other things being equal, it is rational to prefer simpler scientific theories to more complex ones. This principle is called *Ockham's razor,* and the aim of this thesis is justify it.

Justifying Ockham's razor requires answering at least three questions. First, what makes a scientific theory simple? Second, what criteria make it "rational to prefer" one scientific theory to another? That is, what are the costs of endorsing, believing, or pursuing research in a particular scientific theory? Finally, why does a systematic preference for simpler theories minimize such costs?[4] In

[1] In the *Posterior Analytics,* Aristotle writes "We may assume the superiority ceteris paribus of the demonstration which derives from fewer postulates or hypotheses." See Aristotle (1971). Similarly, in arguing that a geocentric model of the solar system is most plausible, Ptolemy claims, "For the same [observations] would result as if [the earth] had another position than at the center [of the solar system]. And so it seems to me *superfluous* to look for the causes of the motion to the center when it is once for all clear from the very appearances that the earth is in the middle of the world" (my italics). See Ptolemy (1958) pp. 12.

[2] In defending a heliocentric model of the solar system, Copernicus writes, "Just as [nature] especially avoids producing anything superfluous or useless, so it prefers to endow a single thing with many effects." See Copernicus (1995). Galileo uses similar reasoning in defending Copernicus: "Nature does not multiply things unnecessarily; that she makes use of the easiest and simplest means for producing her effects; that she does nothing in vain, and the like." See Galileo (1953). Finally, Newton's first rule of scientific explanation reads as follows: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances." Newton (1964) pp. 398.

[3] Einstein writes, "The grand aim of all science is to cover the greatest possible number of empirical facts by logical deductions from the smallest possible number of hypotheses or axioms." Quoted in Nash (1963) pp. 173.

[4] To these three questions, Alan Baker adds a fourth: what role does simplicity *actually* play in scientific practice? See Baker (2004). Baker's distinction is important: many philosophers have defended Ockham's razor, but they often use the word "simple" in a way that is different from that employed by practicing scientists. Thus, some philosophical defenses of Ockham's razor explain how a preference for simpler scientific theories *might be* justified, but they also provide no defense of why various scientists' appeals to simplicity have been, and continue to be, successful heuristics in inquiry. I provide a defense of Ockham's razor that, I will argue, justifies appeals to simplicity in several scientific problems, including curve-fitting, causal inference, and estimating conserved quantities in particle physics. I do not, however, provide an extensive historical or sociological study to show that past or present scientists understand simplicity and Ockham's razor in the way defended in this paper.

this introduction, I provide an overview of how philosophers, scientists, and statisticians have attempted to answer these three questions. This overview is expanded in the second chapter, where I summarize the philosophical literature dedicated to Ockham's razor in far greater detail.

How should simplicity be defined? The list of proposed definitions is numerous; scientific theories have been characterized as simple if they minimize any of the following: (1) number of causes required to explain a set of phenomena,[5] (2) number of theoretical entities (e.g. fundamental particles, chemical elements, etc.),[6] (3) number of theoretical predicates,[7] (4) number of free parameters,[8] (5) number of laws, hypotheses, and/or axioms,[9] (6) description length,[10] (7) number of models or interpretations of the theory,[11] and (8) number of tests and/or experiments required to falsify or verify the theory.[12] The eight definitions above do not exhaust all philosophical attempts to define simplicity, but the list does capture the bulk of them. Moreover, many philosophers identify one or more of the above definitions. Karl Popper, for instance, argues that the number of free parameters in a theory is directly related to its ability to be falsified.

Despite their *prima facie* differences, the eight definitions of simplicity above are similar in many ways. For example, definitions three through seven characterize simplicity in terms of *linguistic* features of scientific theories. In particular, definitions three through six characterize the *syntactic* complexity of a theory, whereas seven concerns its *semantic* complexity. Moreover, because the number of causes and entities postulated by a theory are often directly related to the number of its parameters, predicates, hypotheses, and models, one might view definitions three through seven as an attempt to make the first two definitions of simplicity rigorous. A second similarity amongst all the definitions is that they involve *minimizing* one quantity or another.[13]

With a host of potential definitions of simplicity in hand, one can begin to address the second and third questions above. What are the values and costs of scientific inquiry? And can a systematic preference for simpler theories help to achieve said values and minimize said costs? Unfortunately, it would be impossible to enumerate all possible goals, values, and costs of science.[14] Scien-

[5] Copernicus (1995), Galileo (1953), and Newton (1964). This definition of simplicity is closely related to that of unifying power, which is defended as a theoretical virtue by Friedman (1974), Friedman (1974), Myrvold (2003), and Kitcher (1976)

[6] See Lewis (1973), Baker (2007), and Nolan (1997).

[7] See Goodman (1950), Goodman (1952), Goodman (1955), Goodman (1958), Suppes (1956), Bunge (1961), and Bunge (1962)

[8] See Wrinch and Jeffreys (1919), Popper (1959), and Kyburg (1961)

[9] For instance, see Aristotle and Einstein's explications of simplicity, cited above.

[10] See Li and Vitanyi (1997), Li and Vitanyi (2001), and Simon (2001).

[11] See Kemeny (1955b) and Bunge (1961).

[12] See Popper (1959), Mayo (1996), and Spanos (2001). For an argument that simplicity and testability can be inversely related, see Schlesinger (1961)

[13] See Baker (2004) for a discussion of so-called "principles of plenitude," which are exceptions to the characterization of simplicity as minimizing quantities.

[14] In fact, some philosophers argue that science has no overarching aims, values, or goals. See Fine (1996). I think this view is misleading, if not altogether false. Although different

tific theories are used to improve medicine, to build bridges, cars and airplanes, to send astronauts into outer space, to provide a deeper understanding of the world, and much more. For simplicity, then, philosophers have focused almost exclusively on the relationship between Ockham's razor and the goals/values of the scientist *qua* scientists.[15] That is, irrespective of their ability to improve technology, medicine, and social policy, scientific theories can be praised if they (a) are true, (b) do not contain erroneous hypotheses, (c) are informative (i.e. contain many empirically testable consequences that are easily deduced from the hypotheses of the theory), (d) are computationally tractable, (e) make accurate predictions, and/or (f) possess other "pragmatic" or "theoretical" virtues like explanatory power, unifying power, elegance, and so on.

The arguments employed by philosophers in defense of Ockham's razor, therefore, can be divided into one of two categories, depending upon which of the goals enumerated in (a) through (f) are taken to be paramount. First, some philosophers take (a) and (b) to be the primary goals of scientific inquiry, and they argue that a systematic preference for simpler theories aids one in finding true theories while at the same time minimizing error. In this sense, simplicity is "truth-tracking." I call this type of argument a *realist defense* of Ockham's razor.[16] In contrast, other philosophers take some combination of (c) through (f) to be the primary values of science, and they argue that simpler theories are *merely useful,* in that they maximize informativeness, computational tractability, and predictive accuracy, regardless of whether they are true or not. I call such arguments *anti-realist* defenses of Ockham's razor. Of course, a systematic preference for simpler theories might be *both* useful and "truth-tracking," which is the view taken in this paper. As will become clear, however, most philosophers, scientists, and statisticians have abandoned the realist defenses of Ockham's razor.

In what ways might simpler theories be useful? One might argue that it is easier to make numerical computations and predictions with theories that contain fewer parameters and hypotheses. Consider the Ptolemaic and Copernican models of the solar system, for instance. If an astronomer wished to calculate the position of Venus using Ptolemy's theory, he would need to make roughly twice as many arithmetic calculations than had he used Copernicus', as he would need to account for Venus' speed on both the deferent and epicycle.[17]

scientists may have different goals, and although sometimes scientists will have no aims in common, many scientists do share aims and values, and it is a non-trivial matter to find and analyze what said values are.

[15] Levi calls these the "cognitive" virtues of science. See Levi (1973).

[16] Throughout this paper, I assume that that scientific hypotheses have truth values. Many philosophers contest this point, arguing that scientific hypotheses are better understood as *tools* for making predictions, for example.

[17] In Ptolemy's theory, planets move in two distinct circular paths simultaneously. One path, called the "deferent," is the circular path that a planet follows around the earth. The second path, called an "epicycle," is the circular motion a planet undertakes around a *point on the deferent.* To visualize this, think of the motion of the moon in the Copernican model. The moon rotates around the earth, but also moves in a path around the sun at the same time (because the earth does!). In Ptolemy's theory, all planets rotate around the earth in much the same way, except that they circulate about a point on the deferent (rather than a planet,

Similarly, if unified scientific theories are more explanatory, then simpler theories might provide better explanations by showing how multiple phenomena can be explained a single underlying cause. For example, Newton simultaneously explains planetary motion, the tides, and projectile motion in terms of one cause: gravity.

One might question why I have distinguished between true theories and ones that provide good predictions and explanations. Shouldn't true scientific theories always make the best predictions, for example? Unfortunately, the answer is "no." Consider the problem of curve-fitting. In curve-fitting, one measures two real-valued variables x and y and plots the resulting points on a graph. For concreteness, let x represent the number of cigarettes an individual smokes in a day, and let y represent the volume of tar in one's lungs. One is then asked to "fit" a polynomial curve (i.e. a curve of the form $f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$) to the data points on the graph. Here, a collection of curves represents a theory about how the two variables are related (e.g. tar increases quadratically in the number of cigarettes one smokes daily). Polynomials with lower degree can be specified by fewer parameters/coefficients, and Ockham's razor is reflected in the standard practice of fitting the best line to the data before trying higher degree polynomials. It turns out, that at low sample sizes, a simpler polynomial might provide better predictions *even if it is known that the true relationship between the two variables is summarized by a polynomial of higher degree.*[18] Thus, the question of whether simpler theories are more likely to be true is independent of whether simpler theories make better predictions. Similar arguments show that true scientific theories need not provide better explanations or be computationally tractable.

Although some of the most prolific scientists of all time, including Newton and Galileo, have argued that a theory's simplicity is an indication of its truth, few philosophers today advocate realist defenses of Ockham's razor. Many philosophers have been convinced by the following argument, due to Bas Van Fraassen, that there is no relationship between simplicity and true scientific theories:

> Some writings on the subject of induction suggest that simple theories are more likely to be true. But it is surely absurd to think that the world is more likely to be simple than complicated (unless one has certain metaphysical or theological views not usually accepted as legitimate factors in scientific inference).[19]

Other philosophers, like Elliot Sober and Margaret Morrison argue that simplicity and truth cannot be related because there is no single definition of simplicity that is common to every scientific discipline: biologists, psychologists, and physicists, for instance, all appeal to simplicity in different ways, and there

as the moon does) while simultaneously moving on a path around the earth. By eliminating epicycles, the Copernican theory only needs to specify the speeds at which the planets rotate around the sun, rather than their speeds on both the deferent and epicycle.

[18] See Vapnik (2000) and Forster and Sober (1994).

[19] See van Fraassen (1980).

will be widespread disagreement about the definition of simplicity even within a scientific discipline.[20] Hence, even if one definition of simplicity were an indicator of the truth of a scientific theory, there is no reason to believe that, in general, Ockham's razor is a principle that aids one in avoiding errors or finding the truth. Amongst philosophers, therefore, there is a growing conviction that while simpler theories might be useful, there is no relationship between simplicity and theoretical truth.

This thesis meets the arguments of van Fraassen, Sober, and Morrison head on. I argue that simpler theories are not only useful, but moreover, by giving systematic preference to simpler theories, a scientist will, in some sense, find a true theory in the "most direct" way. To do so, I first describe a model of scientific inquiry developed by Kevin Kelly, Clark Glymour, and Oliver Schulte (henceforth, the KGS model).[21] I then explain how Kelly, Glymour, and Schulte employ this model to provide a realist defense of Ockham's razor. The centerpiece of their arguments are the *Efficiency Theorems,* which prove that, within the KGS model, a scientist who repeatedly chooses simpler theories minimizes the number of errors and changes in opinion he makes before finding a true theory. The principal contribution of this thesis is a generalization of the Efficiency Theorem; I prove that, even amongst randomized methods for choosing theories from available evidence, a preference for simpler theories is still optimal in minimizing errors and changes in opinion. A similar theorem has already been proven by the author and Kevin Kelly in another paper,[22] but the one conjectured here is far more general. Hence, it has a number of different important philosophical implications that are discussed below.

The structure of the thesis is as follows. In the first chapter, I summarize several definitions of "simplicity" that have been offered by philosophers, scientists, and statisticians. Here, I outline three criteria that any definition of simplicity ought to satisfy so that one can meaningfully ask whether there is a relationship between theoretical simplicity and truth, and I argue that existing definitions of simplicity generally fail to satisfy at least one of the criteria.

In Chapter two, I provide a detailed description of the KGS model (including the definition of theoretical simplicity within the model), and I state the Efficiency Theorems. To show how the KGS model is relevant problems of interest to working scientists, I explain how the it can be used to represent learning causal relationships between a finite number of variables. In outlining the KGS model, I argue that its definition of simplicity satisfies the three criteria discussed in the first chapter.

In the third chapter, I generalize the Efficiency Theorem to consider arbitrary, randomized strategies for inferring theories from data. To do so, I represent the KGS model of inquiry as a two-person, strictly competitive game

[20] See Sober (1985) and Morrison (2000). Although Morrison's book is about "unification" rather than simplicity, she clearly thinks that her arguments apply to *any* standard theoretical virtue (including simplicity) discussed in philosophy of science.

[21] See Kelly (2007), Kelly and Glymour (2004), and Schulte (1999a).

[22] See Kelly and Mayo-Wilson (2008).

between a scientist and a player called "Nature."[23] The game-theoretic representation of the KGS model is mathematically useful, but I should stress that, unlike Descartes' use of an evil demon to justify the method of doubt, the philosophical conclusions that I reach do need not depend upon thinking of "Nature" as an active, villainous agent trying to thwart a scientist's attempts to understand the world. Rather, one can interpret the (mixed) strategies employed by Nature in my game as representing *the scientist's beliefs* about the likelihood of various possible worlds. This will be made more clear in subsequent chapters.

Given the game-theoretic framework, I lay the foundation for a generalization of the Efficiency Theorems, which would be philosophically important for two at least reasons. First, there is a multitude of existing Bayesian arguments for Ockham's razor, all of which conclude that simpler theories are better confirmed. All such arguments to date, however, implicitly *assign higher prior probability* to simpler theories, thereby begging the question as to why one ought to think simpler theories are more likely to be true.[24] I prove a series of theorems that suggest that, even when complex theories are assigned high probability, a preference for simpler theories remains rational.

Second, Kelly, Glymour, and Schulte's arguments justify Ockham's razor by use of a decision rule that is a variant of maximin. In other words, the Efficiency Theorems prove that heeding Ockham's razor only minimizes errors and changes of the opinion *in the worst-case*, in some sense. This leaves open the question, "Is there any (prior) probability distribution on possible worlds under which heeding Ockham's razor actually *maximizes expected utility*?"[25] The central theorem of this paper suggests that the answer to this question is "yes," thereby proving that a preference for simpler theories not only minimizes particular costs in the worst-case, but moreover, such a preference is in fact *optimal* in certain cases (i.e. when one has a particular prior distribution on worlds).

[23] Game-theoretic terms like "strictly competitive" are defined at the outset of Chapter 3.

[24] See Kelly (2009a) for an analysis of this argument.

[25] To be clear, the costs of inquiry in the KGS model are not representable by utilities, but rather by vectors of real numbers representing several different, often incomparable costs. Again, this will be made clear in subsequent chapters, but one can ask a similar question about whether there is a prior distribution under which Ockham's razor minimizes expected cost, where expectation is taken coordinate-wise in the vectors.

Chapter 1

Simplicity Defined

Although it has long been recognized that simplicity plays an important role in scientific theorizing, it was not until the twentieth century that philosophers, scientists, and statisticians attempted to rigorously analyze the concept. This chapter summarizes the attempts of twentieth century thinkers to properly define simplicity and complexity. All definitions of simplicity, I argue, fall into one of three groups.

Philosophers, scientists, and statisticians in the first group analyze simplicity in terms of the linguistic features of scientific theories. For such philosophers, a theory is generally defined to be a set of sentences expressed in a first-order language with equality. For brevity, I call such definitions of simplicity *linguistic complexity measures.* Advocates of linguistic complexity measures include Nelson Goodman, Patrick Suppes, Henry Kyburg, Harold Jeffreys, Dorothy Wrinch, and John Kemeny. For these philosophers, linguistic complexity is measured either by syntactical features of first-order theories, like the number of free parameters or predicates in the theory, or by semantic properties, like the inter-definability of predicates and the number of models of the theory.

The second group of philosophers and statisticians, including Karl Popper and Deborah Mayo, identify simplicity with *testability* or *falsifiability in principle.* Here, the qualification "in principle" is important: the hypothesis that an object weighes five pounds is no more falsifiable than the claim that it weights one tenth of a pound, even if I do not possess a sufficiently precise scale. Unless there are mathematical or scientific reasons, that I cannot, in principle, measure a tenth of a pound, then the two claims are equally falsifiable. Note that the identification of simplicity and falsifiability or testability may be justified via an appeal to measures of syntactic simplicity. For instance, Popper argues that scientific hypotheses are simpler if they contain fewer quantifiers, and hence, the connection between simplicity and falsifiability is mediated, in part, by the connection of the two with the number of quantifiers in a hypothesis.

Importantly, by defining simplicity in terms of the linguistic features of scientific theories or testability in principle, philosophers like Goodman and Popper assume that the simplicity/complexity orderings *do not depend* on what data have been recorded or which phenomena have been observed. In contrast, the final group of philosophers, who view simplicity as "data-reduction" or "unification," implicitly define a simplicity ordering of scientific theories *as a function of the data or phenomena to be explained.* This approach is taken by Herbert Simon, Paul Vitanyi, Ming Li, R.A. Fisher, Aris Spanos, and to a certain extent, Michael Friedman and Wayne Myrvold.[1] How does the approach of these philosophers and statisticians differ from that of the first group? When simplicity is a function of observed phenomena, it is possible for a theory T_1 to be simpler than T_2 relative to phenomenon P but more complicated relative to P'. Such a situation is impossible for philosophers who define simplicity in terms of the linguistic features of scientific theories.

The distinction between the approach taken by the first two groups of philosophers and that taken by the third has largely been ignored by philosophers of science. Even those, like Friedman, who implicitly recognize that simplicity might be a function of the phenomena to be explained often vacillate when explaining their definition of simplicity. The distinction is important for at least two reasons. First, those who advocate linguistic complexity measures assume that the complexity of a scientific theory can be determined *a priori;* philosophers in the second group are not committed to the view. In defending Ockham's razor, then, the first group of philosophers must defend significantly stronger claims. Suppose one adopts a linguistic complexity measure and defends Ockham's razor by arguing that simpler theories are more likely to be true. As linguistic simplicity can be determined *a priori,* then one is committed to the claim that simplicity is an *a priori* indicator of truth. That is an extremely strong claim, and it's one that those who define simplicity in terms of data reduction are not committed. Second, Sober's and Morrison's argument that there is no single definition of simplicity that applies in all contexts clearly attacks the arguments presented by the first group of philosophers and statisticians, but it's not clear that it undermines the second group's arguments at all, as the second group would allow what constitutes a simple theory to differ from one discipline or problem to the next.

Ultimately, I advocate a definition of simplicity (namely, that of the KGS model) that borrows insights from all of the definitions presented in this chapter, but which differs, at least formally, from each in substantial ways. For instance, the definition of simplicity that I advocate, when used to analyze complexity of various curves in non-linear regression, agrees with many of the linguistic and falsifiability complexity measures that lower degree polynomials are simpler than higher degree ones. And in causal inference, the definition of simplicity that I advocate resembles that of the third group of philosophers and statisticians who argue that simpler theories are more unified, thereby explaining multiple phenomena by reference to a few causes. Studying previous attempts to explicate

[1] Unfortunately, to keep this thesis at a manageable length, I will focus entirely on the philosophical literature dedicated to "unification," and I will omit expositions of (i) Fisher's and Spanos' identification of simplicity with "data reduction" and (ii) the huge literature on minimum description length, which is employed in philosophical arguments by Simon, Vitanyi, and Li. A more thorough literature review might be available later.

simplicity, therefore, makes clear in what ways my arguments borrow ideas from others, and in what ways they differ.

Yet a central question remains: why is there any debate over the *definition* of "simplicity?" Definitions, as standardly understood, are neither true nor false, and so one might wonder why one cannot see the existing philosophical literature as a summary of different *ways* in which a theory might be simple. Recall that the central reason that philosophers have sought to define simplicity is so that they can ask two related questions: (1) What role ought theoretical simplicity play in scientific theorizing? i.e. what reasons (if any) are there for preferring simpler theories in inquiry? (2) In particular, is there any sense in which simpler theories are "more likely" to be true? If these questions have meaningful and non-trivial answers, then the definition of theoretical simplicity ought to satisfy at least three criteria:

1. **Clarity:** Any definition of theoretical simplicity ought to allow one to determine, in some class of scientific problems or historical case studies, precisely which theories (if any) are simpler than others and which theories (if any) are incomparable in terms of complexity. Otherwise, such a definition of simplicity is not sufficiently precise to allow one to evaluate the use of simpler theories in practice.

2. **Relevance to Scientific Practice:** Any definition of theoretical simplicity ought to agree with intuitive assessments (i.e. those of some practicing scientists) of simplicity in some class of scientific problems. Otherwise, definitions of "theoretical simplicity" are merely explications of some concept other than what is called simplicity in science.

3. **Data Driven:** The simplicity or complexity of a theory ought to be a function of what phenomena and/or data have been observed. Otherwise, theories that are refuted by existing evidence might still be deemed "simplest," and the question of whether simplicity is an indicator of truth is settled. On the other hand, any definition of simplicity ought not identify "simple" with "more probable" or "better confirmed." Otherwise, the definition of simplicity renders trivial the question of the relation between simplicity, truth, probability, and confirmation.

As I review each of the three families of definitions of simplicity, I argue that most fail to meet (at least) one of the above criteria. Goodman's and Suppes' definitions of simplicity, for instance, are of questionable value in assessing the complexity of any "real-world" scientific theories, and moreover, neither definition renders theoretical complexity a function of existing evidence; hence, they fail both of the latter criteria. On the other hand, while those who identify simplicity with "unifying power" seem to have picked out a definition of simplicity that is both relevant to scientific practice and makes complexity a function of observed phenomena (e.g. in the writings of Friedman and Kitcher), applying their definitions to new problems or historical case studies is often nearly impossible for lack of sufficiently clear definitions of simplicity. In the third chapter, I argue that the definition of simplicity in the KGS model meets all three criteria, thereby allowing one to reasonably ask about the possibility of a realist justification of Ockham's razor.

1.1 Linguistic Simplicity

1.1.1 Syntactic Simplicity

The most systematic attempt to analyze the simplicity of scientific theories in terms of syntax was made by Nelson Goodman in the 1950's (See Goodman (1955), Goodman (1958)). In a series of articles, Goodman argues that the complexity of a scientific theory can be defined as a function of its *predicate basis.* Roughly, one can think of a predicate basis of a scientific theory as the collection of adjectives and terms that one might use to describe objects in the theory. For example, a theory from particle physics might contain predicates for mass, charge, spin, and momentum, as these are important properties of particles. Similarly, a theory describing the basic elements appearing in the periodic table might contain predicates for atomic weight and atomic mass.

More rigorously, Goodman analyzes the complexity of scientific theories (expressed in a natural language) by imagining that they have been translated into a first order language $\mathcal{L}$, which contains equality and a countable number of predicate symbols of any finite arity. He assumes that each extra-linguistic predicate (i.e. all predicates except equality) in the the scientific theory is assigned a predicate symbol in $\mathcal{L}$. For instance, the assertion "particle p has mass r" might be translated as the first order formula, "$M(p,r)$." When all the predicates of a scientific theory have been expressed in a first order language $\mathcal{L}$, Goodman assumes that it makes sense to analyze the complexity of the *scientific* predicate (e.g. mass) using the most commonly studied properties of predicates in *mathematical* theories, mainly, their arity[2] and whether they possess properties like transitivity, symmetry, and reflexivity.[3]

Goodman fails to explicitly define a number of terms that he uses in developing an axiomatic theory of simplicity. In particular, he omits definitions of the two most important concepts in his axiomatization: relevant kinds and the

[2] The *arity* or a predicate is the number of objects it relates. For instance, "red" is a unary predicate because it describes only one object (e.g. "roses are red.") In contrast, "is redder than" is a binary predicate because it is used to describe the relationship between two objects (e.g. "a rose is redder than a violet"). We can always find predicates of arbitrarily large finite arity. For example, the predicate "forms a Baker's dozen" has arity 13, and the predicate "forms a flock of geese" might take 27 arguments (depending upon where one wishes to draw the line as to how many geese constitute a flock!).

[3] A binary predicate P is *transitive* just in case xPy and yPz implies that xPz. For example, "taller than" is a transitive relation because if Jimmy is taller than Johnny, and Johnny is taller than Jane, then Jimmy is taller than Jane. A binary predicate is *symmetric* just in case xPy implies yPx. The predicate "is a sibling of" is symmetric because if Jimmy is a sibling of Jane, then Jane is a sibling of Jimmy. Finally, a binary predicate is *reflexive* if it relates every object to itself. The predicates "is at least as tall as" and "is identical to" are both reflexive.

notion of being always replaceable. I provide the most charitable reconstruction of his argument that I can below. Let $\mathcal{L}$ be a first order language with an infinite number of predicate symbols of any finite arity, and let $\mathcal{L}^P$ denote the set of all predicates in $\mathcal{L}$. Fix a natural language, say, English. Because the number of English sentences is countable (as English sentences are finite strings from a finite alphabet), we may suppose that $\mathcal{L}$ contains a predicate symbol corresponding to every predicate in English. If B is a finite subset of $\mathcal{L}^P$, we define a *predicate basis* to be the interpretation B^M of the corresponding English predicates. For convenience we drop the superscript M from now on. Define $K_{m,n} = \{B \subset \mathcal{L}^P : B \text{ is a predicate basis with } m \text{ elements and}$ $\text{every predicate in } B \text{ has arity } n\}$. We then define the set $\mathcal{K}$ of *relevant kinds* inductively as follows:

1. For any natural numbers m and n, the set $K_{m,n}$ is a relevant kind.

2. Let B be a predicate basis, and B_{sym}, B_{trans}, and B_{ref} respectively denote the set of symmetric, transitive and reflexive predicates of B. Then for any relevant kind, $K_{sym} := \{B_{sym} : B \in K\}$, $K_{trans} := \{B_{trans} : B \in K\}$, and $K_{ref} := \{B_{ref} : B \in K\}$ are all relevant kinds.

3. If K and K' are relevant kinds, then $K \wedge K' := \{B \cup B' : B \in K, B' \in K'\}$, $K * K' =: \{B \cap B' : B \in K, B' \in K'\}$, and $K|K' := \{B - B' : B \in K, B' \in K'\}$ are all relevant kinds..

For example, the set $\hat{K} = \{\{P, Q\} : P \text{ is a unary predicate and } Q \text{ is a}$ $\text{transitive binary predicate}\}$ is a relevant kind. What is the motivation for the definition of relevant kinds? Goodman's insight is that a predicate basis of a particular relevant kind K is often *routinely* replaceable by a predicate basis of a different relevant kind K'. For example, three unary predicates "x is red," "y has teeth," and "z behaves wildly" are replaceable by a single ternary predicate "x is red, y has teeth, and z behaves wildly." In general, any n unary predicates $\{P_1, \ldots, P_n\}$ can be replaced by a single n-ary predicate Q such that $P_1(x_1) \wedge \ldots \wedge P_n(x_n) \leftrightarrow Q(x_1, \ldots, x_n)$.

Essentially, a predicate basis B can be replaced by another basis B' only if everything that is expressible in terms of B is expressible in terms of B'. Intuitively, predicate bases with greater expressive power ought to be more complex, and so replacing one predicate basis by another cannot effect any simplification. Goodman's major axiom in defining simplicity, therefore, states that routine replacement of one predicate basis by another can never decrease complexity. Unfortunately, Goodman never explicitly states what counts as routinely replacing one predicate basis by another. I make the notion as precise as I can. First, note that for any relevant kind K, any basis $B \in K$ has the same size. As relevant kinds are defined recursively, this can be proven by induction. Let n_K denote the number of predicates in a given basis $B \in K$.

Now, for any relevant kind K, there is a second order formula $A_K(X_1, \ldots, X_{n_K})$ that holds of every predicate basis $B = \{P_1, \ldots, P_{n_K}\}$ in K, when the predicates of B are enumerated in a particular order. For example, given the relevant kind $\hat{K}$ defined above, the second order formula $A_{\hat{K}}(X_1, X_2)$ defining K is as follows.

$$\exists x_1 X_1(x) \wedge \exists x_1 \exists x_2 X_2(x_1, x_2) \wedge \forall z_1 \forall z_2 \forall z_3 [X_2(z_1, z_2) \wedge X_2(z_2, z_3) \rightarrow X_2(z_1, z_3)]$$

The first two conjuncts in the above formula state that X_1 and X_2 are nonempty, which Goodman requires, and the last conjunct states X_2 is transitive. Given a relevant kind K and any arbitrary predicate basis $B \in K$, therefore, let $B_i \in B$ be an enumeration of B such that $A_K(B_1, \ldots, B_{n_K})$ holds, and let a_i be the arity of B_i.

Now we say a relevant kind K is *always replaceable* by a kind K' if there exist n_K many second-order formula $\psi_i(X_1, \ldots, X_{n_{K'}}, x_1, \ldots, x_{a_i})$ such that for all $B = \{B_1, \ldots, B_{n_K}\} \in K$ the following sentence is consistent:

$$\exists X_1 \ldots \exists X_{n_{K'}} (\varphi_{K'}(X_1, \ldots, X_{n_{K'}})) \wedge [\forall x_1 \ldots \forall x_{a_i} \bigwedge_{1 \leq i \leq n_K} [\psi_i(X_1, \ldots, X_{n_{K'}}, x_1, \ldots, x_{a_i}) \leftrightarrow B_i(x_1, \ldots, x_{a_i})]]$$

In other words, relevant kind K is always replaceable by K' if there is a method for defining all the predicates in a basis $B \in K$ in terms of the predicates in a basis $B' \in K'$ *knowing only that B is of kind K, regardless of its interpretation.* Given these definitions, Goodman offers the following axioms that constrain a choice of a complexity function:

Axiom 1.1.1. Every extralinguistic predicate has positive complexity

Axiom 1.1.2. The complexity of a predicate basis is the sum of the complexity of the predicates it contains.

Axiom 1.1.3. The complexity of a relevant kind is the greatest complexity of the predicate bases it contains.[4]

Axiom 1.1.4. The complexity of a predicate basis is equal to the complexity of the narrowest kind that contains it.

Axiom 1.1.5. Let K and L be relevant kinds. If every K is always replaceable by L, then K does not have greater complexity than L.

Given the above axioms, Goodman argues that there are three properties most relevant to evaluating the complexity of a predicate: (1) arity, (2) "symmetricity degree," and (3) "self-completeness degree." For instance, from the above axioms, Goodman claims that one can prove that if P is an n-ary predicate, Q an m-ary predicate, and $n > m$, then the predicate basis $\{P\}$ is more complex than the basis $\{Q\}$.

For scientists and philosophers without extensive training in formal logic, Goodman's proposal to analyze the complexity of predicates might seem odd.

[4] Though he never says so explicitly, Goodman's third axiom assumes no relevant kind has infinite complexity, and hence, by Axioms 1 and 2, the predicate bases in a relevant kind must have a bounded size. This is why I have chosen to formalize Goodman's notion of relevant kind by letting the sets $K_{m,n}$ be a basis for relevant kinds instead of the more natural basis consisting of all finite sets of n-ary predicates

It might seem arbitrary, for example, to say that "red," "orange-smelling," are any less complex than "bigger than a breadbox" or vice versa. Is the complexity function that Goodman describes relevant to judgments of simplicity made in various sciences? Or does Goodman argue that his complexity measurement will be useful for any purpose?

Goodman admits that theories might be judged "simple" in any number of ways, that judgments of simplicity are made for various reasons (e.g. making rough estimates versus making predictions), and what a particular scientist considers to be simple might depend on subjective factors. In response, Goodman argues that several millennia ago, judgments of size might have been viewed similarly. Thin yet tall objects might be judged large, and similarly for short but wide objects. Moreover, judgments of size can be affected by lighting, distance, perspective, and so on. By fixing a unit length (say a meter), however, one can start to make precise judgments of size, and one can precisely differentiate between length, width, and depth. Of course, different types of measurement of size will be relevant in different circumstances (the height of a giraffe is relevant to explaining why it can eat leaves from tall trees; its girth is not), but this does not mean we shouldn't try to provide precise measurements of size.

Similarly, the axiomatization presented above, Goodman argues, may only measure one "dimension" of simplicity, but it is a first step towards making one such precise measurements of the simplicity of scientific theories. Still, even on its own accounts, Goodman's approach faces a number of problems. The first, and perhaps most important, is raised by Suppes (1956): Goodman never uses his axiomatization to analyze the complexity of any well-known, axiomatized mathematical theory (let along a scientific one!), so one might question whether his axiomatization captures *any* of the relevant "dimensions" of simplicity at all. Goodman never provides an example from the history of science in which his axiomatized concept of simplicity was invoked (knowingly or unknowingly) by a scientist or metaphysician. Further, Goodman never explains how his axiomatization might be a useful measurement of complexity, say, for making predictions, even if it had never actually been implicitly invoked by any scientist.

Ignorance, however, is not a good objection: although Goodman does not describe an episode from the history of science in which his definition of simplicity was used, there still might be one. So how does Goodman's theory do when we try to apply it to examples from the history of science? Not well. Kyburg convincingly argues that Goodman's measure assigns equal complexity to the Ptolemaic and Copernican models, as neither theory's predicate basis is more complex than the other. I argue Goodman's theory also fails to capture any relevant sense of simplicity in curve-fitting as well. Consider the set of all functions $f : \mathbb{R} \to \mathbb{R}$. According to Goodman's measure of complexity, the function $f(x) = x$ is uniquely simplest amongst all functions because no other functions are reflexive, symmetric, transitive, or self-complete. For the same reason, every other function is equally complex according to Goodman's measure, and so, for example, equal complexity values are assigned to $g(x) = 3x$ and $h(x) = \sum_{n=0}^{\infty} \frac{\sin(x)}{n!\pi^n}$. Moreover, because binary relations that are symmetric, transitive, or self-complete are never *functions,* if scientists preferred Goodman-simpler theories, they would almost never accept a hypothesis that asserted a functional relationship held between two variables. Finally, even if we are to accept Goodman's highly un-intuitive simplicity ranking of functions, it's clear his definition is useless in curve-fitting: faced with the choice between two functions that explain the data equally well, Goodman's measure of simplicity cannot be used to break ties unless one of the functions is the identity mapping.

The second problem, raised by both Bunge (1961) and Suppes (1956), is that Goodman's measure of complexity does not adequately take into account the *relationship between* predicates in a basis. To illustrate this point, I develop an example that is simpler than those developed in Suppes. Consider a basis consisting $B = \{\preceq, \prec\}$ of a strict linear order $\prec$ and its reflexive closure $\preceq$ (i.e. $x \preceq y \leftrightarrow x \prec y \vee x = y)$. Then consider another linear order consisting of two *unrelated* linear orders $B' = \{\leq, \sqsubset\}$, where again the second predicate is non-reflexive but the first is. According to Goodman's axioms, both bases are equally complex, as they belong to the relevant kind consisting of two linear orders, one of which is reflexive and one of which is irreflexive. However, because the two predicates in B can be defined in terms of one another, one might think that former basis is simpler than the latter. If the axioms governing the relationship between predicates in a basis B are relevant to determining its complexity, therefore, we need not argue that Goodman's measure of complexity fails to capture enough features of scientific *theories:* Goodman's axiomatization may fail to meet its stated goal of measuring complexity for predicate bases.

Though obviously problematic, Goodman's approach inspired Patrick Suppes, Mario Bunge, Henry Kyburg, and John Kemeny to use the tools of formal logic to develop alternative axiomatic definitions of complexity. In this section, I describe Suppes, Bunge's, and Kyburg's alternative proposals, as their proposals, like Goodman's, focus on the syntactical features of a scientific theory. In the next section, I summarize Kemeny's proposal.

In light of the second criticism above, Suppes defines a complexity ordering $\preceq$ on *ordered pairs* (P, A) where P is a predicate basis and A is a set of axioms describing the relationship between the predicates. For example, Suppes first three axioms state that $\preceq$ is a total, transitive ordering and that $(P_1, A_1) \preceq (P_2, A_2)$ whenever $A_1 = A_2$ and $P_1 \subseteq P_2$ (this last axiom implies $\preceq$ is reflexive). Note that Suppes does not argue that ordered pairs (P, A) can be assigned a *numerical* value as Goodman does, but rather, argues that complexity orderings may be entirely *qualitative.*[5] Thus, for Suppes, simplicity is primarily a *relation*

[5] In fact, at the time Suppes wrote his article, he said it was unknown whether the axioms implicitly defined a real-valued function f such that $f(P, A) \leq f(P', A')$ if and only if $(P, A) \preceq (P', A')$. Suppes' axioms, however, were chosen to be formally analogous to de Finnetti's and Savage's axioms for a binary qualitative probability relation, and furthermore, Savage and de Finnetti's axioms do induce strictly agreeing quantitative probability functions under various assumptions about the structure of the σ-algebra on which the qualitative relation is defined. See Savage (1972) for a discussion of these issues and an extended bibliography concerning the various axiomatizations for qualitative probability. I have not been able to verify whether Suppes' axioms do meet the conditions guaranteeing the existence of a quantitative complexity measure, though I think it is likely the question has been resolved.

amongst theories. In other words, Suppes is not committed to the view that the complexity of a single theory can be assessed in isolation, but rather, its simplicity may only be determined *in relation to that of other existing theories.* This is an interesting distinction between Suppes' proposal and, for instance, that of Goodman. The definition of simplicity that I advocate also shares this feature, but to my knowledge, no philosopher has analyzed he importance of a relational versus non-relational analysis of simplicity; further work is needed here.

Unfortunately, as his axiomatization is proposed only as an alternative to Goodman's, Suppes likewise does not provide any reason to believe his axioms reflect any useful measurement of simplicity or have ever been employed in science to decide between competing theories. Moreover, Suppes' axiomatization can easily be seen to fail to provide any useful measure of complexity of various functions in curve-fitting.

Like Suppes, philosopher Mario Bunge argues that Goodman's definition of simplicity fails to consider several ways in which certain predicates are (intuitively) simpler than others. For example, Bunge distinguishes between *atomic* predicates like "is black" and *molecular* predicates like "is a black crow," which can be expressed using boolean combinations of atomic predicates. Intuitively, a molecular predicate ought to be more complex than the atomic predicates from which it is constructed, but Goodman's formalism does not seem to capture this distinction. Bunge also distinguishes between predicates of different *order.* For example, "red" might be a first order predicate, as it describes objects. "Dark," on the other hand, might be a second order predicate, as one can use it to describe the first order predicate "red." Finally, Bunge criticizes Goodman for failing to distinguish between discrete predicates (e.g. "has x ears") from continuous ones (e,g, "has mass x"); the latter, he thinks, are infinitely more complex than the former, although this intuition is unmotivated.

Bunge's most important contribution to the simplicity debate, however, is to note that it is not straightforward to use a linguistic complexity measure for *predicate bases* to define a complexity measure for scientific *theories.* One might try measure the complexity of a sentence in terms of the predicates it mentions, the number of connectives it contains, and the number and order of quantifiers that appear. A complexity measure on sentences could then be used to define the complexity of first-order theories (i.e. sets of sentences). However, Bunge argues that, intuitively, what matters in measuring the complexity of a sentence is the *relationship* between the predicates, connectives, and quantifiers. He concludes "no adequate measure of complexity of propositions or propositional functions is available."[6] Because measuring the syntactic complexity of theories requires both (a) a complexity measure for sentences and (b) an analysis of the relationship between the sentences of a theory, Bunge concludes that, like propositions, no syntactic complexity for theories is available. Bunge does leave open the possibility that there may be alternative, non-syntactic measures of complexity.

[6] See Bunge (1962), pp. 121.

Like Bunge, several other mid-twentieth century philosophers and statisticians, including Henry Kyburg, Harold Jeffreys, and Dorothy Wrinch, realized that analyzing complexity in terms of predicate bases was both excessively complicated and tenuously motivated. As a result, they shifted their focus to a different syntactic feature of scientific theories: the number and order of quantifiers. For instance, Henry Kyburg makes the "modest proposal" that the complexity of a scientific theory can be measured exclusively by the *number* of quantifiers contained in a its axioms and observation sentences, and he claims this is a more obvious way of understanding Ockham's razor: "Do not multiply entities needlessly" (See Kyburg (1961)). Unfortunately, Kyburg's theory seems incapable of distinguishing between many infinite theories. Let T be a theory, $A \subseteq T$ be its axioms and observation sentences, and $Q \subseteq A$ be the set of sentences of A containing a quantifier. If Q is infinite, then, according to Kyburg's measure, T has infinite complexity. Moreover, scientific theories with an infinite set of axioms are quite common, as almost any phenomenon that is described mathematically will require a theory to contain arithmetical axioms, and hence, an induction schema.[7]

Harold Jeffreys and Dorothy Wrinch proposes a slightly more complicated measure of complexity that accounts for both quantifier order and number, but their proposal faces the same problem as Kyburg's.[8] Finally, in recursion theory and descriptive set theory, there is a long tradition of measuring complexity in terms of quantifier order, as is clear from the study of the analytic, arithmetic, and Borel hierarchies. According to some philosophers, the simultaneous use of quantifier order as a measure of complexity in recursion theory and in science suggests a deep connection between the problem of induction and computation.[9]

1.1.2 Semantic Simplicity

Despite Goodman's insight that simplicity is a multi-faceted concept, both he and Suppes attempt to provide a *single* definition of simplicity. In contrast, Mario Bunge and John Kemeny provide multiple measures of complexity, and they argue that the different measures are indicators of different values in scientific inquiry. [10] For example, Bunge provides a laundry list of theoretical virtues, like "coherence," "testability," and "forecast power," and he argues that different measures of simplicity will simultaneously be indicators of these different virtues. In the previous section, I described how Bunge analyzes the syntactic complexity of a *theory* by expanding upon Goodman's complexity measure for a collection of *predicates*. Like Bunge, Kemeny also refines Goodman's complexity measure for predicate bases. However, both Kemeny and Bunge argue that

[7] What is standardly called "Peano Arithmetic," for instance, contains an axiom of the following form for every formula $\varphi(x)$ with a single free variable:

$$[\varphi(0) \wedge \forall n(\varphi(n) \to \varphi(S(n))] \to \forall n(\varphi(n))$$

[8] See Wrinch and Jeffreys (1923), for instance.

[9] See Kelly and Schulte (1997).

[10] See Bunge (1961), Bunge (1962), and Kemeny (1955b).

there are other measures of complexity of scientific theories (again, expressed in a first-order language) that require analysis of their *semantic* properties. This section describes the semantic measures of complexity.

Kemeny defines the simplicity of a scientific theory (again, expressed in a first-order language) in terms of the number of models it has.[11] For Kemeny, models represent "worlds," and so different models of the same scientific theory represent different "possible worlds," in Leibniz's sense. The obvious problem with this model-theoretic approach, which Kemeny himself realizes, is that most first order theories have infinite models. Moreover, if a first-order theory T has an infinite model, then the Löwenheim-Skolem theorem implies that T has a model of every infinite cardinality. Hence, according to Kemeny's second measure, not only do many theories have infinite complexity, but further, their complexity cannot be bounded by *any* infinite cardinal.

Kemeny claims that the type of theories that occur within science, however, will not have infinite complexity. Scientific theories may describe an extremely large number of objects, such as the theory that describes the number of atoms in the universe between 10 Billion B.C.E. and 100 Billion A.D., but such theories, at root, are still only concerned with a *finite* number of objects. Thus, Kemeny argues that typical scientific theories will not have infinite complexity, as for a fixed first-order language $\mathcal{L}$ with a finite number of predicate and function symbols, there are only a finite number of $\mathcal{L}$-structures of a fixed, finite cardinality.

Kemeny's argument, however, is clearly problematic. In virtually every scientific domain, widely accepted theories require representing physical quantities by real numbers. For example, consider any theory containing any of the following predicates: mass, space, time, pressure, volume, and temperature. Models of these theories will contain a continuum number of elements. Moreover, the use of real numbers is not a defect, nor does it complicate such theories. In fact, the use of continuous quantities instead of discrete ones is often view as an attempt to *simplify* an otherwise unwieldy theory (Get a reference from Ben on fluid mechanics). Furthermore, some physical theories go so far as to employ complex numbers as a simplifying tool, even when the imaginary parts of said numbers have no physical interpretation, as for instance, in the theory of RC-circuits. While Kemeny might be correct that, at root, many physical quantities are discrete and can only take on a finite number of values, any complexity measure that fails to recognize the simplifying role that continuous quantities play in scientific theories fails to accurately capture the way Ockham's razor is used in science.

[11] Technically, Kemeny uses the word "interpretations." See Kemeny (1956a). However, Kemeny's definition of an interpretation is now the standard definition of a model. At the time of Kemeny's article, the word "model" was used to describe Tarski's definition of a model, which is no longer standard. See Etchemendy (1999), pp. 166, footnote number two, for a discussion of the differences.

1.2 Falsifiability and Testability

Like the first group of philosophers, Popper likewise measures the complexity of a scientific theory in terms of its dimension or number of free parameters. However, he also argues that simplicity ought to be identified with falsifiability, which is the key characteristic of scientific hypotheses.[12] To understand Popper's discussion of simplicity, however, one must first review other key concepts in his epistemology.

In stark contrast to the majority of scientists, philosophers, and statisticians, Popper argues that experiments and observational studies can never provide support for a scientific hypothesis. Rather, hypotheses may only be *falsified.* For Popper, scientific theories, which are conjunctions of hypotheses, *logically entail* that particular phenomena will be observed under particular circumstances. Newton's theory of gravity, for example, might logically entail that scientists will observe high tides during night. For Popper, then, a theory is falsified just in case it entails a phenomenon that is not observed.

Why might falsifiability be a good measure of simplicity? Popper provides several examples to prime his readers' intuitions., Here, I discuss Popper's analysis of curve-fitting, which is illustrative but also notoriously problematic.[13] Consider the problem of finding which polynomial curve best describes the relationship between two fixed variables x and y. Here, Popper claims, a hypothesis is a particular polynomial f, and the hypothesis f is falsified just in case a point (x, y) is observed that is not a member of the graph of $f(x) = y$. If three non-collinear points are observed, then, assuming the data contains no errors, the every linear hypothesis would be falsified. In contrast, there is always a quadratic equation compatible with three points. Hence, Popper concludes that linear hypotheses are simpler than quadratic ones, and similarly, the falsifiability of a polynomial is directly related to its degree.

Of course, Popper's analysis of curve-fitting is problematic. In the absence of statistical error, observing the point $(-1, 0)$ is sufficient to refute the quadratic hypothesis $y = x^2$. In general, if a hypothesis is identified with a *single* polynomial, then there exists a single point that is sufficient to falsify any given hypothesis (no polynomial covers every point on the plane!). Had Popper defined a hypothesis to be the disjunction of the set of polynomials of a fixed degree, then the hypothesis "There is a linear relationship between x and y" would be falsified by three non-collinear points, while the hypothesis "There is a quadratic relationship between x and y" could be falsified only by four or more points. In this way, falsifiability might be an intuitive measure of simplicity.

Popper's definition of simplicity, however, faces two serious problems. First and foremost, experimental data is hardly ever conclusive enough to falsify a

[12] Popper himself notes that he is often misinterpreted as claiming *meaningful* statements are falsifiable. According to Popper, falsifiability is a criterion for demarcating scientific statements from non-scientific ones.

[13] See Schlesinger (1961) pp. 487-490, Sober (2010), pp. 7, and Baker (2004) for various criticisms of Popper. Importantly, Baker criticizes not only Popper, but also works like Sober and Forster's, that attempt to explain all of the various forms of Ockham's razor in science by appeal to statistical methods.

particular hypothesis. Because of measurement error, chance variation, and limits in computational precision, data is never exact. For instance, observing 400 non-collinear points in curve-fitting may not be taken as evidence against the hypothesis that the relationship between the two variables is linear, so long as the 400 points are clustered in a certain way. Popper needs a more refined notion of falsifiability, one which does not rule out a theory because of practical limits in data collection.

Second, Popper's notion of falsifiability also conflicts with intuitive judgments of theoretical simplicity. Schlesinger proposes a particular compelling counterexample.[14] In the early nineteenth century, astronomers struggled to explain perturbations in Uranus' orbit, and several rival hypotheses were proposed. One hypothesis, independently developed by John Couch Adams and Heinrich d'Arrest, was that an eighth planet, now known as Neptune, caused said perturbations. But one can also imagine a maverick nineteenth century scientist who might have hypothesized that the irregularities in Uranus' orbit were caused by the gravitational pull of twelve yet undiscovered planets. Intuitively, Adams' and d'Arrest's hypothesis seems simpler than that of the imagined scientist, as it postulates fewer causes of the phenomena in question. If one identifies simplicity with falsifiability, however, then the imagined scientist would have proposed a simpler than either Adams or d'Arrest: it is surely easier to show that one of the twelve planets does not exist rather than to show that a single planet does not.

For most philosophers of science, these two difficulties were fatal for Popper's definition of simplicity. Popper's arguments, however, inspired a small group of philosophers to address the above two issues and to defend a more refined analysis of simplicity in terms of falsifiability or testability. For example, Aris Spanos and, to an extent, Deborah Mayo, employ techniques from classical statistics to address the first deficiency of Popper's definition and to rescue the spirit of Popper's insistence on submitting scientific hypotheses to rigorous (or "severe") testing. Ultimately, however, Spanos defines simplicity in terms of number of parameters and "informational content" of a statistical model, and he argues that greater testability is a *consequence* of simplicity, not its defining feature. Arthur Schlesinger tries to address the second issue by turning Popper's definition on its head: Schlesinger claims simpler theories are, over time, more *difficult* to falsify. In the remainder of this section, I summarize Schlesinger's analysis of simplicity, and I return to Spanos' analysis later.

In evaluating the relationship between truth and simplicity, Schlesinger argues, the relevant property of scientific theories is *dynamic simplicity,* which is the property of being capable of being updated with fewer, less complicated hypotheses. Notice that Schlesinger tacitly employs two different definitions of simplicity. That is, say a theory T in is *dynamically simplest* relative to new data D if when T is brought to explain D, then its revision T_D is *statically simpler* than the revision of any other theory T' relative to data D. Notice that dynamic simplicity must be relativized to a set of not yet observed data.

[14] See Schlesinger (1961) pp. 491-492.

Schlesinger never fully defines what would make a theory T statically simpler, but he provides an example to motivate the

Suppose that you have two new neighbors, Tom and Dick, who you know, via neighborhood gossip, to be members of the same family. Suppose, however, that you are unsure of how Tom and Dick are related. One day you meet Tom and Dick, and you find that Tom and Dick have similar appearances. In particular, Tom and Dick appear to be approximately 35 years old. You hypothesize that Tom is Dick's brother, but your skeptical friend insists that you cannot be sure of your hypothesis. Tom, your friend hypothesizes, is Dick's father, but through extensive exercise, dieting, and meticulous care for his body, he has retained a youthful appearance. Call your hypothesis that Tom and Dick are brothers H_1, and call your friend's hypothesis that Tom is Dick's father H_2.

Now suppose that the day after you first meet Tom and Dick, you see Tom driving a car with a sticker advertising membership in the American Association for Retired Persons (AARP). You call your friend to assert that you now know that Tom is Dick's father, but your friend quickly retorts that Tom is Dick's father, but he likely drives a car that he inherited from a grandparent. Call your friend's new hypothesis H_2'. One can continue the story indefinitely. Suppose Tom declares "I am Dick's father" one day in conversation, and your friend adjusts his hypothesis to be the claim H_2'' that Tom is a pathological liar, drives a car that he inherited from his grandparents, and is excessively vain. Schlesinger argues that, in light of the evidence, H_2 is a dynamically less simple hypothesis than H_1, as it requires more and more *ad hoc* assumptions to explain the phenomena. Because true theories, according to Schlesinger, will not require additional, *ad hoc* assumptions, there is a straightforward relationship between (dynamic) simplicity and truth.

Despite the intuitive appeal of Schlesinger's argument, his definition of dynamic simplicity seems to be equivalent to "better confirmed by existing evidence." In the story of Tom and Dick, the hypotheses H_2' and H_2'' are not only more complicated than their rival H_1, but moreover, they, intuitively, seem less likely to be true than H_1 given the available evidence. As understood by most philosophers, the problem of justifying Ockham's razor, however, requires explicating the relationship between simplicity and other values of inquiry, all other things, *including confirmation,* being equal.

In the past two decades, Aris Spanos and Deborah Mayo have provided a more nuanced analysis of the relationship between simplicity, testability, and falsifiability.[15] Spanos, however, argues that simplicity is defined, roughly, in terms of *compression of data* and that greater testability is a consequence of being simple. Spanos' analysis of simplicity is one of many that analyze complexity in terms description length, unification, and data reduction. In all such proposals, simpler theories are defined to be those that provide the most concise summary of the phenomena and data. Although I cannot discuss Spanos' and Mayo's proposals in depth, the next section discusses the philosophical literature in which simplicity is identified with unification.

[15] See Spanos (2001) and Mayo and Spanos (2006).

1.3 Simplicity as Unification

With the exception of Mario Bunge, all of the philosophers, scientists, and statisticians discussed thus far have explicated the concept of simplicity in terms of either (a) linguistic properties of scientific theories or (b) testability or falsifiability of a theory. Importantly, both types of definitions explicate simplicity as a function of a scientific theory *alone.* In contrast, many philosophers have argued that the simplicity of a theory is both a function of properties of the theory *and available evidence.* For example, Aris Spanos writes:

> The second dimension of simplicity concerns the informational content of the statistical model ... [However,] this "compression of information" cannot be done in the abstract, but in conjunction with *relevant information in the observed data.*[16]

In this section, I summarize attempts, like Spanos', to explicate simplicity as a relation between theory and data. I begin with those philosophers, such as William Whewell and Michael Friedman, who identify simplicity with unifying power. While it is less clear that unifying power cannot be evaluated *a priori,* one can understand these philosophers as advocating a view that simpler theories are ones that can explain the phenomena that have been observed in the shortest space. In this way, unifying power is relevant similar to data reduction, a second type of definition of simplicity. Spanos, who takes inspiration from RA Fisher, is the prominent defender of this analysis of simplicity. Finally, recently, computer scientists and statisticians have attempted to identified simplicity with minimum description length, which shares motivations with the analysis of simplicity in terms of data-reduction. For reasons of length, I will focus solely on the philosophical discussion of unification and simplicity; a fuller literature review would address these latter two attempts to analyze simplicity.

In his two-volume treatise *Philosophy of The Inductive Sciences,* William Whewell identifies simplicity with the *generality* of a scientific hypothesis, where a theory T is more general than T' if T entails T'.[17] Whewell calls the process of finding more and more general hypotheses the "consilience of inductions," and he argues that, as more consilience of inductions occur, scientists can be more and more assured that they are developing *true* theories. Whewell calls the relationship between consilience and truth "the constant Tendency to Simplicity observable in true theories."[18] As an example, Whewell discusses how

[16] See Spanos (2001), pp. 83.

[17] Here, I leave open the question of how strong the notion of "entailment" ought to be understood. For example, a scientific theory T might *logically* entail T'. A theory T might entail T' only modulo standard mathematical axioms (e.g. Zermelo's axioms for set theory). A weak (but perhaps most plausible) way of defining entailment between scientific theories is to say that T entails T' just in case T' is provable from T modulo standard mathematical axioms **and** other widely accepted scientific theories. The example of the kinetic theory of gases described in this section is of this character, as one can "derive" the ideal gas law from the kinetic theory of gases only modulo axioms of real analysis **and** Newton's laws.

[18] See Whewell (1966), 77.

observable astronomical facts, known to the ancient Greeks and Babylonians, are entailed by the heliocentric model of the solar system:

> Thus, that the stars, the moon, and the sun, rise, culminate, and set are facts *included* in the proposition that the heavens, carrying with them all the celestial bodies, have a diurnal revolution about the axis of the earth. Again, the observed monthly motions of the moon, and the annual motions of the sun, are *included* in certain propositions concerning the movements of the luminaries with respect to the stars. But all these propositions are really *included* in the doctrine that the earth, revolving on its axis, moves round the sun, and the moon round the earth.[19]

Here, the (true) heliocentric hypothesis is more general than propositions describing the "monthly motions of the moon and annual motions of the sun" because the latter is entailed by the former. I describe Whewell's argument in greater detail below, and I also explain the concept of one proposition being "included" in another. For now, it is important to note that the Whewell's concept of generality is closely related to an oft discussed theoretical virtue: *unifying power.* In contemporary philosophical jargon, a theory is said to have unifying power, roughly, if it simultaneously provides an explanation of many observable phenomena and/or many other successful scientific theories. Clearly, generality and unifying power are closely related: if one identifies scientific explanations with deductions, as in the Deductive-Nomological (DN) model, then theories with unifying power will also have greater generality and vice versa.[20] Hence, Whewell's discussion of consilience might help us to understand some hotly contested debates in 20^{th} century philosophy of science. Moreover, Whewell's view is strikingly similar to that of two 20^{th} century philosophers: Michael Friedman and Philip Kitcher. In particular, I will argue that Whewell's understanding of consilience of inductions is actually *more general* (in a non-Whewellian sense) than the definition of unification provided by Friedman.

Whewell distinguishes between two parts of the inductive process: (1) invention of a conception, and (2) deduction of less general laws from more general ones. I describe the two parts in turn. Introductory textbooks in philosophy of science often use examples like the following to describe the problem of induction as follows. If one knows that all swans are white, then she can (deductively) infer that any particular swan that she observes will be white. In contrast, if one has observed five thousand swans, all of which are white, she cannot infer that the next swan that she observes will be white. After all, it is possible that, by sampling error or bad luck, one has viewed all and only the white swans in the world. Induction, according to such textbooks, is the process of making inferences that *do not preserve truth,* and the problem of induction is to explain when (or if) such inferences are reliable and/or justified.

[19] *Ibid*, pp. 75.

[20] See Hempel and Oppenheim (1948) for a discussion of the DN model, and Putnam (1973) for a criticism as to why deductions are not necessarily explanatory.

Whewell thinks the above description of induction is extremely simplified and misleading. According to Whewell, in order to observe whether a swan is white or not, one first needs to possess the concepts "swan" and "white," and in particular, one must know under what conditions an object may said to be a swan or said to be white. As such, induction is not simply the process of inferring "All objects of types X possess property P" from the observation that a particular sample of objects of type X are all P. In inductive reasoning, one must first develop the appropriate concepts before any inferences can be made whatsoever.

As an example, Whewell considers the hypothesis "All planets follow elliptical paths around the sun." In making observations of the night sky, an astronomer does not see heavenly bodies conveniently labeled with tags reading "planet," "moon," "comet," "star," and so on. Instead, she must *group* heavenly bodies according to their size, brightness, movement in the night sky, and so on. The astronomer might then form the concept "planet" to describe those bodies other than the moon which, on the basis of their size, brightness, and movement, appear to be closest to the earth in the sky. Similarly, an astronomer does not observe dotted elliptical lines in the sky indicating the paths followed by individual planets. Rather, she must document the position of different planets during different times of the month, year, and so on. She might then *classify* the movement of a planet as elliptical because, when the astronomer's observations are (i) augmented by interpolated values and (ii) represented graphically, the planet's movement resembles the geometric shape of an ellipse. Hence, before inferring that "all planets move in elliptical paths about the sun" from observations that Venus and Mars do, an astronomer must develop the concepts "planet" and "elliptical" and specify under what conditions a planet can be said to moving in an elliptical path. Whewell calls this process "invention of a conception."[21]

Although Whewell does not say so explicitly, he seems to indicate that certain concepts are directly linked to observable properties of objects. For example, the concept of "weighing five pounds" might be directly connected to observable properties by specifying a scale and conditions under which the scale is said to register five pounds as the weight of the object. In contrast, Whewell indicates that certain concepts like elliptical orbit are derivative of more basic concepts (here, cyclical motion), and hence, a hypothesis such as "the earth follows an elliptical orbit" can only be tested by testing the truth of hypotheses that contain concepts that are directly linked to observable properties. The hypothesis that"the earth follows an elliptical orbit," for example, might be tested by testing the (conjunctive) hypothesis, "the earth occupies position p_1 in the night sky at time t_1, p_2 during time t_2, and so on." In general, Whewell seems to indicate that there is a hierarchy of concepts such that concepts at the lowest level are connected with directly observable properties of objects, and the meaning of higher-level concepts is specified by a series of logical relationships (perhaps definitions?) to directly observable properties.

[21] See Whewell (1966), pp. 49-54.

Understanding Whewell in this way makes intelligible his claim that certain scientific propositions are included in more general ones. If one understands more general propositions as those containing higher level concepts, then the logical relationships between concepts will induce logical relationships between scientific hypotheses employing said concepts. For example, Whewell indicates that Newton's law of gravitation is more general than hypotheses about the tides and planetary motion. According to the above reconstruction, one can understand Whewell as saying that the meaning of Newton's universal law is specified, in part, by the predictions it entails about the tides and planetary motion. Hence, Newton's law of gravitation can be verified and/or refuted by testing less general hypotheses about tides and planetary motion, and these less general hypotheses are again verified and/or refuted by finding even less general hypotheses. The process continues until one find a hypothesis that is directly testable. Whewell writes:

> That each of these particular propositions is true may be ascertained . . . when the propositions is resolved into *its* more special propositions. And thus we may proceed, til' the most general truth is broken up into small and manageable portions. Of these portions, each may appear by itself narrow and easy; and yet, they are so woven together by hypothesis and conjunction, that the truth of the parts necessarily assures us of the truth of the whole.[22]

One can now see the relationship between Whewell's view and contemporary discussions of unifying power. For Whewell, general hypotheses are devised to entail many different well-tested hypotheses, and they contain terms denoting concepts that are (for a lack of a better term) generalizations of concepts occurring in other extant theories. General scientific theories are simpler because they reduce the number of laws and terms required to explain disparate phenomena.

Like Whewell, the contemporary philosopher Michael Friedman identifies simplicity with unifying power. To motivate his discussion of unifying power, Friedman provides several paradigmatic examples of unification: the kinetic theory of gases, the atomic theory of matter, and Newton's theory of gravitation. For brevity, I will merely summarize Friedman's view on the kinetic theory of gases, which explicates macroscopic properties of gases in terms of the velocity and collisions of the molecules that compose it. Prior to and during the development of the kinetic theory, scientists had developed numerous laws describing the relationship between macroscopic properties of gases like pressure, volume, and temperature. For example, Boyle discovered that volume and pressure were inversely proportional given a fixed temperature. Scientists have also long known that volume and temperature were directly related at a fixed pressure, and that temperature and pressure are inversely related at a fixed volume; these facts are now known respectively as Charles' and Gay-Lussac's laws.

The kinetic theory of gases unifies such gas laws in three ways. First, by introducing quantities for the number of molecules in a sample of gas, the kinetic

[22] *Ibid*, pp. 80.

theory permits all gas laws to be expressed in a single *mathematical* expression, mainly, the ideal gas law $PV = nrT$.[23] That is, in the absence of the variable n, which represents the number of moles in a sample of an ideal gas, the relationship between pressure and volume, for example, can only be expressed by the equation $P = k \cdot \frac{1}{V}$, where k is a constant the depends on a *fixed* sample of gas. If two different samples of nitrogen are added, therefore, the pre-molecular theoretic gas laws (e.g. Boyle's law) do not possess the resources to express the resulting relationship between the pressure and volume of the new sample.

Second, by identifying gases with collections of molecules, the kinetic theory also permits one to identify *all* macroscopic properties of gases, like temperature, with properties of aggregates of molecules, like root mean kinetic energy. Third, under certain assumptions the ideal gas law (and hence Boyle's, Charles'm and Gay-Lussac's laws) is derivable from laws that govern the behavior of molecules.[24] For these three reasons, both the length of the mathematical description and the number of primitive properties of gases are reduced by the kinetic theory. Friedman then identifies theoretical simplicity (he calls it "parsimony") with fewer primitive entities and properties, and he and concludes that unified theories are simpler.[25]

In general, Friedman argues that, in cases in which scientific theories are unified, there is a model $\mathcal{M} = \langle |\mathcal{M}|, R_1^M, \ldots, R_n^M, f_1^M, \ldots, f_k^M \rangle$ of observable phenomena (e.g. the macroscopic property of gases) that is a substructure of a theoretical model $\mathcal{N} = \langle |\mathcal{N}|, R_1^N, \ldots, R_s^N, f_1^N, \ldots f_r^N \rangle$ where $s \geq n, r \geq k$, $|\mathcal{M}| \subseteq |\mathcal{N}|$, and $R_i^N \restriction |\mathcal{N}| = R_i^M$ for all $i \leq n$. Note, Friedman carefully distinguishes between a structure $\mathcal{M}$ being *embeddable* in $\mathcal{N}$ and it being a *substructure.* The former he calls a *representation,* the latter a *reduction.* The distinction will be important when discussing Friedman's argument that unified theories, over time, are more confirmed than disjoint collections of theories. I'll return to this argument in the next chapter.

Friedman's identification of simplicity with unification is both intuitive and plausible. However, I would argue that the mathematical representation of unification in terms of the submodel relation is problematic for at least three

[23] Here, P denotes the pressure of the gas, V its volume, T its temperature, n the number of moles of molecules in the gas, and r a called the universal gas constant. Gases are called *ideal* so long as intermolecular forces, like Van der Waal forces, are negligible in comparison to the effects of the the pressure of the gas.

[24] Two related philosophical issues could be mentioned, but I cannot pursue them here. First, the last two reasons supporting the claim that the ideal gas law unifies previous gas laws suggest that there might be a close relationship between unification and reductionism in science. Hence, one might ask, "are reduced theories always more unified?" Second, attempts at reduction within science, that is, attempts (i) to identify observable, macroscopic phenomena with behavior of microscopic ones (e.g. fundamental particles), and (ii) to derive laws governing macroscopic phenomena from laws governing the corresponding microscopic entities are analogous to logicist attempts in foundations of mathematics (i') to define mathematical concepts with logical ones and (ii') to prove mathematical theorems from only logical axioms and rules of inference. Are the desires for unification in science and logical foundations of mathematics both instances of a more general principle concerning explanation?

[25] Friedman (1983), 335-336

reasons.[26] First, in the cases discussed by Friedman, elements of the observable model (e.g. gases) are simply not elements of the theoretical model (e.g. molecules). Gases are collections of molecules, and so strictly speaking the subset relation is lacking. Similarly, properties of gases (e.g. temperature) are not also properties of molecules (e.g. velocity); they are properties of collections of molecules. Third, the motivation for the use of the submodel relation is to make rigorous the distinction between representations and reductions; that is, Friedman wants to distinguish objects of the theory (e.g. space-time points) from representations of those objects (e.g. a four-tuple of real numbers representing a point of space-time). However, unless Friedman wishes to admit the existence of a continuum number of predicates, the extensions of predicates like "is x meters long" must include both an object and real numbers describing the object. Why? Descriptions of the form "is x meters long" can either be represented as a continuum number of unary predicates $L_x(O)$, each of which keeps the real number x fixed, or they can represented as a binary relation $L(x, O)$ which relates a variable real number x to a variable object O. In the former case, every model of a physical system will have continuum many predicates, and so scientific theories do not appear to be very parsimonious under Friedman's account. In the latter case, however, both concrete physical objects (e.g. tables) and abstract mathematical objects will lie in the extension of the relation $L(x, O)$. Hence, both concrete physical objects and abstract mathematical objects need to be elements of any given model, which destroys the distinction between reduction and representation that Friedman hoped to stress.

Friedman can address all three deficiencies by appropriate use of model theory. For instance, the first two problems can be ratified by defining a reduction to be an appropriate relation between two models, where one structure contains second-order predicates ranging over subsets of the universe of the first. However, I am unsure what role the model theory actually plays in Friedman's arguments; as long as the distinction between reduction and representation is clear, the choice of mathematical formalization is less important.

This concludes a summary of the major attempts, in philosophy, to define the notion of simplicity as it pertains to scientific theories. Having highlighted both the difficulties and strong points of different attempts to define simplicity, it will be clear how the definition of simplicity advocated in the third chapter (namely that of the KGS model) avoids many of the downfalls of the proposals just considered, and capitalizes on their insights.

[26] Morrison (2000) has criticized Friedman's proposal in other ways. At the informal level, I find Friedman's definition of simplicity to be insightful, though his attempts to make it rigorous by employing formal logic are clearly problematic.

Chapter 2

KGS Model

In this chapter, I describe a model of scientific inquiry developed by Kelly, Glymour, and Schulte (henceforth, the *KGS model*). I provide one paradigmatic example of how the model can be used to formally represent interesting problems in scientific inquiry; the example is causal discovery. I then state and explain the *Ockham Efficiency Theorems,* which prove that scientists who heed Ockham's razor retract their opinions less often than do their complexity preferring counterparts.

2.1 Empirical Effects, Problems, and Worlds

In scientific inquiry, theories are often refuted by small, difficult to detect phenomena. Newton's explanation of planetary motion, for example, might be taken to have been refuted by the observation of precession of the perihelion of mercury, which could only be accounted for by the addition of a number of *ad hoc* hypotheses. Such phenomena, like the precession of Mercury's perihelion, are often difficult to observe because of a lack of sensitive experimentation, ingenuity in experiment design, or perhaps because the sample size of observations is too small to compensate for intrinsic randomness in observations. For brevity, I use the word *effect* to refer to such difficult to detect phenomena.

To represent these ideas formally, let E be a non-empty, countable (finite or countably infinite) set whose elements are called *empirical effects*, and define a *problem* to be a set $K \subseteq 2^E$ such that each member of K is finite. A problem is intended to represent those sets of effects that a scientist would observe were he or she to live forever. In this paper, K is a variable that is often held fixed, and thus, reference to K may be dropped to ease notation. A *world* w in K is an infinite sequence in $(2^E)^\omega$ such that

1. $w_n \subseteq w_{n+1}$ for all $n \in \mathbb{N}$
2. There exists an $n \in \mathbb{N}$ such that $w_n \in K$ and $w_m = w_n$ for all $m \geq n$.

In other words, a world is a non-decreasing sequence of sets of effects such that there is some point at which no further effects are observed. The point $n \in \mathbb{N}$ at which a world w "stabilizes" to some set of effects for eternity is called the *modulus of continuity* for w and is denoted $mod(w)$. Informally, a world represents the evidence available to a scientist at discrete points in inquiry.

Let W_K be the set of worlds, and let $w \upharpoonright n$ denote the finite initial segment $\langle w_0, \ldots, w_{n-1} \rangle$ of w. In particular, $w \upharpoonright 0 = \langle \rangle$, the empty sequence. For arbitrary set $R \subseteq E$, let R^∞ denote the infinite sequence that is constantly R. Let:

$$\begin{aligned} W_{K,n} &= \{w \upharpoonright n : w \in W_K\}; \\ W_{K,fin} &= \bigcup_{i \geq 0} W_{K,i}. \end{aligned}$$

Let $e, e' \in W_K \cup W_{K,fin}$, and let $l(e)$ be the length of the sequence e (so that $l(e) \in \mathbb{N} \cup \{\omega\}$). In general, if σ is any sequence of order type ω, we will use $l(\sigma)$ to denote its length. Write $e \leq e'$ if e an initial segment of e' and let $e < e'$ hold just in case e is a proper initial segment of e'. Let $*$ denote sequence concatenation, so that, for example, if $e = \langle \emptyset, \emptyset \rangle$ and $e' = E_0^\infty$ for some $E_0 \in K$, then $e * e' = \langle \emptyset, \emptyset, E_0, E_0, E_0, \ldots \rangle$. Define:

$$\begin{aligned} W_{K,fin,e} &= \{e' \in W_{K,fin} : e \leq e'\}; \\ W_{K,e} &= \{q \in W_K : e \leq q\}. \end{aligned}$$

The set of effects presented along e is:

$$E_e = e_{l(e)}.$$

The restriction of K to effect sets compatible with e is then:

$$K_e = \{K_0 \in K : E_e \subseteq K_0\}.$$

As argued above, scientific theories are often *refuted* by small, difficult to detect effects. But what about *confirmation*? Can one say that a given set of effects $K_0 \in K$ confirms a given theory? In the KGS model, the answer is "no." When many common scientific problems are represented formally in the KGS model, it turns out that (under the most realistic assumptions) any theory might be refuted by some yet-unobserved effect.

However, if a scientist *knew* that no future effects would be observed − in other words, if he knew the world w in which he lived − then there should be a unique scientific theory true of that world. As such, it is reasonable to let each effect set $K_0 \in K$ determine a scientific theory. Formally, let Th be any countable set, and call elements of Th *theories*. To represent the fact that each member of K determines a theory, let $T : K \to \mathsf{Th}$ be any function, and for ease of notation, define $T_{K_0} := T(K_0)$ for all $K_0 \in K$. For any sequence $e \in W_K \cup W_{K,fin}$ such that $e_{l(e)} \in K$, define $T_e \in \mathsf{Th}$ to be the theory $T(e_{l(e)})$.

If $w \in W_K$ is a world, then say T_w is the *theory true of w.* Finally, define a *problem* to be a triple $\langle K, \mathsf{Th}, T : K \to \mathsf{Th} \rangle$.[1]

It will be helpful to discuss some examples of effects, theories, and worlds as described above.

2.1.1 An Example - Causal Discovery

One central task in scientific research is to discover causes. Medical researchers, for instance, might ask, "what are the causes of heart disease?" Psychologists and sociologists might be interested in the causes of suicide, and economists often study the causes of a recession. How might one formally represent causal discovery problems like these in the KGS model?

In recent years, a number of philosophers, computer scientists, and statisticians have argued that one can discover causal relationships amongst variables of interest by studying the *conditional probabilistic dependencies* that hold amongst said variables; such probabilistic relationships are often visually represented in *directed acyclic graphs* (DAGs) like the one below.[2] For example, suppose that smoking causes increased tar in the lungs, and further, suppose that tar causes cancerous cell growth in the lungs. Finally, suppose that smoking only causes cancer by increasing tar, so that there is no other mechanism by which it raises the incidence of cancer. One might represent this graphically in the DAG below. Of course, given this causal knowledge, one should expect that smokers will have higher incidence of lung cancer so that

$$p(\text{lung cancer} = \text{`Yes'} | \text{ Smoker '} = \text{`Yes'}) > p(\text{lung cancer} = \text{`Yes'} | \text{ Smoker '} = \text{`No'}).$$

Importantly, if one *knew* that amount of tar in John's lungs, then one would not need to know whether or not John smokes in order to predict his chances of lung cancer: because the only mechanism by which smoking causes cancer is through increasing tar in the lungs, the knowledge of his smoking habits is irrelevant once one knows how much tar is in her lungs. In this case, the probabilistic dependence between smoking and lung cancer disappears when one conditions on tar in the lungs. In other words,

$$p(\text{lung cancer} = \text{`Yes'} | \text{ `Smoker'} = \text{`Yes'} \,\&\text{`Tar in Lungs'} = \text{`Yes'}) = p(\text{lung cancer} = \text{`Yes'} | \text{`Tar in Lungs'}).$$

The above example is intended to suggest that causal relationships leave a "footprint" in the probabilistic dependencies amongst variables. Yet in the social sciences and medical research, variables of interest are frequently very weakly correlated, and as a result, statistical dependencies are often incapable of being detected without sensitive instrumentation, large sample sizes, or ingenious experimental design. Moreover, the most subtle and difficult to detect dependencies may be of critical importance to discovering the true causal theory

[1] This definition of problems differs from that in Kelly's papers, as it is slightly more general.

[2] See Spirtes, et. al. (2000) for a discussion of the use of DAGs in representation of causal relationships amongst variables. Appendix 2 contains all mathematical details necessary to understand this paper.

governing the variables under investigation. Therefore, it makes sense to define an effect in causal discovery to be a conditional dependence statement of the form:

$$v_1 \text{ is conditionally dependent on } v_2 \text{ given some set of variables } V_0$$

Keeping these ideas in mind, I now formalize how one might represent the problem of causal discovery in the KGS model. Let V be a finite or infinite set of random variables on a common probability space $\langle \Omega, \mathcal{F}, p \rangle$.[3] Let $\textsc{dag}_V$ be the set of directed acyclic graphs with vertices in V. For any two variables $v_1, v_2 \in V$ and any subset of variables $V_0 \subseteq V \setminus \{v_1, v_2\}$, I use the following abbreviation for the claim "v_1 is **not** probabilistically dependent on v_2 given V_0":

$$D(v_1, v_2 \mid V_0)$$

Let $\textsc{cdc}_V$ be the set of all such assertions over a set of variables V; here $\textsc{cdc}$ stands for "conditional *dependence* constraint." If G is a DAG, let $D_G \subseteq \textsc{cdc}_V$ be the set of conditional dependence constraints implied G. Then, to represent causal inference in the KGS model, one lets:

$$\begin{aligned} E_V &= \textsc{cdc}_V \\ K_V &= \{D_G \subseteq \textsc{cdc} \ : \ G \text{ is a DAG } G \text{ over } V\} \end{aligned}$$

Here, elements of K_V are bijective correspondence with Markov Equivalence classes or patterns over V, where patterns are simply the graphical representations of Markov equivalence classes. Importantly, note that, although the set of variables V may itself be infinite, because each element of K_V is finite by stipulation, it follows that there are only finitely many edges in the pattern corresponding to any element $K_0 \in K_V$. Given the above definition of K_V, worlds $w \in W_{K_V}$ are sequences of non-decreasing sets of conditional dependence statements, which are learned as weaker and weaker probabilistic dependencies between the variables are discovered.

Now one may be interested in any number of different questions concerning the variables under investigation. For instance, one might be interested in discovering all of probabilistic dependencies that held amongst the variables V. In this case, theories are simply elements of K, so that $\mathsf{Th}_V = K_V$, and so one is interested solving the problem $\langle K, \mathsf{Th}_V, id : K \to \mathsf{Th}_V \rangle$, where id is the identity map.

On the other hand, one might only be interested in determining whether v causes v' or vice versa, where $v, v' \in V$ are two variables of special interest. In this case, partition K_V into four classes $K_{v \to v'}, K_{v' \to v}, K_{v - v'}, K_{v \not\to v'} \subseteq K_V$ where $K_{v \to v'}$ consists of sets of conditional dependence constraints that imply v causes v' (i.e. there is a directed edge from v to v' in the pattern corresponding to K), $K_{v' - v}$ consists of sets conditional dependence constraints from which one cannot detect whether v causes v' or vice versa (i.e. there is an undirected edge from v to v' in the pattern corresponding to K), and $K_{v \not\to v'}$ consists of

[3] See Appendix 1 for definitions of random variable and probability space.

sets conditional dependence constraints such that neither v causes v' nor vice versa (i.e. there is no edge from v to v' in the pattern corresponding to K). Accordingly, define:

$$\mathsf{Th}_{v,v'} = \{K_{v \to v'}, K_{v' \to v}, K_{v - v'}, K_{v \not\to v'}\}$$

Then one is interested in solving the problem $\langle K, \mathsf{Th}_{v,v'}, T_{v,v'} : K \to \mathsf{Th}_{v,v'} \rangle$ where $T_{v,v'}$ is the function that takes some set $K_0 \in K$ to the element of the partition $\mathsf{Th}_{v,v'}$ to which it belongs. For brevity, in the future I will refer to the first problem as the *full causal discovery problem* and the latter as the *edge discovery problem.*

2.2 Simplicity and Ockham's Razor

Given the above examples, one can now discuss the problem of induction in the KGS model. In the KGS model, the problem of induction amounts to this:

> For any world $w \in W$ and any finite point in inquiry $w \upharpoonright n$, there are effect sets $K_0, K_1 \in K$ such that $w \upharpoonright n \subseteq K_0, K_1$ and $T_{K_0} \neq T_{K_1}$.

In other words, the problem of induction is that, for many real-world problems, at any point of scientific inquiry, there are distinct theories compatible with all available evidence. Hence, the central task of a methodologist is to prescribe how, given available evidence, one ought to choose from competing theories that are compatible with current observations. Of course, scientists are not always proponents of a single theory, as they may think that the available evidence is insufficient to warrant believing one theory rather than another. Accordingly, let $\mathsf{Ans} = \mathsf{Th} \cup \{`?'\}$ be the set of *answers* a scientist may produce during inquiry; here '?' represents refusal to commit to a single theory.[4] Say that a *method* is any function of the form $M : W_{K,fin} \to \mathsf{Ans}$, so that a method M produces an answer given some finite amount of evidence. The purpose of this paper is to present a novel argument for Ockham's razor, and hence, to assess the merits and demerits of methods that select only the *simplest* theories compatible with available evidence. To do so, however, one first needs a definition of simplicity and complexity.

Recall, elements of the set E are intended to represent subtle, difficult to detect empirical effects, such as weak correlations between random variables (as discussed above). Thus, it makes sense to characterize the complexity of a given world and theory, roughly, by the number of effects it postulates *in addition to*

[4] Teddy Seidenfeld has remarked that '?' fails to capture the wide variety of ways in which a scientist might be hesitant to commit to a single theory. For example, one might believe that either T_1 or T_2 is true, but think that T_3 is distinctly false. Alternatively, one might think T_1 or T_3 is true, but think that T_2 is false. Ideally, a model of scientific inquiry ought to be able to capture the distinction between these two ways of failing to commit to a single theory rather than using a single symbol, '?', to represent both. I agree with these suggestions wholeheartedly; Kevin Kelly and Hanti Lin have made promising steps in generalizing the model presented here so that Ans consists of arbitrary, finite disjunctions of theories, as for instance, Levi's model permits.

those already observed. Accordingly, for any problem $\langle K, \mathsf{Th}, T : K \rightarrow \mathsf{Th}\rangle$ and for any $e \in W_{K,fin}$, define a *path* to be a finite sequence $\langle K_j \rangle_{j \neq n}$ (where $n \in \mathbb{N}$) of elements of K_e such that for all $j < n$:

$$\begin{array}{rcl} K_j & \subset & K_{j+1} \\ T_{K_j} & \neq & T_{K_{j+1}} \end{array}$$

In particular, because the entire path lies in K_e, notice that $E_e \subseteq K_0$. For each $K_0 \in K_e$, let $\mathsf{path}(S|e)$ be the set of paths in K_e terminating in S. Abusing terminology, define *a path to a theory* $T \in \mathsf{Th}$ given e to be a path $\langle K_j \rangle_{j \neq n}$ in K_e such that $T = T_{K_n}$. Given a set of effects $E_0 \subseteq E$, define:

$$\mathsf{Ock}_K(E_0) := \{\pi \in \mathsf{path}(S|e) : l(\pi) - 1 = 0\}$$

where, as always, $l(\pi)$ is the length of the path. The above definition might be rephrased more perspicuously as follows. The set $\mathsf{Ock}_K(E_0)$ contains all paths of length 1 starting with E_0 such that there is no proper subpath. Say a theory $T \in \mathsf{Th}$ is *simplest* relative to e if there exists $K_0 \in \mathsf{Ock}_K(E_e)$ such that $T = T_{K_0}$. Define the *set of Ockham answers given* e as follows:

$$\mathsf{Ock}_e = \{T \in \mathsf{Th} \ : \ T \text{ is simplest relative to } e\} \cup \{\text{‘?’}\}$$

Finally, define an *Ockham method* $M : W_{K,fin} \rightarrow \mathsf{Ans}$ to be a function such that $M(e) \in \mathsf{Ock}_e$ for all $e \in W_{K,fin}$.

According to the above definition, there are still many Ockham methods with pathological behavior. For example, one Ockham method may choose the simplest theory compatible with the data (whenever one exists) at every even stage of inquiry, and return ‘?’ at every odd stage of inquiry. There is a more useful class of methods in the KGS model called *normal Ockham methods* that have the following three properties:

1. Every normal Ockham method is an Ockham method simpliciter.

2. If M is normally Ockham, then it is *eventually informative,* which means that for all $w \in W_K$, there is some $n \in \mathbb{N}$ such that $M(w \upharpoonright n') \neq$ ‘?’ for all $n' \geq n$.

3. If M is normally Ockham, then it is *stalwart*, which means that for all $w \in W_K$ and $n \in \mathbb{N}$, if T is simplest at $w \upharpoonright n$ and $w \upharpoonright n+1$, and moreover, if $M(w \upharpoonright n) = T$, then $M(w \upharpoonright n+1) = T$

In other words, normally Ockham methods are those that (i) never return a theory other than that which is simplest with respect to available evidence, (ii) eventually stop returning ‘?’, and finally, (iii) having produced some theory T, continue to return T until it fails to be simplest with respect to available evidence.

One final series of definitions will be helpful before providing examples. Given the above definition of path, one can characterize three types of problems that will be of special interest in this thesis. Say a problem $\langle K, \mathsf{Th}, T : K \rightarrow \mathsf{Th}\rangle$

- *has no short paths* if for every $e \in W_{K,fin}$ and for all $K_0 \in \mathsf{Ock}_K(E_e)$, there exists a path of maximal length in K_e beginning with K_0.
- *has an infinite path* if there exists a path of infinite length $K_0 \subset K_1 \ldots$, in K.
- is *bounded nowhere* if and only if for each $e \in W_{K,fin}$ and each $K_0 \in K_e$, there exists $K_1 \in K_e$ such that $K_0 \subset K_1$ and $T_{K_0} \neq T_{K_1}$

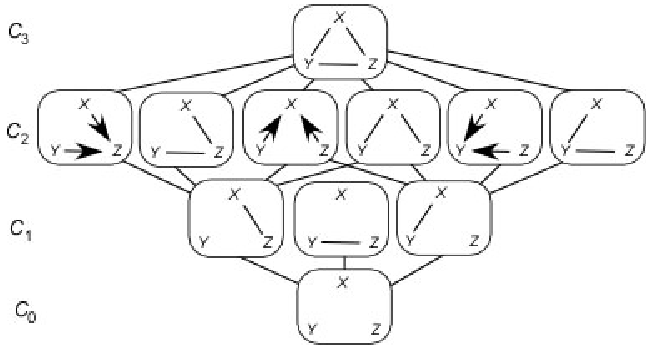
Roughly, a problem is bounded nowhere if for every $e \in W_{K,fin}$, all paths in K_e can be extended infinitely. If $\langle K, \mathsf{Th}, T : K \to \mathsf{Th}\rangle$ is bounded nowhere, then it has an infinite path and it has no short paths. In general, the converse of both assertions is false.

The next section shows that this definition of simplicity captures intuitive definitions.

2.2.1 Methods and Simplicity in Causal Discovery

Although small correlations may be extremely difficult to detect, scientists and policy makers cannot wait indefinitely to form causal hypotheses. As a result, scientists must choose amongst competing causal theories as they detect more and more subtle statistical dependencies amongst the variables under study. Hence, a method a $M : W_{K,fin} \to \mathsf{Ans}$ returns causal theory given a particular set of observed statistical dependencies, where Ans might be either $\mathsf{Th}_V \cup \{\text{`?'}\}$ or $\mathsf{Th}_{v,v'} \cup \{\text{`?'}\}$ depending on whether one is interested in solving the full causal discovery problem or the edge discovery problem.

Intuitively, simple causal theories postulate few of these weak statistical dependencies, and extremely complex ones imply that many subtle, yet-unobserved statistical dependencies hold between the variables under investigation. Hence, the definition of simplicity in terms of paths in the previous section captures important intuitions about the relative complexity of differing causal graphs. When there are three variables under investigation, the figure below depicts how differing DAGs are characterized in terms of complexity.



Interestingly, the causal discovery problems of standard interest are of the form discussed in the previous section.

Theorem 2.2.1. When V is finite, then both the full causal discovery problem $\langle K, \mathsf{Th}_V, T_V : K \rightarrow \mathsf{Th}_V \rangle$ has no short paths.

Proof: The proof follows from Chickering's Theorem (See Appendix 2), but is omitted for ease of exposition.

□

Theorem 2.2.2. When V is infinite, then both the full causal discovery problem and the edge discovery problem are bounded nowhere. Consequently, both problems have no short paths as well.

Proof: First, consider the full discovery problem. Given any $e \in W_{K,fin}$ and any set $K_0 \in \mathsf{Ock}_{K_V,e}$, there is an infinite path in K_V beginning with K_0, as one can simply add more and more conditional dependencies to produce patterns with more and more edges. For instance, as K_0 is finite, there are infinitely many pairs of variables $\{\{v_n, v'_n\}\}_{n \in \mathbb{N}}$ such that no conditional dependence constraints of the form $D(v_n, v'_n | V_0)$ are members of K_0 (where $V_0 \subseteq V \setminus \{v, v'\}$). Define:

$$C(v, v') = \{D(v, v'|V_0) \in \text{CDC}_V \ : \ V_0 \subseteq V \setminus \{v, v'\}\}$$

Then because there is an edge between two variables v and v' in some Bayes net if and only if the two variables are probabilistically dependent given any subset of the variables, the set $K_0 \cup C(v, v')$ is a set of conditional dependence constraints for some set of DAGs. One can define an infinite path $\langle K_n \rangle_{n \in \mathbb{N}}$ in K_V by recursion as follows:

$$\begin{aligned} K_0 &= K_0 \\ K_{n+1} &= K_n \cup C(v, v') \end{aligned}$$

In the case of the edge discovery problem, for any $e \in W_{K,fin}$ and any set $K_0 \in \mathsf{Ock}_{K_V,e}$, there is an infinite path in K_V beginning with K_0 by the proof of Proposition 16 in Kelly and Mayo-Wilson (2007).

□

2.2.2 A Brief Discussion of the KGS definition of simplicity

Recall in the first chapter, I outlined three criteria that any account of theoretical simplicity ought to satisfy:

1. **Clarity:** Any definition of theoretical simplicity ought to allow one to determine, in some class of scientific problems or historical case studies, precisely which theories (if any) are simpler than others and which theories (if any) are incomparable in terms of complexity. Otherwise, such a definition of simplicity is not sufficiently precise to allow one to evaluate the use of simpler theories in practice.

2. **Relevance to Scientific Practice:** Any definition of theoretical simplicity ought to agree with intuitive assessments (i.e. those of some practicing scientists) of simplicity in some class of scientific problems. Otherwise, definitions of "theoretical simplicity" are merely explications of some concept other than what is called simplicity in science.

3. **Data Driven:** The simplicity or complexity of a theory ought to be a function of what phenomena and/or data have been observed. Otherwise, theories that are refuted by existing evidence might still be deemed "simplest," and the question of whether simplicity is an indicator of truth is settled. On the other hand, any definition of simplicity ought not identify "simple" with "more probable" or "better confirmed." Otherwise, the definition of simplicity renders trivial the question of the relation between simplicity, truth, probability, and confirmation.

I argued that previous attempts to analyze theoretical simplicity failed to meet at least on of the above three criteria. Given the discussion in the last few sections it should now be clear why the definition of simplicity within the KGS model meets all three. With respect to the clarity criterion, the KGS model provides an unambiguous axiomatic definition of simplicity.[5] Hence, "simplicity" is whatever satisfies said axioms for a given problem, and thus, given a particular question, one can always construct a partial ordering on theories in terms of their degree of complexity.

The KGS definition of simplicity meets the second criterion in that it provides a reasonable definition of simplicity in causal discovery and curve-fitting, two central problems in all areas of scientific inquiry. Finally, as the KGS model only defines simplest given a set of empirical effects E, it satisfies the first half of the data-driven criterion by not allowing simplest theories to have been refuted. Simultaneously, as theories are not assigned probabilities in the KGS model, and hence, simpler theories cannot be more probable, it meets the second half of the data-driven criterion in that it does not beg the question as to why it is rational to prefer simpler theories.

Importantly, the reader should note that the KGS definition of simplicity agrees extensionally with many other analyses of simplicity in a number of problems. Like those who argue simplicity is a matter of minimizing free parameters, or any number of other measures of syntactic complexity, the KGS analysis of simplicity in curve-fitting identifies simplicity with lower degree polynomials, and complexity with higher-degrees. Like those who identified simplicity with falsifiability, the KGS analysis of simplicity in curve-fitting and causal inference identifies simplicity with models that are more easily rejected at lower sample sizes (and hence, more falsifiable), and identifies complexity with models that might take enormous amounts of data to refute; for example, in curve-fitting, it is considerably more difficult to rule out the possibility that *some* cubic polynomial explains available data than it is to rule out that *some* line explains the

[5] To be fair, the above definition of simplicity in terms of "empirical effects" is not axiomatic, but it is a special case of an axiomatic definition of simplicity available in Kelly (2009b).

data. Finally, like those who identify simplicity with unifying power, the KGS analysis of simplicity in causal inference implies that it is often simpler to suppose that there is one cause of many different phenomena when many variables are correlated.

2.3 Costs of Inquiry and the Efficiency Theorem

In the KGS model, there are three virtues of methods. The first and most important virtue is *convergence*; a method $M : W_{fin} \to \mathsf{Ans}$ is said to be *convergent* provided that for all $w \in W$ there exists an $n_w \in \mathbb{N}$ such that $M(w \upharpoonright m) = T_w$ for all $m \geq n_w$. In other words, a method M is convergent if M eventually (i.e. at some point of inquiry) produces the true theory in every world w, and continues (from that point of inquiry) to produce the true theory forever.[6]

A second virtue of a method M is that it minimizes errors. Informally, say a method M errs in some world w whenever it produces a theory other than T_w, the theory true of w. More precisely, say a method M *errs in w at stage n* if $M(w \upharpoonright n) \in \mathsf{Th} \setminus \{T_w\}$. Notice, that '?' does not constitute an error. This allows one to define a function $\epsilon : \mathsf{Ans}^{W_{K,fin}} \times W_K \times \mathbb{N} \to \{0, 1\}$ such that:

$$\epsilon(M, w, n) = \begin{cases} 1 \text{ if errs at } w \text{ at stage } n \\ 0 \text{ otherwise} \end{cases}$$

Let $\epsilon(M, w) = \sum_{n \in \mathbb{N}} \epsilon(M, w, n)$ be the total number of errors committed by M in w. Favoring methods that converge to the true theory and minimize total errors is fairly uncontroversial. Most philosophers, working scientists, and statisticians would recognize the importance of eventually finding true theories (or at least approximating such theories), and of minimizing errors whenever possible. Of course, some philosophers might think that there are other virtues of scientific theories that, in some circumstances, are more important than error minimization and convergence.[7]

However, there is third, less standard virtue of methods in the KGS model that is crucial for proving the Efficiency Theorems: minimization of "retractions" or "mind-changes." A scientist is said to *retract* his opinion at time n if he advocates some theory T at time $n - 1$, and some other answer (whether it be another theory or '?') at time n. In symbols, say M *retracts in w at stage n* (where $n \geq 2$) if $M(w \upharpoonright n) \neq M(w \upharpoonright (n - 1))$ and $M(w \upharpoonright (n - 1)) \neq$ '?'. Define:

$$\rho(M, w, n) = \begin{cases} 1 \text{ if M retracts in } w \text{ at stage } n \\ 0 \text{ otherwise} \end{cases}$$

[6] Importantly, the notion of convergence in the KGS model is a minor variant of the more general concept of *statistical consistency* when one (i) considers Th as a metric space with the discrete metric, (ii) considers initial segments of some world w as (deterministic) samples of increasing size given the unknown world w.

[7] See Sober (1985) for a discussion of why statistical consistency, for instance, need not be considered a paramount virtue of estimators.

Hence, $\rho(M, w) := \sum_{n \in \mathbb{N}} \rho(M, w, n)$ is the total number of retractions committed by some method M in the world w.

Penalizing a scientist for retracting his previous opinions might seem, on first glance, to reward dogmatism and penalize open-mindedness. Yet abandoning an old theory in favor of a new one often comes with heavy costs, both epistemically and pragmatically. In retracting previously held beliefs, a scientist must learn all of the theoretical consequences and commitments of the newly advocated theory and unlearn those of the old.[8] One must learn new vocabulary and explanations of familiar phenomena; one must often rework mathematical calculations in a new framework, repeat computer simulations, and so on.

In particular areas of scientific inquiry, retracting previously held scientific theories can also lead to substantial practical costs. Consider causal inference, and imagine a federal education policy, for example, that needed to be abandoned in light of a new causal theory concerning the efficacy of a particular type of mathematics curriculum. Retraining instructors, revising textbooks, and so on is financially expensive and administratively burdensome. Hence, there are significant reasons to avoid retracting previously held scientific theories.

Importantly, one should note that retracting a previously endorsed theory can be costly *even if one is abandoning a false theory in favor of a true one.* The epistemic costs discussed above (e.g. re-calculating the theoretical consequences of one's theory, learning new vocabulary, etc.) are the result of *changing* one's mind and occur independently of whether the newly advocated theory is true or not.

Before developing the KGS model in any greater detail, I want to anticipate an important objection. One might object that retractions are not so costly (either epistemically or pragmatically) in every area of scientific inquiry as they might be in the example involving development of educational policy discussed above. I agree with this so-called objection. Recall the purpose of the KGS model is to (i) discover those costs that would make a systematic preference for simpler theories rational, and (ii) to explain under what circumstances such costs are reasonably part of scientists' concerns and goals in a particular area of inquiry. If one is unconcerned with truth or avoiding error, or if retracting previously advocated theories is particularly cheap, then the arguments below provide little reason to heed Ockham's razor. The force of the Efficiency Theorem as an argument for Ockham's razor lies in the fact that, in more areas of scientific inquiry than not, retractions, errors, and/or failures to find true theories are far more costly than the alternatives.

Given these definitions, one can define the cost vector $\lambda(M, w)$ of the method M in the world w to be the ordered pair $\langle \epsilon(M, w), \rho(M, w) \rangle$.[9] Define a partial

[8] Strictly speaking, according to the above definition, a scientist may not advocate a new theory at the time he retracts an old one. This is unimportant, as the costs in KGS model require that a scientist eventually advocate another theory, lest he fail to converge to the truth in some world.

[9] In previous papers by Kevin Kelly, as well as one jointly written with the author, cost vectors contained an optional component for the "times" at which particular retractions occur. As evaluating the times at which various methods retract is no necessary for the proofs of the

ordering on cost vectors as follows:

$$\lambda(M,w) \preceq \lambda(M',w) \text{ if and only if } \epsilon(M,w) \leq \epsilon(M',w) \text{ and } \rho(M,w) \leq \rho(M',w)$$

Further, say $\lambda(M,w) \prec \lambda(M',w)$ if and only if $\lambda(M,w) \preceq \lambda(M',w)$ and either (a) $\epsilon(M,w) < \epsilon(M',w)$ or (b) $\rho(M,w) < r(M',w)$ (or both).

In the absence of well-defined probabilities on worlds, a scientist interested in minimizing errors and retractions might wish to pick methods that minimize the number of errors and retractions he commits in the worst-case. Accordingly, define the *worst case cost bound over* a set of worlds $W_0 \subseteq W$ to be $\lambda(M, W_0) := \sup\{\lambda(M,w) : w \in W_0\}$. One can then compare worst-case cost bounds over a set of worlds W_0 in the same way one compared cost bounds over a single world. Namely, employing M is held to be *at least as desirable* as employing M' in W_0 if M commits no more retractions or errors in the worst case in W_0; employing M is *strictly preferred* to employing M' in W_0 if M is at least as desirable, and moreover, M commits either strictly fewer errors or strictly fewer retractions in the worst-case.

Above, the supremum is defined coordinate-wise, so that $\sup\{\lambda(M,w) : w \in W_0\} := \langle \sup\{\epsilon(M,w) : w \in W_0\}, \sup\{\rho(M,w) : w \in W_0\}\rangle$. This supremum is always defined as $\langle \omega, \omega \rangle$ bounds all possible cost bounds.

Proposition 2.3.1. If M is convergent, then the worst case cost bound $\lambda(M, W)$ over the set of all worlds is $\langle \omega, \omega \rangle$.

By the above proposition, all convergent methods have the same worst case cost bounds over the set of *all* worlds. So we cannot expect Ockham's razor to minimize worst-case retractions over all worlds. A similar situation arises in computer science. In general, there are no finite bounds on the amount of time and space an algorithm might use, as input sizes are unbounded. For example, if it takes n^2 seconds to multiply two n-digit numbers, then we can force a computer to take arbitrarily long to multiply by feeding it larger and larger numbers. To analyze complexity, then, computer scientists compare the worst case performance of algorithms on *inputs of a fixed size.* The KGS model pursues similar approach, by comparing the performance of methods on worlds of *identical empirical complexity.* How is complexity defined? For all $e \in W_{fin}$, define the n^{th} *complexity class* $\mathsf{Comp}_{n,e}$ *relative to* e to be the set of worlds $w \in W_{K,e}$ such that there exists (i) a path $\langle K_j \rangle_{0 \leq j \leq n}$ of length $n+1$ in K_e and (ii) $m_0 < m_1 \ldots < m_n \in \mathbb{N}$, where $m_0 > l(e)$, such that$w_{m_j} = K_j$ for all $j \leq n$. In other words, complexity is, roughly, a function of the number of subtle empirical effects predicted by a theory.

Now it is possible to compare two methods' M and M' performances by comparing their worst case bounds on complexity classes. Let $\lambda^e_n(M) = \lambda(M, \mathsf{Comp}_{n,e})$, and define a partial ordering on methods as follows:

$$M \preceq_e M' \text{ if and only if for all } n \in \mathbb{N}, \text{ we have } \lambda^e_n(M) \preceq \lambda^e_n(M')$$

Efficiency theorems, they are ignored for the remainder of this paper. That is not to say that one should not care about when a method retracts, but rather, that the exposition in this thesis is greatly condensed by avoiding adding more dimensions to the cost-vectors above.

Define $M \prec'_e M$ to hold if and only if $\lambda^e_n(M) \preceq \lambda^e_n(M')$ for all $n \in \mathbb{N}$ and there exists an n_0 such that $\lambda^e_{n_0}(M) \prec \lambda^e_{n_0}(M')$. That is, $M \prec'_e M$ if M performs at least as well (in terms of errors and retractions) as M' on every complexity class (with respect to e), and performs strictly better in at least one complexity class. Of course, one may wish to know whether M does strictly better than M' in *every* complexity class, and for this, we introduce one more abbreviation:

$$M << _e M' \text{ if and only if for all } \forall n \in \mathbb{N}(\lambda^e_n(M) \prec \lambda^e_n(M'))$$

Deciding whether to switch from a method M to M' after acquiring some evidence $e \in W_{K,fin}$ requires evaluating whether M or M' accrues more errors or retractions in the future. But in such cases, analyzing the cumulative costs of each method is misleading: M may have fewer total costs but accrue more costs than M' in the future. Hence, if we treat past costs as fixed, one can only evaluate methods that have identical costs in the past. One easy way to do so is to compare methods that have identical behavior up to the point of inquiry e in question. Accordingly, if M and M' are methods, write $M \approx_e M'$ if and only if $M(e \upharpoonright j) = M'(e \upharpoonright j)$ for all $j < n$. Now we can state the first efficiency theorem.

Theorem 2.3.1 (Weak Efficiency Theorem)**.** Let $e \in W_{K,fin}$. If M is normally Ockham (i.e Ockham, stalwart, and eventually informative), then $M \preceq_e M'$ for all convergent methods M' such that $M \approx_e M'$.

Of course, the above theorem only proves that Ockham methods perform no worse than do other methods. Yet a stronger theorem is true. Assuming that the problem K has not short paths, an Ockham method M performs *strictly* better (in the worst-case) than does any non-Ockham method M' from *any* point of inquiry at which the M' fails to be Ockham onward. More formally, say M *violates Ockham's razor at* e if $M(e) \notin \mathsf{Ock}_e$. With these definitions, one can now state the strong efficiency theorem which is proven in Kelly (2007) and in more general form in Kelly and Mayo-Wilson (2008).

Theorem 2.3.2 (Strong Efficiency Theorem)**.** Suppose M is a normally Ockham method (i.e M is Ockham, stalwart, and eventually informative). Further, assume that M' is a convergent method. Let $e \in W_{fin}$, and suppose that M' violates Ockham's razor time at e. Then $M << _e M'$.

It follows immediately from Theorems 2.2.1 and 2.2.2 that Ockham methods are most efficient in the full causal discovery problem when V is finite, and in the full causal and edge discovery problems when V is infinite.

Chapter 3

Randomized Strategies and Ockham's Razor

In the last chapter, I sketched the KGS model of inquiry, and described the Efficiency Theorems, which prove that a systematic preference for simpler theories is *truth-tracking* in that scientists who employ normal Ockham methods make fewer errors and retract their opinions less often (in the worst case) than those who do not. The Efficiency Theorems, however, are limited in two primary ways. First, effects are assumed to be *unambiguous* in the sense that they are not subject to measurement error or chance deviation from their true value. Second, in the KGS model, a scientist must *deterministically* choose theories given available data; she is not permitted to use any sort of randomizing device.

In this chapter, I generalize the KGS model to consider *randomized* methods for inferring theories from data, thereby addressing the second limitation. To do so, I represent the KGS model as a strategic game, so that randomized methods are formally represented by *mixed strategies* in the game. In the game-theoretic representation of the KGS model, there are two players, the scientist and Nature, whose interests are strictly competitive. That is, any gain for the scientist is a loss for Nature and vice versa. In the upcoming sections, I will rigorously define the actions available to the scientist and Nature, what their preferences are, and so on. However, for now, it is important to explain the purpose of employing game theory at all, and how one should interpret a game in which "Nature" is a competitor.

First, and most importantly, representing the KGS model as a game does not commit one to the (untenable) assumption that there is an intelligent agent called "Nature" whose purpose is to wreak havoc on scientific inquiry. Rather, one can imagine a mixed strategy for Nature as a scientist's "prior" probability on possible states of world (so that a pure strategy for Nature represents the scientist's belief that a particular state is true with probability 1). The game-theoretic concepts of "best response" and "Nash equilibrium," then, allow one to investigate what types of methods a scientist would regard as rational given various candidate prior probabilities on states of the world.

Second, representing the KGS model game theoretically suggests a deep extension of Kelly's efficiency theorem. Although a full proof is not yet available, the results of this chapter suggest the following conjecture:

Conjecture: When the KGS model is appropriately represented as a two-person, strategic form game, then the only Nash equilibria are those in which the scientist plays a mixture of Ockham methods.

If the conjecture is true, then one can conclude that, even amongst *randomized* methods for selecting theories from data, a systematic preference for simpler theories is still rational (in the sense that a scientist whose randomized method favors simpler theories will minimize errors and retractions during inquiry). Moreover, the conjecture would imply that a systematic preference for simpler theories is rational *even when one believes the truth is arbitrarily complex with high probability.* Because, as I have argued, various Bayesian defenses of Ockham's razor beg the question by assuming the truth to be simple with high probability, the arguments I present provide a novel defense of Ockham's razor.

The structure of this chapter is as follows. In the first section, I review basic game theory. The section can be skipped by anyone with familiarity with the subject. Next, I analyze a class of games, which I call *quasi-games,* which relax the standard game-theoretic assumption that players' preferences are linearly ordered. I then prove a generalized version of Nash's theorem for games in which player's preferences are quasi-orders with a particular structure. In the second section, I represent the KGS model as a quasi-game in three different ways, and I analyze the equilibria in each of the three representations.

3.1 Strategic Games

In game theory, it is standard to distinguish between *strategic games* and *extensive-form games.* In the former, a finite number of players each *simultaneously* choose *one* action to perform, and the payoffs to each player are a function of the set of all actions chosen by all of the players. In contrast, in extensive-form games, players make *sequential* decisions, often informed by the actions taken by other players in the past. Moreover, unlike strategic games, the same player may act more than once in an extensive-form game. In this section, I provide three ways of representing the KGS model as a strategic game, and argue the second is most appropriate. I first review some basic game theory.

3.1.1 Nash Equilibria in Finite and Infinite Strategic Games

A *strategic game* G is a triple $G = \langle N, \{A_i\}_{i \in N}, \{\succeq_i\}_{i \in N} \rangle$ where:

1. N is a finite set of *players,* which, for simplicity, will be labeled by natural numbers $\{1, 2, \ldots, N\}$

2. For each $i \in N$, the set A_i is the set of *actions* available to player i

3. For each $i \in N$, the relation $\{\succeq_i\}_{i \in N}$ is a linear, quasi-order on $A = \times_{i \in \mathbb{N}} A_i$. The relation $\{\succeq_i\}_{i \in N}$ is called the *preference relation* of player i.[1]

A game G is called *finite* if A is a finite set. For all $a, a' \in A$, write $a \approx_i a'$ if $a \succeq_i a'$ and $a' \succeq_i a$. Condition (3) is sometimes strengthened so that $a = a'$ whenever $a \approx_i a'$. In other words, players' preference relations are often assumed to be *anti-symmetric.* In such cases, $\succeq_i$ is called a *total order.* Say that Player i *weakly prefers* a to a' if $a \preceq_i a'$, and say that he *strictly prefers* a to a', written $a' \prec_i a$, if $a \preceq_i a'$ but $a \not\succeq_i a'$. For the purposes of modeling scientific inquiry, in upcoming sections we will drop the requirement of linearity in Condition (3), so that each player's preference relation is assumed only to be a *quasi-order* (i.e. not every strategy profile need be comparable under a given player's preference relation).

Many of the fundamental results of elementary game theory, however, no longer hold if player's preferences are not linearly-ordered, so we first review standard game-theoretic results in the presence of the linear ordering assumption. Then we show how they change when the requirements on preference relations are relaxed. Standard examples of games include the following:

Example: Bach Or Stravinsky Suppose that two friends, Manuel and Deborah, wish to attend a classical musical concert tonight with one another, but neither Manuel nor Deborah is able to contact the other before the concert (perhaps Manuel has no cellphone). There are two different concerts taking place: a Bach concert and a Stravinsky performance. Manuel would like to see the Bach concert with Deborah; Deborah would prefer to see the Stravinsky concert with Manuel, and both prefer seeing a concert with the other than going to the symphony alone. Formally, there are two players $N = \{M = Manuel, D = Deborah\}$, each of whom can choose to attend the Bach or Stravinksy concert $A_M = A_D = \{B = Bach, S = Stravinsky\}$, and their respective preference orderings are as follows:

$$(B, B) \succ_M (S, S) \succ_M (B, S) \succ_M (S, B)$$

$$(S, S) \succ_D (B, B) \succ_D (B, S) \succ_D (S, B)$$

Example: Prisoner's Dilemma Suppose that two suspects for a crime, Bonny and Clyde, are taken to separate holding cells to be interrogated. Each suspect is then offered the following deal. If neither suspect confesses to committing

[1] Here, a quasi-order is a reflexive, transitive binary relation. In mathematics, quasi-orders are more frequently called *pre-orders.* An ordering on a set is linear just in case any two elements of the set are comparable in the ordering. Note that one does not assume that player's preferences are anti-symmetric, as there might be two distinct outcomes that are equally desirable. For example, if a decision-maker enjoys coffee as much as he does tea, then "coffee $\preceq$ tea" and "tea $\preceq$ coffee" might both be true, but this does not imply that coffee and tea are the same beverage.

the crime, then both will receive a mandatory 2 years in prison. If one suspect confesses while the other remains silent, then the confessor will receive only a year sentence for his or her cooperativeness and the silent one will receive 10 years in prison. Finally, if both parties confess, then both will receive 3 year sentences, as neither suspect's confession provides information to the police that they would have failed to obtain had only one party decided to remain silent. Both Bonny and Clyde know that the other has been offered the same deal, and each must decide indepedently whether to confess or remain silent without knowledge of the other's choice. Outcomes of the game are represented by ordered pairs from $\{S, C\} \times \{S, C\}$, where (S, C), represents the outcome where Bonny stays silent (S) and Clyde confesses (C). Both Bonny and Clyde prefer to minimize the time they spend in jail, and so their preference relations are as follows:

$$(C, S) \succ_B (S, S) \succ_B (C, C) \succ_B (S, C)$$

$$(S, C) \succ_C (S, S) \succ_C (C, C) \succ_D (C, S)$$

Example: Matching Pennies Suppose Manuel and Deborah decide to play a second game called matching pennies. In this game, each player has two possible moves: "heads" and "tails." If both players choose the same move (i.e. both players play heads, or both players play tails), then Manuel wins a point and Deborah loses a point. Otherwise, Manuel loses a point and Manuel wins a point. Again, the outcome space is a set of ordered pairs where each coordinate is either H or T, representing heads and tails respectively, and Manuel and Deborah's preference relations are as follows:

$$(H, H) \approx_M (T, T) \succ_M (H, T) \approx_M (T, H)$$

$$(T, H) \approx_D (H, T) \succ_D (H, H) \approx_D (T, T)$$

Game theorists argue that, under particular models of rationality, players will utilize particular strategies in the above games. To see why, it will be helpful to introduce some notation. Call an element $a \in A$ a *strategy profile.* For any player i, let $a_{-i} = (a_1, \ldots a_{i-1}, a_{i+1}, \ldots, a_N)$ be the $N-1$-tuple consisting of all coordinates of a except a_i. For any strategy profile $a \in A$ and any $a_i^* \in A_i$, let (a_{-i}, a_i^*) denote the result of replacing the i^{th} coordinate of a with a_i^*. Then say a strategy profile a is a *Nash equilibrium* if $a \succeq_i (a_{-i}, a_i^*)$ for all $i \in N$ and all $a_i^* \in A_i$.

There are several alternative definitions of a Nash equilibrium. For each player $i \in N$, define the set $B(a_{-i})$ of *best responses* given a_{-i} as follows:

$$B(a_{-i}) = \{a_i^* \in A_i \mid (a_{-i}, a_i^*) \succeq_i (a_{-i}, a_i^{**}) \text{ for all } a_i^{**} \in A_i\}$$

Alternatively, define a player's $uB(a_{-i})$ *unbeaten responses* given a_{-i} as follows:

$$uB(a_{-i}) = \{a_i^* \in A_i \mid \neg \exists a_i^{**} \in A_i \text{ such that } (a_{-i}, a_i^{**}) \succ_i (a_{-i}, a_i^*)\}$$

When each player's preference relation is linear, it's clear that $uB(a_{-i}) = B(a_{-i})$, and thus, $a \in A$ is a Nash equilibrium if and only if $a_i \in B(a_{-i})$

for all $i \in N$ if and only if $a_i \in uB(a_{-i})$ for all $i \in N$. When one drops the linearity assumption, however, a player may have unbeaten responses that are not best responses in the sense defined above, and so the equivalence of the above definitions does not hold.

For examples of Nash equilibria, again consider the above three games. By definition, Bach or Stravinsky has two Nash equilibria: (B, B) and (S, S). The prisoner's dilemma has a unique Nash equilibrium, namely, when both players confess. Finally, the matching pennies game has no Nash equilibria. It turns out, however, that if one allows players to employ randomizing devices, one can likewise find Nash equilibria for the matching pennies game. In fact, under very general conditions, any finite game has Nash equilibria when one allows players to employ randomizing devices. These results are reviewed in the next section, and I then discuss their importance to an analysis of the KGS model of inquiry.

3.1.2 Mixed Strategies and Nash's Theorem for Strategic Games

In order to model the use of randomizing devices in games, let $\mathcal{A}_i$ be a σ-algebra over A_i. In most applications, it is assumed that $\mathcal{A}_i$ contains every singleton of the form $\{a_i\}$, where $a_i \in A_i$. Define a *mixed strategy* for player i to be a probability measure p over $\mathcal{A}_i$. When G is a finite game, as in many applications, A_i is a finite set and it is standard to define $\mathcal{A}_i$ to be 2^{A_i} so that each action $a_i \in A_i$ has a well-defined probability. If p is countably additive, then a mixed strategy is called a *countable mixture,* and if p is purely finitely additive, then p is called a *purely finite mixture.* In most standard applications, it is assumed that p is countable mixture.

How does one assess the desirability of playing a mixed strategy? Assume that, for each player $i \in N$, each strategy profile $a \in A$ can be assigned a numerical utility $u_i(a)$.[2] For each player $i \in N$, let $\mathcal{P}(A_i)$ represent the set of probability measures on A_i. Define player i's utility for $p \in \times_{i \in N} \mathcal{P}(A_i)$ as follows:

$$u_i(p) = \sum_{a \in A} [(\Pi_{j \in N}\ p_j(a_j)) \cdot u_i(a)]$$

That is, the utility of a mixed strategy p is the weighted average of the utilities it assigns to each (pure) strategy profile, where the weights are exactly the probabilities of a particular strategy profile being being the outcome of the game. Here, the probability that a strategy profile $a \in A$ is played is $\Pi_{j \in N}\ p_j(a_j)$; this assumption of probabilistic independence is justified by the fact that players choose their actions simultaneously in strategic games.

Why are mixed strategies important? Again, consider the matching pennies game. Suppose Manual employs the strategy $p = (\frac{1}{2}, \frac{1}{2})$, where p represents the mixed strategy in which heads and tails are each played with equal probability of

[2] In decision theory, it is standard to first axiomatize plausible properties of preference relations of rational agents, and then to prove that such axioms imply that one's preferences are representable as numerical utilities. See Savage (1972).

one half. What is Deborah's best response? It turns out that, whatever strategy Deborah employs, she can expect to win half the times she plays matching pennies if Manuel plays p. Suppose that Deborah plays the strategy $q = (q_1, 1 - q_1)$. Deborah wins exactly when Manuel and she play different actions (i.e. one player chooses heads while the other chooses tails). Hence, given that Manuel and Deborah play simultaneously (and thus, their actions are probabilistically independent), then the probability that Deborah wins is exactly:

$$q_1 \cdot \frac{1}{2} + (1 - q_1) \cdot \frac{1}{2} = \frac{1}{2}$$

If Deborah likewise chooses the strategy $p = (.5, .5)$, therefore, then Manual can expect to win exactly half of the games whatever strategy he employs, and so he has no reason to choose a strategy other than p. It follows that the strategy profile (p, p) is a mixed strategy Nash equilibrium of the matching pennies game.

The example of matching-pennies is not unique. In light of the following theorem (due to Nash), extending a game by mixed strategies creates mixed strategy Nash equilibria under very general circumstances.

Theorem 3.1.1 (Nash 1950)**.** Let G be a finite strategic game in which each player's preferences are representable as numerical utilities. Then G has a countably additive mixed strategy Nash equilibrium.

Each assumption of Nash's theorem is necessary to obtain the desired result. For example, infinite games need not have countably additive mixed strategy Nash equilibria. Consider a two player game called, "Pick the biggest natural number." Every natural number represents a possible action for each player, and a player wins in this game if and only if he picks a larger natural number. Clearly, there are no pure strategy Nash equilibria in this game because, for if a player chooses a smaller integer than his opponent, then he would have benefited from choosing a different, larger integer.

Moreover, it can be shown that there there are no countably additive Nash equilibria. An informal argument is given here. Let $p_1(n)$ represent the probability that player one plays the natural number n. Because player one's mixed strategy must be countably additive, it follows that:

$$\sum_{n=0}^{\infty} p_1(n) = 1$$

In other words, for any ϵ, there is some $n_\epsilon \in \mathbb{N}$ such that player one places probability at least $1 - \epsilon$ on all natural numbers before n_ϵ. Player 2, then, can achieve an expected payoff of at least $1 - \epsilon$ so long as he chooses a mixed strategy that assigns probability 1 to numbers greater than n_ϵ. In such a case, Player 1's expected payoff is ϵ. But notice that Player 2's strategy must likewise be countably additive, and hence, there is some $m_\epsilon > n_\epsilon$ such that Player 2 plays numbers less than m_ϵ with probability $1 - \epsilon$. Thus, Player 1 could improve his expected earnings from ϵ to $1 - \epsilon$ by playing an alternative mixed strategy that assigns probability one to numbers greater than m_ϵ. And so on. So there are no countably additive Nash equilibria either.

There is an extensive literature on infinite games, and the mathematical conditions under which such games have equilibria.[3] However, these results standardly assume, like Nash's theorems, that players preferences are representable by numerical utilities, or at the least, player's preferences over outcomes are linearly ordered (i.e. confronted with two outcomes, a player weakly prefers one outcome to the other).

Yet one can also see the failure of Nash's theorem when player's preferences are not required to be linearly ordered. Consider a one-person game in which the sole player has two possible actions, neither of which he prefers to the other. That is, the two possible outcomes of the game are *incomparable* to the player; he is **not** indifferent between the two outcomes. Then, by definition, no action is preferred to all others, and so there are no Nash equilibrium simpliciter in such a game. Both actions, however, are what I call *quasi-Nash* equilibrium, which will be defined more precisely below. More complicated examples can be constructed in multi-player games in which there may exist distinct equilibria, some of which are Nash and some of which are merely quasi-Nash.

In formalizing the KGS model of inquiry, I consider games in which players have both (i) an infinite number of actions and (ii) quasi-ordered preferences. Despite the failure of these two crucial assumptions, which are standardly made in game theory, the games I discuss often have non-trivial Nash equilibria. Before discussing such equlibria, it will be helpful to prove some results about games in which players' preferences over outcomes are not linear, as such games are rarely discussed in game theory.

3.1.3 Quasi-Ordered Preferences in Strategic Games

In this section, I generalize Nash's theorem to consider games in which players desire to maximize a finite set of goods, some of which are comparable, and some of which are not. For example, suppose a player desires to maximize three different goods: wealth, leisure time, and reputation. The imaginary player may prefer, say, wealth to both leisure time and reputation, but she may not have well-defined preferences concerning how leisure time and personal reputation ought to be balanced against or traded for one another. Alternatively, a player may be interested in maximizing different qualities or virtues of a single object. For instance, car buyers often seek cars that are simultaneously reliable, aesthetically-pleasing, and so on.

In either case, a player's preferences may only be *quasi-ordered,* and moreover, they may contain *lexicographical* comparisons, in the sense that some goods or virtues, no matter how small in quantity, are preferred to other goods or virtues, no matter how great the quantity. For example, some car buyers are primarily concerned with reliability, and they only consider aesthetic

[3] See Berge et. al. (1957) Parts IV, and V. See Heath and Sudderth (1972) and Kadane, Schervish, and Seidenfeld (1999) pp. 246-267 for a discussion of finitely-additive equilibria in infinite games.

features of cars to break ties between vehicles that are equally reliable. It is well-known that, in general, there may be no equilibria in games in which players' preferences are lexicographically ordered.[4] However, below I prove that, in several interesting games, equilibria exist even when players preferences contain lexicographically-ordered components. Moreover, I prove that, although relaxing the ordering assumption might seem to be a major revision to the definition of the game, it turns out that when the quasi-orders representing players' preferences are of a particular form, the proof of Nash's theorem remains fundamentally unchanged.

Before embarking on studying games in which players' preferences are only quasi-ordered, it is important to explain why one should be interested in the hybrid lexicographic-pareto preference relations discussed above. Recall that, in the KGS model, a scientist wishes to find a *convergent* method the minimizes errors and retractions during inquiry. Thus, the scientist is interested in maximizing three different virtues of scientific methods, namely, convergence, avoidance of error, and retraction minimization. In the KGS model, convergent methods are preferred to non-convergent ones always, but there is no prescribed way of balancing errors against retractions. Therefore, the costs of inquiry in the KGS model are best represented as one of the preference relations described above, in which some goods and/or values are ranked lexicographically higher than others, while other goods may simply be incomparable.

Define a *strategic quasi-game* G to be a triple $G = \langle N, \{A_i\}_{i \in N}, \{\succeq_i\}_{i \in N}\rangle$, where

1. N is a finite set of *players,* which, for simplicity, will be labeled by natural numbers $\{1, 2, \ldots, N\}$
2. For each $i \in N$, the set A_i is the set of *actions* available to player i
3. For each $i \in N$, player i's preference relation $\{\succeq_i\}_{i \in N}$ is a reflexive, quasi-order on $A = \times_{i \in \mathbb{N}} A_i$.

Notice that only difference between the definition of a strategic game and a quasi-game is Condition (3), as in quasi-games players preference are **not** required to be linearly ordered. Hence, every game is a quasi-game, but not vice versa.

Recall that Nash equilibria exist in finite strategic games in which players preferences are *representable by numerical utilities.* Hence, one should expect Nash's theorem to generalize to quasi-games only when players' preferences are representable in some numerical fashion. In this section, I will focus on a class of quasi-games, which I call *quasi-game for incomparable goods,* in which players preferences are representable by finite *vectors* of real numbers.[5] These vectors will represent quantities of different goods or sources of value available to a given player. For example, if a player wishes to maximize wealth and leisure time,

[4] See Fishburn (1972).

[5] I do not know what axioms a preference ordering must satisfy so that it can be represented in the way I describe below.

then his or her preferences might be represented as ordered pairs $(x, y) \in \mathbb{R}^2$, where x represents some number of dollars, and y represents number of hours. As discussed above, some of the goods the players wish to maximize may be incomparable, as for example, if a player does not know how to compare amounts of wealth and leisure time. Other goods will be lexically ordered, so that, for example, a player who abhors chocolate ice cream may prefer one teaspoon of vanilla ice cream to any arbitrarily large amount of chocolate.

To represent such preference relations, let $\overline{k} = \langle k_1, \ldots, k_n \rangle \in \omega^{<\omega}$ be a finite sequence of natural numbers, and let $l(\overline{k})$ denote its length. Here, a coordinate k_i of $\overline{k}$ represents quantities of a finite number of goods which a player will consider to be incomparable. If $i < j$, then the player prefers all goods in k_i to all those represented in k_j. As in previous sections, for any sequence r, let r_i denote the i^{th} element in the sequence r. So if $r \in \mathbb{R}^k$, then r_i is the i^{th} real number of the vector r (assuming, of course, $i \leq k$).

For any $k \in \mathbb{N}$, define a quasi-order $\succcurlyeq_k$ on $\mathbb{R}^k$ such that $r \succcurlyeq_k r'$ if and only if $r_i \geq r'_i$ for all $i \leq k$. In other words, $\succcurlyeq_k$ is the (weak) Pareto ordering on $\mathbb{R}^k$. For any $\overline{k} \in \omega^{<\omega}$, let $k^* = \sum_{i \leq l(k)} k_i$, and then define $\mathbb{R}^{\overline{k}} = \mathbb{R}^{k^*}$. For any $j \leq lh(k)$, let $\pi_j : \mathbb{R}^{\overline{k}} \to \mathbb{R}^{k_j}$ be the projection:

$$\pi_j(\langle r_1, r_2, \ldots, r_{k^*} \rangle) = \langle r_{\sum_{i<j} k_i}, \ldots, r_{k_j + \sum_{i<j} k_i} \rangle$$

Finally, given $\overline{k} \in \omega^{<\omega}$, define a quasi-ordering $\lesssim^{\overline{k}}$ on $\mathbb{R}^{\overline{k}}$ as follows. For $r, r' \in \mathbb{R}^{\overline{k}}$, say $r \lesssim r'$ if and only if there exists $j \leq l(\overline{k})$ for which

$$\begin{array}{rcl} \pi_i(r) & = & \pi_i(r') \text{ for all } i < j \text{ and} \\ \pi_j(r) & \preccurlyeq_{k_j} & \pi_j(r') \end{array}$$

Suppose that, for each player $i \in N$, there is a $\overline{k}(i) \in \omega^{<\omega}$ and a function $f_i : A \to \mathbb{R}^{k(i)^*}$ such that for $a, a' \in A$:

$$a \preceq_i a' \Leftrightarrow f_i(a) \lesssim^{\overline{k}} f_i(a')$$

Then say $\mathcal{G} = \langle G, \{f_i\}_{i \in N} \rangle$ is a *strategic quasi-game for incomparable goods.* If $l(\overline{k}(i)) = 1$ for all players $i \in N$, then say $\mathcal{G}$ is a *simple, strategic quasi-game for incomparable goods.* In other words, in such simple games, players' preference relations do not contain lexicographic components, but rather, are merely pareto orderings on finitely many goods or types of value. Here, the functions f_i associate each strategy profile with a list of the finite set of goods of interest to player i. Notice that, because all of the information concerning player i's preferences is encoded by the mapping f_i, one can describe a strategic quasi-game for incomparable goods by a triple $\langle N, \{A_i\}_{i \in N}, \{f_i\}_{i \in N} \rangle$.

As noted above, when players' preferences are linearly-ordered, there are several alternative definitions of a Nash equilibrium. Namely, $a \in A$ is a Nash equilibrium if and only if $a_i \in B(a_{-i})$ for all $i \in N$ if and only if $a_i \in uB(a_{-i})$ for all $i \in N$. In strategic quasi-games, these definitions are no longer equivalent and as such, it is important to introduce different terminology for both.

Definition 3.1.1. Say $a \in A$ is a **strong Nash equilibrium** if and only if $a_i \in B(a_{-i})$ for all $i \in N$. Say $a \in A$ is a **quasi-Nash equilibrium** if $a_i \in uB(a_{-i})$ for all $i \in N$.

Every strong Nash equilibrium is a quasi-Nash equilibrium but not vice versa (i.e. $B(a_{-i}) \subseteq uB(a_{-i})$ for any $a \in A$ and $i \in N$). Moreover, if each player's preference relation is linear, then the set of Nash equilibria and quasi-Nash equilibria are identical. In order to generalize Nash's theorem, one needs to define the payoff associated with a mixed strategy in a quasi-game for incomparable goods. The obvious approach is taken, where the payoff is the weighted averaged of the strategy profiles in each coordinate. That is, suppose $\mathcal{G} = \langle N, \{A_i\}_{i \in N}, \{f_i\}_{i \in N} \rangle$ is a finite, strategic game for incomparable goods. Define the mixed extension $\mathcal{G}^*$ to be the quasi-game for incomparable goods such that:

- N is the finite set of players from $\mathcal{G}$
- Player i's actions are $\mathcal{P}(A_i)$, which is the set of probability measures on A_i. Here, A_i is the set of actions available to player i in the quasi-game $\mathcal{G}$
- For each $i \in N$, the function $f_i^* : \times_{i \in N} \mathcal{P}(A_i) \to \mathbb{R}^{\overline{k}(i)}$ extends f_i from the game $\mathcal{G}$ so that $f_i^*(p_1, \ldots, p_N) = \sum_{a \in A} [(\Pi_{j \in N} \; p_j(a_j)) \cdot f_i(a)]$
- For every $i \in N$, player i's (quasi) preference ordering $\succeq_i$ on $P = \times_{i \in N} \mathcal{P}(A_i)$ is encoded by f_i^*. In other words, for all $p, p' \in P$, one has $p \preceq_i p' \Leftrightarrow f_i^*(p) \lesssim^{\overline{k}(i)} f_i^*(p')$

Let $\mathcal{G}$ be the mixed extension of a finite, strategic quasi-game for incomparable goods. Say a player's preference relation $\succeq_i$ is *convex* if and only if for every $p \in P$, the set $\{p_i' \in \mathcal{P}(A_i) \mid (p_{-i}, p_i') \succeq_i p\}$ is convex in $\mathbb{R}^{\overline{k}(i)}$. Further, say $\succeq_i$ is *continuous* if and only if for every pair of sequences $\langle a^k \rangle_{k \in \mathbb{N}}$ and $\langle b^k \rangle_{k \in \mathbb{N}}$ with limits a and b respectively, if $a^k \preceq_i b^k$ for all k, then $a \preceq_i b$.[6]

Proposition 3.1.1. A strategic quasi-game has a quasi-Nash equilibrium if

- for all $i \in N$ the set of actions A_i for player i is a nonempty, compact, convex subset of $\mathbb{R}^k$ for some $k \in N$
- for all $i \in N$, player i's preference relation $\succeq_i$ is continuous and convex on A.

Proof: This is Proposition 20.3 in Osborne and Rubinstein (1994), which is a special version of Theorem in Nikaido and Isoda (1955). The same proof works here, as Nikaido's proof proceeds by contradiction assuming the non-existence of a quasi-Nash equilibrium (in our sense) and then employing the equivalence of a quasi-Nash equilibrium and strong Nash equilibrium in strategic games.

[6] Here, limits are with respect to the standard Euclidean metric on $\mathbb{R}^n$. Also, recall that a subset $K \subseteq \mathbb{R}^n$ is convex provided that for every pair $x, y \in K$ and any number $s \in [0, 1]$, the element $sx + (1 - s)y \in K$. This captures the notion that if x and y are in K, then the line joining x and y also lies entirely in K.

□

Lemma 3.1.1. Let $\mathcal{G}^*$ be the mixed extension of a **simple**, finite, strategic quasi-game for incomparable goods. Then for all $i \in N$, player $i's$ preference relation is continuous and convex on $\mathcal{P}(A_i)$.

Proof: The lemma follows immediately from the continuity and convexity of the relation $\leq$ on $\mathbb{R}$, but we fill in the details for clarity.

Let $(r^n)_{n\in\mathbb{N}}, (q^n)_{n\in\mathbb{N}}$ be two infinite sequences in $\mathbb{R}^k$ such that (i) $r^n \succcurlyeq_k q^n$ for all $n \in \mathbb{N}$ and (ii) $\lim_{n\to\infty} r^n = r$ and $\lim_{n\to\infty} q^n = q$. We must show $r \succcurlyeq_k q$. Suppose not. Then there is some $q_j > r_j$ for some $j \leq k$. Hence, there is some $\epsilon > 0$ such that $q_j = r_j + \epsilon$. As $r^n \to r$, there is some n_0 such that $|r_j^m - r_j| < \frac{\epsilon}{2}$ for all $m \geq n_0$. Similarly, as $q^n \to r$, there is some n_1 such that $|q_j^m - q_j| < \frac{\epsilon}{2}$ for all $m \geq n_1$. Let $n = \max\{n_0, n_1\}$. It follows that $r_j^m < r_j + \frac{\epsilon}{2} < q_j - \frac{\epsilon}{2} < q_j^m$ for all $m \geq n$. In particular, by the definition of the ordering $\succcurlyeq_k$, it follows that $r^n \not\succcurlyeq q^n$, contradicting assumption. So $\succcurlyeq_k$ is continuous.

To show that $\succcurlyeq_k$ is convex, suppose $r, r' \succcurlyeq_k q$. By the definition of $\succcurlyeq_k$, it follows that $r_j, r'_j \geq q_j$ for all $j \leq k$. Hence, for all $j \leq k$ and all $\delta \in [0, 1]$, it follows that $(\delta \cdot r_j) + ((1 - \delta) \cdot r'_j) \geq q_j$, which implies $(\delta \cdot r) + ((1 - \delta) \cdot r) \succcurlyeq q$, as desired. □

Theorem 3.1.2. Every simple, finite, strategic quasi-game for incomparable goods has a mixed strategy quasi-Nash equilibrium.

Proof: If the set of player i's actions $A_i = \{a_1, \ldots, a_n\}$ is a finite set for all $i \in N$, then the a probability distributions $p \in \mathcal{P}(A_i)$ over A_i is representable as a vector in $\mathbb{R}^n$, namely, $\langle p(a_1), \ldots, p(a_n)\rangle$. The set of all probability distributions $\mathcal{P}(A_i)$ on A_i is clearly non-empty, convex, and compact. By Lemma 3.1.1, each player $i's$ preference relation is continuous and convex on $\mathcal{P}(A_i)$. Hence, by Proposition 3.1.1, there exists a quasi-Nash equilibrium $p \in P$. □

Requiring the quasi-game to be simple is necessary for the proof of Lemma 3.1.1, as the relation $\precsim^{\overline{k}}$ is not continuous when $l(\overline{k}) > 1$. For example, let $\overline{k} = \langle\langle 1\rangle, \langle 1\rangle\rangle$, which is the standard lexicographic ordering $\precsim^{\overline{k}}$ on $\mathbb{R}^2$. That is, $\precsim^{\overline{k}}$ is the ordering such that $\langle x, y\rangle \precsim^{\overline{k}} \langle x', y'\rangle$ if and only if (i) $x \leq x'$ or (ii) both $x = x'$ and $y \leq y'$. Let $\langle r^n\rangle_{n\in\mathbb{N}}$ be the sequence $r^n = \langle 2 - \frac{1}{2^n}, 2\rangle$, and let q^n be the constant sequence $\langle 2, 1\rangle$. Then $r^n <^{\overline{k}} q^n$ for all n, but

$$\langle 2, 1\rangle = \lim_{n\to\infty} q^n <^{\overline{k}} \lim_{n\to\infty} r^n = \langle 2, 2\rangle$$

Hence, $\precsim^{\overline{k}}$ is not continuous. It is an open question whether one can amend the above proof to quasi-games for incomparable goods more generally.

3.2 Three Representations of the KGS Model as a Quasi-Game

I now describe three ways in which one might represent the KGS model of inquiry as an infinite strategic game. In each of the three representations, one begins with the assumption that there are two players: Nature and the scientist. I will use N to denote "Nature," and S to denote scientist.[7] The three ways of understanding the KGS model, then, correspond to ways in which one can specify the actions of these two players. After describing the three different representations, I argue the second is the most reasonable model of scientific inquiry, as it captures the *dynamic* aspect of learning over time in a way the first representation does not, and it does not anthropomorphize Nature in the way that the third does.

Before I begin, two remarks are in order. First, one might ask, "if only the second representation of the KGS model as a game is philosophically interesting, why do I spend so much time discussing the other two?" It turns out, for purely formal reasons, that a number of the mathematical results concerning the first representation of the KGS model can be re-used in analyzing the second and third representations. Similarly, the third representation exhibits a mathematical symmetry not possessed by the second, and thus, it contributes to an explanation of why the second model does and does not have equilibria under various assumptions.

Second, all of the games discussed in this paper are *strategic* (also called *normal form*) games rather than *extensive-form* games. Recall, that in strategic games, players act simultaneously, whereas in extensive games, players act sequentially and can (in some circumstances) choose their next move in light of previous moves made in the game. This distinction is important, but the arguments below show that the game-theoretic analogs of the Efficiency Theorem can be proven using only strategic form representations of the KGS model.

3.2.1 G : Actions as Answer and Effect Sequences

Imagine that inquiry is a one-shot, simultaneously play game between nature and the scientist in which the scientist's set of pure strategies is identical to the set of infinite sequences of answers, and that nature's set of pure strategies is the set of infinite sequences of nested sets of effects. In other words:

$$\begin{aligned} A_N &= \{\eta \in (2^E)^\omega \ : \ \eta_n \subseteq \eta_{n+1} \text{ for all } n \in \mathbb{N}\} \\ A_S &= \mathsf{Ans}^\omega \end{aligned}$$

[7] From hereon, I will **not** use N to refer to the set of players in a game. Similarly, A will no longer represent the set of strategy profiles, but rather, will be used as a variable to indicate an element of Ans as defined in the KGS model.

In order to define mixed strategies, however, one must introduce some definitions. For any $\alpha \in A_N$, $\eta \in A_S$, and $m \in \mathbb{N}$, define:

$$\begin{aligned}[][\alpha \restriction m] &= \{\alpha' \in A_S \ : \ \alpha' \restriction m = \alpha \restriction m\} \\ [\eta \restriction m] &= \{\eta' \in A_N \ : \ \eta' \restriction m = \eta \restriction m\}\end{aligned}$$

In other words, $[\alpha \restriction m]$ is the set of infinite answer sequences α' that are identical to α up to stage m. Similarly, $[\eta \restriction m]$ is the set of infinite effect sequences η' that are identical to η up to stage m. Let $\mathcal{A}_S = \sigma(\{[\alpha \restriction n] : \alpha \in A_S, n \in \mathbb{N}\})$ and $\mathcal{A}_N = \sigma(\{[\eta \restriction n] : \eta \in A_N, n \in \mathbb{N}\})$ be the respective σ-algebras created by taking the σ-closure of the events defined above.[8] It is convenient to know that the following events are measurable in these two σ-algebras.

[Add the following to the previous chapter describing the KGS model]

Lemma 3.2.1. *W is countable.*

Proof: Any world $w \in W$ can be identified with a pair (T_w, e_w) where T_w is the theory true of w and $e \in W_{fin}$ is the largest finite initial segment of w before which w converges to the true theory w. Hence, $|W| = |\mathsf{Th}| \cdot |W_{fin}|$ both of which are countable by [just create a large list of countable events in the previous chapter]. □

Proposition 3.2.1. *The following events are measurable:*

1. *For any $A \in \mathsf{Ans}$ and any $m \in \mathbb{N}$, the event $[\alpha_m = A] = \{\alpha \in A_S : \alpha_m = A\}$ is $\mathcal{A}_S$-measurable. For any $E_0 \subset E$, the event $[\eta_m = E_0] = \{\eta \in A_N : \eta_m = A\}$ is $\mathcal{A}_N$-measurable*

2. *For any $\alpha \in A_S$, the singleton $\{\alpha\}$ is $\mathcal{A}_S$-measurable, and analogously, the singleton $\{\eta\}$ is $\mathcal{A}_N$-measurable. In particular, any world $w \in W$ is $\mathcal{A}_N$-measurable.*

3. *For any $E_0 \subset E$, the set $C^{E_0}_m = \{\eta \in A_N \ : \ \eta_{m'} = \eta_m = E_0$ for all $m' \geq m\}$ is $\mathcal{A}_N$ measurable. Similarly, for any $A \in \mathsf{Ans}$, the set $C^A_m = \{\alpha \in A_S \ : \ \alpha_{m'} = \alpha_m = A$ for all $m' \geq m\}$ is $\mathcal{A}_S$-measurable.*

4. *For any set $E_0 \in K$ The set $C^{E_0} = \{\eta \in A_N \ : \ \exists m \forall m' \geq m(\eta_{m'} = \eta_m = E_0)\}$ is $\mathcal{A}_N$-measurable. Similarly, for any $A \in \mathsf{Ans}$, the set $C^A = \{\alpha \in A_N \ : \ \exists m \forall m' \geq m(\alpha_{m'} = \alpha_m = A)\}$ is $\mathcal{A}_S$-measurable.*

5. *The set $C^K_m = \{\eta \in A_N \ : \ \eta_{m'} = \eta_m \in K$ for all $m' \geq m\}$ is $\mathcal{A}_N$ measurable.*

[8] Because Ans and E are countable, the σ-algebras $\mathcal{A}_S$ and $\mathcal{A}_N$ are isomorphic to the standard Borel algebra on ω^ω. See Appendix 1 for definition of "Borel Algebra." That is, because Ans is countable, one can think of a sequence in Ans^ω simply as a sequence of natural numbers in ω^ω (the so called Baire space). Then, just as one defines a topology on ω^ω, one can consider the sets $[\alpha \restriction m]$, for example, for the basis of a topology on A_S, and so the $\mathcal{A}_S$ is simply the Borel algebras with respect to this topology.

6. The set of worlds W is $\mathcal{A}_N$-measurable, and thus, A_N contains the power set $\mathcal{P}(W)$ of worlds.

7. Define $I(\eta)$ to be the greatest natural number n such that there are $m_1 < m_2 \ldots < m_n \in \mathbb{N}$ and $\eta_{m_1} \subset \eta_{m_2} \subset \ldots \subset \eta_{m_n} \in K$ such that $T_{\eta_{m_i}} \neq T_{\eta_{m_{i+1}}}$ for all $i \leq n$. Let $[I(\eta) = n] = \{\eta \in A_N \ : \ I(\eta) = n\}$, and define $Comp_n$ to be the set $W \cap [I(\eta = n)]$. $Comp_n$ is called the "n^{th} complexity class." Then $Comp_n$ is A_N-measurable.[9]

Proof:
(1) As Ans is countable, the set of sequences Ans^{n+1} of answers of length $n+1$ is also countable. Hence, the set Ans_A^{n+1} of finite sequences of answers of length $n+1$ whose last coordinate is A is at most countable, and it's also clearly infinite as Ans is countable. Thus

$$[\alpha_m = A] = \bigcup_{\alpha \restriction n+1 \in \mathsf{Ans}_A^{n+1}} [\alpha \restriction n+1]$$

is a countable union of measurable events, and is therefore measurable. The exact analogous approach applies to showing $[\eta_m = E_0]$ is measurable.
(2) Notice that

$$\alpha = \bigcap_{m \in \mathbb{N}} [\alpha \restriction m]$$

is a countable intersection of measurable events by (1), and hence is measurable. Similarly for η.
(3) $C_m^{E_0} = \bigcap_{m' \geq m} [\eta_{m'} = E_0]$ is a countable intersection of measurable events by (1). Similarly for C_m^A.
(4) $C^{E_0} = \bigcup_{m \in \mathbb{N}} C_m^{E_0}$, and similarly for C^A.
(5) Follows immediately from the fact that $C_m^K = \bigcup_{E_0 \in K} C_m^{E_0}$ and that K is countable.
(6) W is measurable because $W = \bigcup_{m \in \mathbb{N}} C_m^K$, and C_m^K is measurable by (5). To show $\mathcal{P}(W) \subset A_N$, notice that any subset $W_0 \subseteq W$ is at most countable because W is itself countable. Hence, W_0 is at most a countable union of singletons, and so is measurable.
(7) Follows immediately from (6). □

Say a mixed strategy for the scientist is a probability measure p_S on $\mathcal{A}_S$ and that a mixed strategy for Nature is a probability measure p_N on $\mathcal{A}_N$. To finish specifying the representation of the KGS model as a game, one needs to define the preference relations on strategy profiles. Two major caveats are

[9] In Kelly's papers, the definition of complexity class differs slightly from the one presented here. The relation between the two is that for all n, the definition of complexity class n here is a subset of the complexity class n is Kelly's sense, namely, the subset of worlds in which there are no "jumps" in complexity. The reader is urged to consult those papers to examine the difference more formally.

necessary before I introduce said preference relations. First, notice that, in allowing her to choose arbitrary sequences in $\eta \in A_N$, Nature is no longer constrained to stop presenting effects at any point in inquiry, which contradicts the assumption in the KGS model that only finitely many effects occur in any world. By appropriately defining the preference relations below, however, I ensure that there is a cost to presenting infinitely many effects, which ensures that Nature, if behaving rationally, will only choose answer sequences that are worlds. Again, one should be careful not to understand this feature of the game as modeling the preferences of some agent called "Nature." Rather, penalizing Nature for presenting an infinite number of effects is equivalent to forgiving a scientist for his or her mistakes when confronted with empirical problem that is, by definition, unsolvable in any finite amount of time.

Second, throughout the following three sections, I discuss preference relations that contain infinitely long chains, in the sense that one player in the game may prefer action a_{n+1} to a_n for some infinitely long sequence of actions $(a_n)_{n \in \mathbb{N}}$. It is well-known that decision theory, and expected utility theory in particular, yields strange and counterintuitive results when dealing with unbounded utilities. Hence, one may be extremely skeptical of drawing normative conclusions about scientific inquiry from the results below. At the end of the chapter, I discuss why the use of preference relations with infinitely long chains is unproblematic, and even required by the KGS model.

Recall, in Kelly, Glymour, and Schulte's model, the cost of employing method M in world w is represented by a cost-vector $\lambda(M, w) = \langle \epsilon(M, w), \epsilon(M, w) \rangle$, where $\epsilon(M, w)$ and $\rho(M, w)$ are the number of errors and retractions committed by M in w respectively. When comparing the cost of a method M to that of its rivals, the scientist also implicitly evaluates whether M converges in w or not, as she prefers convergent methods (in w) to non-convergent ones, regardless of the number of errors and retractions M commits in w. That is, define $c(M, w)$ to be 0 if M converges in w and 1 otherwise. In the KGS model, the scientist strictly prefers any method that converges in a world to ones that do not, and she then prefers methods that commit fewer errors and fewer, earlier retractions.

Although cost vectors are a function of methods and worlds, one can tweak their definition ever so slightly in order to develop a preference ordering for nature and science in the game in which their actions are answer and effect sequences respectively. Say an effects sequence $\eta \in A_N$ *converges* if $\eta \in W$. Define a function $c_N : A_N \to \{0, 1\}$ such that c_N is zero if $\eta \in W$ and is one otherwise. Similarly, say answer sequence $\alpha \in A_S$ *is convergent* in η if either (a) $\eta \notin W$, or (b) there exists an $n \in \mathbb{N}$ such that $\alpha_m = T_\eta$ for all $m \geq n$. Define $c_S(\alpha, \eta)$ to be zero if α converges in η and one otherwise. Further, define:

$$\epsilon(\alpha, \eta, n) = \begin{cases} 1 \text{ if } \alpha_n \neq T_w \text{ and } \eta \in W \\ 0 \text{ otherwise} \end{cases}$$

$$\rho(\alpha, n) = \begin{cases} 1 \text{ if } \alpha_n \neq \alpha_{n+1} \text{ and } \alpha_n \neq \text{ `?'} \\ 0 \text{ otherwise} \end{cases}$$

Notice that ρ does not depend upon η, as the scientist chooses an infinitely long sequence of answers, **not** a method for selecting theories from data. This will be important later. One can then define the retraction vector and the total errors of α in η as follows:

$$\begin{aligned} \epsilon(\alpha, \eta) &:= \sum_{n=0}^{\infty} \epsilon(\alpha, \eta, n) \\ \rho(\alpha) &:= \sum_{n=0}^{\infty} \rho(\alpha, n) \end{aligned}$$

Finally, the above definitions allow one to define a preference relation for the scientist as follows:

$$\text{(I)} \;\; (\alpha, \eta) \preceq_S (\alpha', \eta') \Leftrightarrow \begin{cases} \eta \in W \text{ and } \eta' \notin W \text{ or} \\ \eta, \eta' \in W \text{ and } c(\alpha, \eta) > c(\alpha', \eta') \text{ or} \\ \eta, \eta' \in W, c(\alpha, \eta) = c(\alpha, \eta'), \epsilon(\alpha, \eta) \geq \epsilon(\alpha', \eta') \text{ and } \rho(\alpha) \geq \rho(\alpha') \end{cases}$$

Notice, the scientist's preference relation is not a total. For simplicity, suppose the game is *strictly competitive,* so that the nature disprefers (α, η) to (α', η') if and only if the scientist prefers (α, η) to (α', η'). With the above definitions, one can define the costs associated with mixed strategies by taking weighted averages of the pure strategies employed in the mixture as follows.

Let p_S and p_N be mixed strategies for the scientists and Nature respectively. Because the scientist and nature choose their respective strategies simultaneously, their choices are probabilistically independent of one another. Therefore, it makes sense to stipulate that the probability that a given pair of strategies (α, η) will be played is simply the product of the probabilities $p_S(\alpha)$ and $p_N(\eta)$. The expected number of errors and retractions of employing a mixed strategy p_S in response to p_N, then, is as follows:[10]

$$\begin{aligned} \overline{\epsilon}(p_S, p_N) &:= \int_{A_N} \int_{A_S} \sum_{n=0}^{\infty} \epsilon(\eta, \alpha, n) \cdot p_S(d\alpha) \cdot p_N(d\eta) \\ \overline{\rho}(p_S) &:= \int_{A_S} \sum_{n=0}^{\infty} \rho(\alpha, n) \cdot p_S(d\alpha) \\ \overline{c}_N(p_N) &:= \int_{A_N} \overline{c}_N \; dp_N \\ \overline{c}_S(p_S, p_N) &:= \int_{A_N} \int_{A_S} \overline{c}_S \; dp_S \; dp_N \end{aligned}$$

If the above series diverge on a set of positive measure, assign the integrals the value ∞. Again, it is important to note that $\overline{\rho}$ is a function of p_S only, which means that in G, the scientist's losses due to expected retractions depends upon his strategy alone, and is not changed if Nature alters her strategy.

Alternatively when p_S and p_N are both countably additive, one can define p to be the product measure on the product space $\mathcal{A}_S \otimes \mathcal{A}_N = \sigma(\{X \times Y \; : \; X \in$

[10] I use the standard definition of the Lebesgue integral. See Appendix 1.

$\mathcal{A}_S,\ Y \in \mathcal{A}_N\}$). That is, p is the *unique* measure on $\mathcal{A}_S \otimes \mathcal{A}_N$ such that for any X and Y such that $X \in \mathcal{A}_S$ and any $Y \in \mathcal{A}_N$:[11]

$$(*)\ p(X,Y) = p_S(X) \cdot p_N(Y)$$

One can then provide an alternative definition $\overline{\epsilon}$, $\overline{\rho}$, and so on, as follows:

$$\begin{aligned}
\overline{\epsilon}(p_S,p_N) &= \int_{A_N \times A_S} \sum_{n=0}^{\infty} \epsilon^n \cdot dp \\
\overline{\rho}(p_S) &= \int_{A_S} \sum_{n=0}^{\infty} \overline{\rho}^n \cdot dp_S \\
\overline{c}_N(p_N) &= \int_{A_N} \overline{c}_N \ dp_N \\
\overline{c}_S(p_S,p_N) &= \int_{A_N \times A_S} \overline{c}_S \ dp
\end{aligned}$$

where $\epsilon^n : A_S \times A_N \to \{0,1\}$ is the map $\epsilon_n(\alpha,\eta) = \epsilon(\alpha,\eta,n)$ which determines whether α errs in η at time n or not. Similar remarks apply for $\overline{\rho}^n$.

A note of warning is appropriate here. These two definitions of costs are **not** equivalent if either p_S or p_N is finitely additive. In such cases, there exist probability measures on the product space $\mathcal{A}_S \otimes \mathcal{A}_N$ satisfying $(*)$, but such measures, are, in general, not unique. When either p_S or p_N is finitely additive, the first definitions of $\overline{\epsilon}$, $\overline{\rho}$, $\overline{c}_N$, and $\overline{c}_S$ above ought to be used.

Given the above definitions of expected retractions, errors, and convergence costs, one can then use the definition of the preference relations in (I) to provide a reasonable comparison of costs of mixed strategies for both the scientist and Nature as follows:

$$\text{(II)}\ \ (p_S,p_N) \preceq_S (p'_S,p'_N) \Leftrightarrow \begin{cases} \overline{c}_N(p_N) < \overline{c}_N(p'_N) \text{ or} \\ \overline{c}_N(p_N) = \overline{c}_N(p'_N) \text{ and } \overline{c}(p_S,p_N) > \overline{c}(p'_S,p'_N) \text{ or} \\ \overline{c}_N(p_N) = \overline{c}_N(p'_N), \overline{c}_S(p_S,p_N) = \overline{c}_S(p'_S,p'_N), \overline{\epsilon}(p_S,p_N) \geq \overline{\epsilon}(p'_S,p'_N) \text{ and} \\ \overline{\rho}(p_S,p_N) \geq \overline{\rho}(p'_S,p'_N) \end{cases}$$

This completes the description of the first way in which the KGS model can be represented as a strategic quasi-game. For brevity, call the quasi-game G. Although the preference relation above is close to that defined in the KGS model in the previous chapter, there is one feature in need of explanation: why does the scientist prefer some pair of mixed strategies (p_S,p_N) to another (p'_S,p'_N) if $\overline{c}_N(p_N) < \overline{c}_N(p'_N)$? In particular, reliable learning (in the sense of converging to the true theory) is, in general, impossible in effect sequences η that are not worlds; by definition, in such η, there is no "true theory" to which one could converge! So shouldn't an adversarial player Nature prefer to thwart the scientist's attempts to learn?

[11] The existence and uniqueness of p follows from Caratheodory's Theorem. See Appendix 1.

In representing the KGS model as a game, one should be careful to avoid speaking of Nature as a player too literally. Even though a Cartesian-like demon might prefer to thwart a scientist's attempts to learn, I do not truly believe that there is an adversarial agent controlling the universe who seeks to undermine human attempts to understand their surroundings. The competitive structure of the games developed in this thesis is intended to symbolize or reflect the arduous character of scientific inquiry. Hence, the scientist's preference relation in G is constructed so as not to penalize the scientist the scientist for failing to converge, for erring, and/or for retracting when learning is impossible. In other words, I assume that "ought" implies "can." In questioning whether a systematic bias for simpler theories is rational in scientific inquiry, one must side-step broader skeptical worries about whether science can ever discover "truth."

Because each player's preferences are only partially-ordered (and hence not representable as utilities), Nash's theorem does not guarantee the existence of a mixed strategy solution to G. If K is finite, however, it turns out that G has a strong Nash equilibrium, and in a certain sense, this equilibrium is unique. If K is infinite, then G has neither a pure strategy nor countably-additive mixed strategy quasi-Nash equilibria (and hence, no strong Nash equilibria either). To prove these facts, two short lemmas are needed.

The first lemma asserts, roughly, that Nature minimizes "convergence costs" by playing some mixture of worlds, rather than a mixture of effect sequences that do not converge to an element of K. In other words, Nature is penalized for playing any sequence of effects for which it would be impossible for the scientist to learn the true theory in some finite amount of time. Moreover, if nature plays a countably additive mixed strategy, then as time elapses, the probability that she will present any new effects in the future approaches zero.

Lemma 3.2.2. If p_N is countably additive, then the following are equivalent:

1. $\overline{c}_N(p_N) = 0$
2. $p_N(W) = 1$
3. $\lim_{m\to\infty} p_N(C^K_m) = 1$

Proof: (1)⇒(2): Suppose $\overline{c}_N(p_N) = 0$. By definition, for any $\eta \in A_N$, one has $\overline{c}_N(\eta) = 0$ if $\eta \in W$ and is 1 otherwise. So:

$$\begin{aligned} 0 &= \int_{A_N} \overline{c}_N \; dp \\ &= \int_{A_N \setminus W} \overline{c}_N \; dp + \int_W \overline{c}_N \; dp \\ &= (p_N(A_N \setminus W)\cdot 1) + (p_N(W)\cdot 0) \\ &= p_N(A_N \setminus W) \end{aligned}$$

It follows that:

$$1 = p_N(A_N) = p_N(A_N \setminus W) + p_N(W) = p_N(W)$$

(2)⇒(1): The above proof is reversible. Explicitly, if $p_N(W) = 1$, then it follows that:

$$\begin{aligned} \overline{c}_N(p_N) &= \int_{A_N} \overline{c}_N \; dp \\ &= \int_W \overline{c}_N \; dp \\ &= p(W) \cdot 0 \\ &= 0 \end{aligned}$$

(2)⇔(3): Note that $C^K_m \subseteq C^K_{m+1}$ for all m. Now because $W = \bigcup_{m \in \mathbb{N}} C^K_m$, by countable additivity it follows that:

$$p(W) = p(\bigcup_{m \in \mathbb{N}} C^K_m) = \lim_{m \to \infty} p_N(C^K_m)$$

The desired equality follows. Notice that countable additivity is only used to prove the third statement equivalent to the first two. □

Together, the next two propositions show that the game G is structurally similar to the game matching pennies discussed above. Why? In G the scientist wishes to "match" the "theory" played by Nature, and Nature desires the exact opposite. Hence, just as in matching pennies, G has no pure strategy equilibria, which is proven in Proposition . Similarly, if the set set of theories is finite, then just as in matching pennies, Nature can minimize her worst-case losses by playing each theory with equal probability, and the scientist maximizes his worst-case gains by guessing each theory with equal probability. Unsurprisingly, then, G has a countably additive mixed strategy equilibrium when Ths is finite.[12] This is proven in Proposition 3.2.3. Of course, if the scientist and Nature's mixed strategies are countably additive, then assigning equal probability to each theory is possible only when there are only finitely many theories, which is why the restriction that Th is finite is necessary.

Proposition 3.2.2. *G has no pure strategy quasi-Nash equilibria.*

Proof: Suppose the scientist employs answer sequence α (a pure strategy) and nature chooses the effect sequence η. If $\eta \notin W$, then Nature would have preferred to play a world regardless of the scientist's choice. Hence, suppose $\eta \in W$. Now either α converges in η or it doesn't. If α does not converge in η, then the scientist would have preferred to play a sequence α' that does converge in η according to the above preference relation. So suppose α does converge in η, which means there is some $n \in \mathbb{N}$ such that $\alpha_m = T_w$ for all $m \geq n$. Let η' be a world such that $T_{\eta'} \neq T_\eta$. Such an η exists because Th has at least two elements. Then α does not converge in η, and so Nature would benefit from playing η' instead of η, keeping α fixed. This shows that, whichever strategies

[12] Strictly speaking, because the scientist's preference relation contains a lexicographic component, the existence of an equilibrium does not follow from the fact that the above game is zero-sum and there are minimax strategies. See Fishburn (1972).

are chosen by the scientist and Nature, one of the two players would benefit from deviating if the other keeps his or her strategy fixed. Hence, there are no pure strategy Nash equilibria in the game.

□

Proposition 3.2.3. If Th is finite, then G has a Nash equilibrium amongst countably-additive strategies.

In order to prove the proposition, we first need a somewhat technical lemma:

Lemma 3.2.3. Let p_N be a countably-additive mixed strategy for Nature in G, and suppose that $\overline{c}_N(p_N) = 0$. Let if p_S is countably-additive strategy for the scientist. Then, there exists a finite collection of theories $T_1, \ldots, T_m \in \mathsf{Th}$ such that p_S is a best response (amongst countably additive strategies) for the scientist if and only if:

$$1 = p_S(\{\alpha \in A_S \ : \ (\exists n)(\exists i \leq m)[\forall n' < n(\alpha_{n'} = \text{`?'}) \text{ and } \forall n' \geq n(\alpha_{n'} = T_i)]\})$$

Proof: By assumption $\overline{c}_N(p_N) = 0$, and so by Lemma 3.2.2, if follows that $p_N(W) = p_N(\bigcup_{E' \in K} C^{E'}) = 1$. Because p_N is countably additive, there exist finitely many effect sets $E_0, \ldots, E_m \in K$ such that p_N assigns maximal probability to C^{E_i} for each $i \leq m$. Let that probability be q. In other words, $p_N(C^{E_i}) > 0$ and $q := p_N(C^{E_i}) \geq p_N(C^{E'})$ for all $E' \in K$. In particular, $p_N(C^{E_i}) = p_N(C^{E_j})$ for all $i, j \leq m$.. Note that for all $E' \in K$ and all $i \leq m$:

$$(\dagger) \quad 1 - p_N(C^{E'}) \geq 1 - p_N(C^{E_i}) = 1 - q$$

where the inequality is strict if $E' \neq E_i$ for some $i \leq m$. Now define:

$$\begin{aligned} a_i &= \{\alpha \in A_S \ : \ (\exists n)[\forall n' < n(\alpha_{n'} = \text{`?'}) \text{ and } \forall n' \geq n(\alpha_{n'} = T_{E_i})]\} \\ a &= \bigcup_{i \leq m} a_i \end{aligned}$$

Let p^* be a countably-additive measure such that $p^*(a) = 1$. Note that $p^*(a) = \sum_{i \leq m} p^*(a_i)$ as the a_i's are disjoint. We claim p^* is a best response to p_N. First we show that p^* maximizes $\overline{c}_S$ if p_N is held fixed. Let p be the unique product measure induced by p^* and p_N, and define:

$$\begin{aligned} U &= \{\langle \alpha, \eta \rangle \in A_S \times A_n \ : \ \overline{c}_S(\alpha, \eta) = 1\} \\ U_1 &= \{\alpha \in A_S \ : \ \langle \alpha, \eta \rangle \in U \text{ for some } \eta \in A_N\} \\ U_\alpha &= \{\eta \in A_N : \langle \alpha, \eta \rangle \in U\} \end{aligned}$$

One can now calculate that:

$$\begin{aligned}
\overline{c}_S(p^*, p_N) &= \int_{A_N} \int_{A_S} \overline{c}_S \; dp^* \cdot dp_N \\
&= \int_{A_N \times A_N} \overline{c}_S \; dp \qquad \text{by Fubini's theorem, as both } p^* \& p_N \text{ are countably additive.} \\
&= \int_U \overline{c}_S \; dp + \int_{(A_N \times A_S) \setminus U} \overline{c}_S \; dp \\
&= \int_U 1 \; dp + \int_{(A_S \times A_N) \setminus U} 0 \; dp \\
&= \int_U dp \\
&= \int_{U_1} \int_{U_\alpha} dp_N \; dp^* \qquad \text{again by Fubini's theorem.} \\
&= \int_a \int_{U_\alpha} dp_N \; dp^* \qquad \text{as } p^*(a) = 1 \\
&= \sum_{i \leq m} p^*(a_i) \cdot \left[\int_{W \setminus C^{E_i}} dp_N \right] \\
&= \sum_{i \leq m} p^*(a_i) \cdot (1 - p(C^{E_i})) \qquad \text{as } p_N(W) = 1 \\
&= 1 - q
\end{aligned}$$

Recall that $\overline{c}_S$ indicates **non**-convergence to the truth, and so the scientist wishes to minimize $\overline{c}_S$. In contrast, let p_S be any countably-additive mixed strategy for the scientist, and let p' be the product measure on $\mathcal{A}_S \otimes \mathcal{A}_N$ induced by p_S and P_N. For any answer sequence $\alpha \in A_S$, define E_α to be the total set of effects presented in α. Abusing notation, let $C^{E_\alpha} = \emptyset$ if $E_\alpha \notin K$, and be defined as before otherwise. It follows that:

$$\begin{aligned}
\overline{c}_S(p_S, p_N) &= \int_{A_N} \int A_S \overline{c}_S \; dp_S \; dp_N \\
&= \int_{A_S \times A_N} \overline{c}_S \; dp' \qquad z \text{ by Fubini} \\
&= \int_U dp \\
&= \int_{U_1} \int_{U_\alpha} dp_N \; dp_S \\
&= \int_{U_1} [\int_{W \setminus C^{E_\alpha}} dp_N] \; dp_S \qquad \text{as } p_N(W) = p_N(C^K) = 1 \\
&\geq \int_{U_1} (1-q) \; dp_S \qquad \text{by } \dagger \\
&\geq 1 - q \\
&= c(p^*, p_N)
\end{aligned}$$

Hence, if $p_S(a) = 1$ for some $i \leq m$, it follows that $\overline{c}_S(p_S, p_N) \leq \overline{c}_S(p'_S, p_N)$ for any mixed strategy p'_S for the scientist. Next we consider errors and retractions. By definition p^* assigns probability one to answer sequences α for which $\overline{\rho}(\alpha, \eta) = 0$ for all $\eta \in A_N$. Hence, p^* has zero expected retractions regardless of the strategy employed by Nature, and thus, $\overline{\rho}(p^*, p_N) = 0 \leq \overline{\rho}(p_S, p_N)$ for any other mixed strategy p_S employed by the scientist.

Finally, consider the expected errors of p^*. There are two cases to consider. First, consider the case that $p_N(W^{T_i}) = 1$. Then by definition, p^* assigns probability one to answer sequences whose entries contains only '?' and T_i as their entries. It follows that $\overline{\epsilon}(p^*, p_N) = 0$, and so p^* has no more expected errors than any other mixed strategy that could be employed by the scientist. Next, consider the case that $p_N(W^{T_i}) \neq 1$. Then there are at least two effect sets E_0 and E_1 such that $p_N(C^{E_0}) = p(C^{E_1}) > 0$. It's easy to check that every convergent strategy employed by the scientist accrues infinite expected errors, and so p^* does as well as any other mixed strategy employed by the scientist. Hence, p^* is a best response to p_N.

Conversely, suppose that p_S is a countably additive measure such $p_S(a) \neq 1$. We want to show that p_S is not a best response to p_N. It suffices to show that $c(p_S, p_N)$ is strictly greater than $1 - q$. Again, noting that $p_N(W) = 1$, the calculation is nearly identical to the two above:

$$\begin{aligned}
\overline{c}_S(p_S, p_N) &= \int_U dp \\
&= \int_{U_1}\int_{U_\alpha} dp_N \; dp_S \\
&= \int_a \int_{W \setminus C^{E_\alpha}} dp_N \; dp_S + \int_{A_S \setminus a}\int_{W \setminus C^{E_\alpha}} dp_N \; dp_S \\
&= p(a)(1-q) + \int_{A_S \setminus a}\int_{W \setminus C^{E_\alpha}} dp_N \; dp_S \\
&> p(a)(1-q) + \int_{A_S \setminus a} (1-q) \; dp_S \qquad \text{by } \dagger \\
&= p(a)(1-q) + (1-p(a))(1-q) \\
&= 1-q
\end{aligned}$$

This completes the proof.

(End Proof of Lemma) □

Proof of Proposition 3.2.3: Let $\mathsf{Th} = \{T_1, \ldots, T_m\}$, and let $\{E_1, \ldots, E_m\} \subseteq K$ be the effect sets corresponding to each of the finitely many theories T_i (i.e. $T_{E_i} = T_i$). Construct a strong Nash equilibrium as follows. Let p_N be a probability measure on $\mathcal{A}_N$ (i.e. mixed strategy for Nature) such that $p_N(C^{E_i}) = \frac{1}{m}$ for all $i \leq m$. Then, by definition, $p_N(W) = 1$ and so $\overline{c}_N(p_N) = 0$ by Lemma 3.2.2. Suppose a_i is defined as in Lemma 3.2.3 (See below), and let p_S be a probability measure on $\mathcal{A}_S$ such that $p_S(a_i) = \frac{1}{m}$. By the same calculations as in the proof of Lemma 3.2.3, it follows that (i) $\overline{c}_S(p_S, p_N) = 1 - \frac{1}{m}$, (ii) $\overline{\rho}(p_S, p_N) = 0$, (iii) $\overline{\epsilon}(p_S, p_N)$ is either 0 or ∞ depending upon whether the cardinality of K is or is not greater than 1, and finally (iv) that p_S is a best response to p_N. We claim that p_N is a best response to p_S.

Whichever strategy p'_N Nature employs, it's clear that $\overline{\rho}(p_S, p'_N) = 0$, so p_N does no worse with regard to expected retractions than other mixed strategies for Nature. Now consider errors. If $m > 1$, then $\overline{\epsilon}(p_S, p_N) = \infty \geq \overline{\epsilon}(p_S, p'_N)$ for any mixed strategy p'_N. If $m = 1$, then $p'_N(C^{E_0}) = 1$ for any unbeaten mixed strategy for nature, in which case $\overline{\epsilon}(p_S, p_N) = 0 = \overline{\epsilon}(p_S, p'_N)$. So it suffices to show that $\overline{c}_S(p_S, p_N)$ is maximized by p_N. It turns out that $\overline{c}_S(p_S, p'_N) = 1 - \frac{1}{m}$

no matter the strategy p'_N chosen by Nature:

$$\begin{aligned}
\overline{c}_S(p_S, p'_N) &= \int_{A_S \times A_N} \overline{c}_S \; dp \\
&= \sum_{i \leq m} p_S(a_i) \cdot \sum_{j \neq i} p'_N(C^{E_j}) \\
&= \frac{1}{m}(p'_N(C^{E_2}) + \ldots + p'_N(C^{E_m})) + \frac{1}{m}(p'_N(C^{E_1}) + p'_N(C^{E_3}) \ldots + p'_N(C^{E_m})) + \ldots \\
&\quad + \frac{1}{m}(p'_N(C^{E_1}) + \ldots + p'_N(C^{E_{m-1}})) \\
&= \frac{1}{m} \sum_{i \leq m} (m-1) p'_N(C^{E_i}) \\
&= \frac{m-1}{m} \\
&= 1 - \frac{1}{m}
\end{aligned}$$

□

The above proof is informative as it constructs a Nash equilibrium for the game G. However, one can provide a second, shorter proof of the existence of a quasi-Nash equilibrium by employing Proposition 3.1.1. Suppose K contains m-many elements $E_1, \ldots, E_m$. By Lemmas 3.2.2 and 3.2.3, if p_N is an un-dominated mixed strategy for Nature (in the sense that there is no p'_N for Nature that is preferred to p_N regardless of the scientist's choice of strategy), then $p_N(W) = 1$. Similarly, if p_S is an un-dominated strategy for the scientist, then $p_N(a) = 1$ where $a = \bigcup_{i \leq m} a_i$ is defined as in Lemma 3.2.3. Now, one can check that, for any fixed mixed strategy for Nature p_N, and for any two answer sequences $\alpha, \alpha' \in A_S$ such that $\alpha, \alpha' \in a_i$:

$$\begin{aligned}
\overline{c}_S(\alpha, p_N) &= \overline{c}_S(\alpha', p_N) \\
\overline{\epsilon}(\alpha, p_N) &= \overline{\epsilon}(\alpha', p_N) \\
\overline{\rho}(\alpha, p_N) &= \overline{\rho}(\alpha', p_N)
\end{aligned}$$

It follows that, for any mixed strategy p_N for nature, and any two mixed strategies p_S and p'_S such that $p_S(a_i) = p'_S(a_i)$ for all $i \leq m$:

$$\begin{aligned}
\overline{c}_S(p_S) &= \overline{c}_S(p'_S) \\
\overline{\epsilon}(p_S, p_N) &= \overline{\epsilon}(p'_S, p_N) \\
\overline{\rho}(p_S, p_N) &= \overline{\rho}(p'_S, p_N)
\end{aligned}$$

Hence, no generality is lost by restricting the scientist to the *finite* set of pure strategies $\{\alpha^i\}_{i \leq m}$, where α^i is the constant sequence T_i. Similarly, for fixed a mixed strategy p_S, the cost of a mixed strategy p_N for Nature is determined exclusively by the probabilities $p_N(C^{E_i})$ for $i \leq m$. In other words, if p_N and

p'_N are two probability measures such that $p_N(C^{E_i}) = p'_N(C^{E_i})$ for all $i \leq m$, then for a mixed strategy p_S for the scientist:

$$\begin{aligned} \overline{c}_N(p_N) &= \overline{c}_N(p'_N) \\ \overline{\epsilon}(p_S, p_N) &= \overline{\epsilon}(p_S, p'_N) \\ \overline{\rho}(p_S, p_N) &= \overline{\rho}(p_S, p'_N) \end{aligned}$$

Hence, again, no generality is lost by restricting Nature to the finite set of pure strategies $\{w^i\}_{i \leq m}$ where w^i is the constant sequence E_i. When Nature and the scientist are restricted to these strategies, the resulting game is a **simple**, finite quasi-game for incomparable goods, and therefore, has a quasi-Nash equilibrium by employing Proposition 3.1.1.[13]

Notice that, in the proof of Proposition 3.2.3, one could select any of an uncountable number of probability measures p_S and p_N for scientist and Nature respectively, and the two mixed strategies would still constitute a Nash equilibrium. However, the above proof indicates that there is a sense in which there is only **one** such equilibrium. Namely, if p_S and p_N do not assign uniform probabilities to the a_i's and C^{E_i}'s respectively, then their strategies will not constitute a Nash equilibrium. If p_S does not assign uniform probabilities to the a'_is, for instance, then there is at least one a_j that is assigned smaller (or equal) probability than all other a_i's. As such, Nature can cause the scientist to accrue greater losses (than had he played a uniform distribution on the a'_is) by picking a distribution p_N that assigns probability 1 to C^{E_j}. An entirely symmetric argument shows that Nature ought to place uniform probabilities on the events of the form C^{E_i}.

These remarks bolster the claim that G is structurally similar to the game matching pennies. In order to avoid being dominated, Nature must choose a strategy that assigns probability 1 to the set of worlds. When $\mathsf{Th} = \{T_1, \ldots, T_m\}$ is finite, therefore, there are, in a sense, only finitely many available options to Nature, namely, the possible probability assignments to the m-events C^{E_1} to C^{E_m} (where, again, $T_i = T_{E_i}$). And above, we argued that in any Nash equilibrium, Nature places a uniform distribution on these events. Similarly, in matching pennies, there is a single Nash equilibrium in which both players place uniform distributions on their possible actions. Moreover, if either player in matching pennies chooses a mixed strategy other than a uniform distribution on heads and tails, then there is a pure strategy for the second player that is a best response.

However, the assumption that Th is finite is absolutely necessary for the the existence of countably additive equilibria in G. When Th is infinite, not only does the above proof fail, but further, the proposition is demonstrably false:

Proposition 3.2.4. If Th is infinite, then G has no countably additive mixed strategy quasi-Nash equilibria.

[13] Strictly speaking, there preference relations of scientist and nature are not of the form specified in simple games, as they contain a lexicographic coordinate for convergence. However, this coordinate is constant for both players when they play un-dominated strategies by the arguments above, and so the two players' preference orders are isomorphic to simple ones.

Proof: An analogous argument to Proposition 3.2.2 works here. Let p_N be a mixed strategy for Nature. By Lemma 3.2.2, Nature maximizes $\overline{c}_N(p_N)$ by picking a mixed strategy p_N such that $p_N(W) = 1$. By Lemma 3.2.3, there is some finite collection of theories $T_1, \ldots, T_m$ such that the set of best replies for the scientist are measures such that $p_S(a) = 1$, where a is defined as in the previous lemma. Let $T \neq T_i$ for all $i \leq m$. Such a T exists, as Th is infinite. Let E_T be the effect set corresponding to T, and let η be the pure strategy for nature that consists in playing the constant sequence E_T. Clearly, $\overline{c}_N(\eta) = 0$. Moreover:

$$\overline{c}_S(p_S, \eta) < \overline{c}_S(p_S, \eta) = 1$$

And so, holding p_S fixed, Nature prefers η to p_N. This shows there are no countably additive mixed strategy Nash equilibria in G.

□

But there is a curious feature about the above proofs. Only the convergence and error components of the preference relations were used in the above arguments. Thus, the number of retractions committed by the scientist were totally irrelevant to the existence of Nash equilibria when the KGS model is represented as a game in any of the above ways. In particular, a quick examination of the Nash equilibria constructed above shows that a scientist is rewarded for rolling a die to determine a theory, and then dogmatically asserting the result for eternity! This suggests that the above model is not the correct way representation of inquiry. In the next sections, I represent the KGS model in a way that makes the costs of retractions relevant to assessing the desirability of particular actions.

3.2.2 G^*: Actions as Methods and Worlds

In the previous section, scientific inquiry was represented as a one-shot game between Nature and a scientist, where Nature and the scientist simultaneously play an infinite data sequence and an infinite sequence of theory guesses respectively. Clearly, this is a terrible model of inquiry: the idealized scientist in the above game does not form theories on the basis of evidence, but rather, decides *a priori* which theories she will guess for eternity. In this section, I describe a more realistic representation of inquiry as a strategic game in which the scientist's set of possible actions consists of *methods,* or functions from data sequences to theories. I call the resulting game G^*.

Let A_N be the set of possible actions for Nature as defined in the previous section. Define A_N^{fin} to be the set of finite initial segments of sequences in A_N, and let A_S^*, which is the set of possible actions for the scientist, be defined as follows:

$$A_S^* = \mathsf{Ans}^{A_N^{fin}}$$

In other words, A_S^* is the set of functions from finite effect sequences to answers. So a pure strategy for the scientist is a *method* or a mapping $M : A_N^{fin} \to \mathsf{Ans}$. In the same ways as the previous section, define a σ-algebra $\mathcal{A}_N$ on A_N.

Analogously, one can define a σ-algebra $\mathcal{A}_S^*$ on A_S^* as follows. For any $M \in A_S^*$, and $k \in \mathbb{N}$, define:

$$[M(\eta, k) = \alpha] = \{M' \in A_S^* \ : \ M'(\eta \upharpoonright k') = \alpha_{k'} \text{ for all } k' \leq k \text{ where } \alpha \in \mathsf{Ans}^k\}$$

Let $\mathcal{A}_S^* = \sigma(\{[M(\eta, k) = \alpha] : M \in A_S^*, k \in \mathbb{N}\})$ be the σ-algebra formed by taking the σ-closure of the events $[M(\eta, k) = \alpha]$. Then mixed strategies for the scientist and Nature are respectively measures on $\mathcal{A}_N$ and $\mathcal{A}_S^*$. In order to define the two player's preference relations in G^*, one can employ the preference relations for the game G as defined in Equation (II) as follows. Let $\alpha^{M,\eta} \in \mathsf{Ans}^\omega$ be the answer sequence such that $\alpha_n^{M,\eta} = M(\eta \upharpoonright n)$ for all $n \in \mathbb{N}$. In other words, $\alpha^{M,\eta}$ is the answer sequence that is produced when the scientist plays M and Nature plays η. Then, one can identify the errors and retractions committed by M in response to η in G^* as those that would be committed by $\alpha^{M,\eta}$ in response to η in G. Formally, for all $M \in A_S^*$, $\eta \in A_N$, and $n \in \mathbb{N}$ define:

$$\begin{aligned}
\epsilon(M, \eta, n) &= \overline{e}(\alpha^{M,\eta}, \eta, n) \\
\rho(M, \eta, n) &= \overline{r}(\alpha^{M,\eta}, \eta, n) \\
c_S(M, \eta) &= c_S(\alpha^{M,\eta}, \eta)
\end{aligned}$$

In the same way as for G, one can extend the cost functions c_S, e, and r on pure strategies to functions c_S, $\overline{\epsilon}$, and $\overline{\rho}$ on mixed strategies as follows. Let p_S^* be a measure on $\mathcal{A}_S^*$ and p_N be a measure on $\mathcal{A}_N$. Then define:

$$\begin{aligned}
\overline{\epsilon}(p_S^*, p_N) &= \int_{A_N} \int_{A_S^*} \sum_{n=0}^{\infty} \overline{e}(M, \eta, n) \cdot p_N(d\eta) \cdot p_S^*(dM) \\
\overline{\rho}(p_S^*, p_N) &= \int_{A_N} \int_{A_S^*} \sum_{n=0}^{\infty} \overline{\rho}(M, \eta, n) \cdot p_N(d\eta) \cdot p_S^*(dM) \\
\overline{c}_S(p_S^*) &= \int_{A_N} \int_{A_S^*} c_S(\alpha^{M,\eta}, \eta) p_N(d\eta) \cdot p_S^*(dM)
\end{aligned}$$

Again, note the order of integration in the above definitions. In contrast to G, notice that the scientist's retraction costs in G^* **do** depend upon Nature's strategy η. Therefore, the above cost functions are a more realistic representation of scientific inquiry than those defined in the game G. Now that we have defined the error, retraction, and convergence cost functions on for mixed strategies, we can now use the ordering defined in Equation (II) to define the scientist's and Nature's respective preference orderings. This completes the second formalization of the KGS model as a game. Call this game G^*.

Before characterizing equilibria in G^*, three notes are in order. First, in later discussions, it will be helpful to evaluate the expected number of retractions and errors of a given mixed strategy p_S^* for the scientist (with respect to some measure p_N) *over a fixed complexity class.* To this end, for all $n \in \mathbb{N}$, define:

$$\begin{aligned}
\overline{\epsilon}_n(p_S^*, p_N) &= \int_{\mathsf{Comp}_n} \int_{A_S^*} \sum_{n=0}^{\infty} \overline{e}(M, \eta, n) \cdot p_N(d\eta) \cdot p_S^*(dM) \\
\overline{\rho}_n(p_S^*, p_N) &= \int_{\mathsf{Comp}_n} \int_{A_S^*} \sum_{n=0}^{\infty} \overline{\rho}(M, \eta, n) \cdot p_N(d\eta) \cdot p_S^*(dM)
\end{aligned}$$

Second, when p_S^* and p_N are both countably additive measures, a second, cleaner characterization of costs is available. Define a mapping $\varphi^* : A_S^* \times A_N \to A_S \times A_N$ as follows:

$$\varphi^*(M, \eta) = (\alpha^{M,\eta}, \eta)$$

Proposition 3.2.5. The function φ^* is $\sigma(\mathcal{A}_S^* \times \mathcal{A}_N)/\sigma(\mathcal{A}_S \times \mathcal{A}_N)$-measurable.

Proof: For brevity, define:

$$\begin{aligned}
\mathcal{G} &= \{[\alpha \restriction m] \ : \ \alpha \in A_S\} \\
\mathcal{H} &= \{[\eta \restriction m] \ : \ \eta \in A_N\} \\
\mathcal{G}^* &= \{[M(\eta, m) = \alpha] \ : \ M \in A_S^*, \eta \in A_N, \alpha \in A_S, m \in \mathbb{N}\}
\end{aligned}$$

Then by Schilling (2005) Theorem 13.3, it follows that $\sigma(\mathcal{A}_S \times \mathcal{A}_N) = \sigma(\mathcal{G} \times \mathcal{H})$ and $\sigma(\mathcal{A}_S^* \times \mathcal{A}_N^*) = \sigma(\mathcal{G}^* \times \mathcal{H})$. Hence, by **?** Theorem 7.2, it suffices to show that for any $g \in \mathcal{G}$ and $h \in \mathcal{H}$, the set $\varphi^{-1}(g \times h)$ is $\sigma(\mathcal{A}_S^* \times \mathcal{A}_N)$-measurable. This is confirmed by the following calculation:

$$\begin{aligned}
\varphi^{-1}([\alpha \restriction m_1] \times [\eta \restriction m_2]) &= \varphi^{-1}(\{(\alpha', \eta') \ : \ \alpha' \restriction m_1 = \alpha \restriction m_1 \text{ and } \eta' \restriction m_2 = \eta \restriction m_2\}) \\
&= \{(M, \eta') \in A_S^* \times A_N \ : \ \alpha^{M,\eta} \restriction m_1 = \alpha \restriction m_1 \text{ and } \eta' \restriction m_2 = \eta \restriction m_2\} \\
&= [M(\eta, m_1) = \alpha] \times [\eta \restriction m_2]
\end{aligned}$$

This event is not only in $\sigma(\mathcal{A}_S^* \times \mathcal{A}_N^*) = \sigma(\mathcal{G}^* \times \mathcal{H}^*)$, but moreover, it is a member of the generator set $\mathcal{G}^* \times \mathcal{H}$ itself!

□

Thus, when p_S^* and p_N are countably additive measures on $\mathcal{A}_S^*$ and $\mathcal{A}_N$ respectively, Proposition 3.2.5 motivates the following approach to evaluating mixed strategies. Let p^* be the induced product measure on $\sigma(\mathcal{A}_S^* \times \mathcal{A}_N)$. Let p be the measure induced on $\sigma(\mathcal{A}_S \times \mathcal{A}_N)$ by the mapping φ. That is, for all $X \in \sigma(\mathcal{A}_S \times \mathcal{A}_N)$, define:

$$p(X) = p^*(\varphi^{-1}(X))$$

The function p is a probability measure because φ is measurable, in light of proposition 3.2.5. Finally, for $X \in \mathcal{A}_S$ and $Y \in \mathcal{A}_N$, define:

$$\begin{aligned}
p_S(X) &= p(X \times A_N) \\
p_N(Y) &= p(A_S \times Y)
\end{aligned}$$

Then one can use the preference relations defined in Equation II to order mixed strategies in this new game as follows:

$$\begin{aligned}(p_S^*, p_N) \preceq_S (q_S^*, q_N) &\Leftrightarrow (p_S, p_N) \preceq_S (q_S, q_N)\\ (p_S^*, p_N) \succeq_N (q_S^*, q_N) &\Leftrightarrow (p_S, p_N) \preceq_S (q_S, q_N)\end{aligned}$$

Again, the reader is warned that these two approaches for evaluating costs of mixed strategies are only equivalent when the measures are countably additive.

A third, more troublesome issue arises concerning the definition of convergence costs above. To explain the problem, it will be helpful to have a list of measurable events handy:

Lemma 3.2.4. The set $\mathcal{M}_c = \{M \in A_S^* \ : \text{ for all } \eta \in W, M \text{ converges in } \eta\}$ of methods that converge is countable.

Proof: Routine. The assertion follows from the fact that a countable union of countable sets is countable. □

Proposition 3.2.6. The following events are measurable:

1. For any $A \in \mathsf{Ans}$ and any $m \in \mathbb{N}$, the event $[M(\eta \upharpoonright n) = A] = \{M \in A_S^* : M(\eta \upharpoonright n) = A\}$ is $\mathcal{A}_S^*$-measurable.
2. For any $M \in A_S^*$, the singleton $\{M\}$ is $\mathcal{A}_S^*$-measurable.
3. The set $\mathcal{M}_{c,w} = \{M \in A_S^* \ : \ M \text{ converges in } w\}$ is $\mathcal{A}_S^*$-measurable.
4. For any set of worlds $W_0 \subseteq W$, the set $\mathcal{M}_{c,W_0} = \{M \in A_S^* \ : \ M \text{ converges in every } w \in W_0\}$ is $\mathcal{A}_S^*$-measurable. In particular, the set $\mathcal{M}_c = \{M \in A_S^* \ : \ M \text{ converges in } w \text{ for all } w \in W\}$ is $\mathcal{A}_S^*$-measurable.
5. The set $\mathcal{M}_{Ock} = \{M \in A_S^* \ : \ M \text{ is Ockham } \}$ is $\mathcal{A}_S^*$-measurable.

Proof Sketch: To write the above events as countable unions and intersections of known measurable events, one can simply replace every universal quantifier in the above definitions with an intersection symbol and every existential quantifier with a union. For example, a method M converges if $(\forall w \in W)(\exists n \in \mathbb{N})(\forall m \geq n)[M(w \upharpoonright m) = T_w]$, and so one can write this as a union of known measurable events (using (1)) as $\bigcap_{w \in W} \bigcup_{n \in \mathbb{N}} \bigcap_{m \geq n} [M(w \upharpoonright m) = T_w]$. Hence, when the quantifiers in the definitions of the above events range over countable sets (which is guaranteed by the previous lemma), the above events are measurable in their respective σ-algebras.

□

Given the above lemma, we can now describe a philosophical worry concerning the definition of G^*. In every area of empirical science, practitioners must make inferences from data that are known to have randomly distributed errors; such error may be due to experimental error, lack of precision in measuring devices, or simply indeterminism in the phenomena under investigation. Regardless of the source of the error in the data, scientists often require that the methods that they employ are **statistically consistent**. What is statistical consistency? In many common problems, there is some metric on the set of theories, which provides a quantitative measure of how much two theories T_1 and T_2 differ. For example, suppose one is interested in estimating the mean of a normal distribution with known variance of one. Then one can say that two different theories T_1 and T_2, which here are represented by real numbers r_1 and r_2, are $|r_1 - r_2|$ close to one another. A method is said to be statistically consistent if, for every $\epsilon > 0$, the probability that the method produces a theory T that differs from the true theory T_0 by more than ϵ approaches zero as the amount of data (usually quantified in terms of sample size) approaches infinity. In other words, a method for guessing theories from data is statistically consistent if it produces theories closer and closer to the truth with probability approaching one as time elapses.

When the scientist and Nature employ countably additive mixed strategies in G^*, then the convergence cost $\overline{c}_S(p^*S, p_N)$ of employing some mixed strategy p^*_S is exactly zero if and only if the scientist is statistically consistent with respect to the worlds in the support of p_N. This is expressed by the following lemma:

Lemma 3.2.5. Suppose p_N and p^*_S are countably additive measures on $\mathcal{A}_N$ and $\mathcal{A}^*_S$ respectively. Further, suppose that $W_0 \subseteq W$ is the support of p_N. Then the following are equivalent:

1. $\overline{c}_S(p^*_S, p_N) = 0$
2. $p^*_S(\mathcal{M}_{c,W_0}) = 1$
3. $\lim_{n\to\infty} p^*_S([M(w \upharpoonright n) = T_w]) = 1$ for all $w \in W_0$

Proof: Apply Lemma 3.2.2.

□

That is, when the measures under consideration are countably additive, any mixture of convergent methods is statistically consistent, and any statistically consistent mixed strategy is a mixture of convergent methods. However, this equivalence fails to hold when the scientist's methods are only finitely additive. To see why, consider a finitely additive strategy for the scientist that does the following.[14] The scientist first fixes a convergent method M. Then the scientist rolls an infinite-sided die with sides labeled $1, 2, 3, \ldots$ and so on. Here, because we are interested in representing finitely-additive mixed strategies, imagine that the die has equal probability of landing on any given side. Thus, the probability that the die lands on any particular side is zero, even though the probability that it lands on at least one of the sides is one. If the die lands on side n, then the scientist then says '?' until time n, at which point he begins employing the method M. Call this mixed strategy p^*_S.

[14] Thanks to Teddy Seidenfeld for suggesting the following example.

Given this mixed strategy, for any fixed natural number $n \in \mathbb{N}$, the probability that the scientist will have guessed some theory T other than '?' before time n is zero. So the mixed strategy is clearly not statistically consistent, under any reasonable metric of closeness of theories. However, it's easy to check that $\overline{c}_S(p_S^*, p_N) = 0$ regardless of the mixed strategy employed by Nature. Hence, if one wishes to penalize the scientist for employing methods that fail to be statistically consistent, then above cost functions will simply not do.

One might object that the described mixed strategy converges **almost surely**, which is to say that it assigns probability one to convergent methods. Almost sure convergence, moreover, is often considered to be a virtue of scientific methods, and in many circumstances, it is considered to be more valuable than convergence in probability. One reason that almost sure convergence is prized, however, is that it implies convergence in probability when one's measure is countably additive. That is, in many practical applications, almost sure convergence is strictly stronger than convergence in probability. However, as is exhibited by the mixed strategy p_S^*, almost sure convergence does **not** imply convergence in probability when one's measure is purely finitely additive. Because, I assume, most scientists value methods that will provide some information in some finite amount of time, I assume that convergence is probability is a more realistic representation of cost in scientific inquiry in general.

With these arguments in mind, one can adjust the preference relations of the scientist and Nature as follows. Say a mixed strategy p_S^* for the scientist is **convergent in probability** if for every $w \in W$:

$$\lim_{n \to \infty} p_S^*([M(w \upharpoonright n) = T_w]) = 1$$

Then define a preference relation $\preceq_S$ for the scientist as follows:

$$\text{(III)} \quad (p_S^*, p_N) \preceq_S (q_S^*, p_N') \Leftrightarrow \begin{cases} \overline{c}_N(p_N) < \overline{c}_N(p_N') \text{ or} \\ \\ \overline{c}_N(p_N) = \overline{c}_N(p_N') \text{ and } q_S^* \text{ converges in probability and } p_S^* \text{ does not, or} \\ \\ \overline{c}_N(p_N) = \overline{c}_N(p_N'), \text{ and both } p_S^* \text{ and } q_S^* \text{ converge in probability, and} \\ \overline{\epsilon}(p_S^*, p_N) \geq \overline{\epsilon}(q_S^*, p_N') \text{ and} \\ \overline{\rho}(p_S^*, p_N) \geq \overline{\rho}(q_S^*, p_N') \end{cases}$$

From here onward, I shall refer to the game with these preference relations as G^*. It should be emphasized that there are many finitely additive mixed strategies that also converge in probability. Hence, levying a tax on the scientist for not converging in probability does nothing to inhibit him from employing finitely additive mixed strategies.

It is an open question whether, in general, G^* has any countably additive quasi-Nash equilibria. When K contains an infinite path, the game G^* does have finitely-additive mixed strategy equilibria, which is proven below. Before embarking on the proof, however, an informal outline of the proof will be helpful. Suppose Nature has two dice, both of which have countably many sides. Label the sides of the dice $1, 2, 3, \ldots$. Suppose the first die is countably additive with respect to which of the sides on which it lands, but the second is not. In fact, for any natural number n, the probability that the second die lands on a number less than or equal to n is zero (and hence, the probability that it lands on a side numbered $n+1$ or greater is one). Nature throws the first die to determine the complexity of the world that she plays. If the first die lands on n, she then throws the second, finitely-additive die, n-times to determine when (i.e. the stages of inquiry) to present some new set of effects that make the world appear to be in the next highest complexity class. In other words, suppose the rolls of the second die are $m_1, m_2, \ldots m_n$. We may assume that $m_1 < m_2 < \ldots m_n$, as Nature can redo the j^{th} roll if the number rolled is not greater than previous ones. Then Nature picks a world in complexity class n that (i) resembles a zero complexity world until time m_1, (ii) resembles a one-complexity world until time m_2, (iii) resembles a three-complexity world until time m_3, and so on. Notice that because the second die is finitely additive, the probability that the $j+1^{st}$ set of effects will be presented "infinitely late" after the j^{th} set of effects is one.

Now let p_S be a mixed strategy for the scientist that is convergent in probability. Then as the first set of effects occur "infinitely late", the scientist guesses with near unit probability the theory of the zero-complexity world before he sees the first set of effects. Once he sees the first set of effects, however, he begins to converge towards the theory of the one-complexity world. Now, because the second set of effects occur "infinitely late" after the first set, the scientist guesses with near unit probability the theory of the one-complexity world before he sees the second set of effects. And so on. This suggests that the scientist retracts at least n-times in when nature plays according to the strategy p_N and selects a world in the n^{th} complexity class. If the countably additive die employed by Nature assigns approximately $\frac{1}{2^n}$ probability to picking a world in the $(2^n)-th$ complexity class, then a St. Petersburg-like argument proves that Nature forces the scientist to accrue an infinite number of expected retractions and errors. The extensive technical details in the proof below mainly concern the computation of probabilities of the relevant events governed by the finitely additive die.

The proof that G^* has finitely-additive mixed strategy equilibria is quite involved. We first prove a shorter proposition, which itself requires a new definition. Let $\langle K_j \rangle_{j \leq n+1}$ be a path of length $n+1$ in K. For every measure on $\mathcal{A}^*_S$ that is convergent in probability, and every $\epsilon, x \in (0,1]$ such that $x < 1 - (1 - \frac{\epsilon}{n})^{1/n}$. Define:

$$\begin{aligned}\mathsf{Ret}_n(p^*_S, \langle K_j \rangle_{j \leq n+1}, \epsilon, x) \quad &= \quad \{w \in \mathsf{Comp}_n : \exists m_0 < m_1 < \ldots < m_n \in \mathbb{N} \\ &\qquad \forall j \leq n (p^*_S([M(w \restriction m_j) = T_{K_j}]) > 1 - x)\}\end{aligned}$$

That is, $\mathsf{Ret}_n(p^*_S, \langle K_j \rangle_{j \leq n+1}, \epsilon, x)$ is the set of worlds such that p^*_S first produces theory T_{K_0} at some stage of inquiry m_0 with probability at least $1-x$, and then produces theory T_{K_1} at some later stage m_1 with probability at least $1-x$, and so on. The following proposition shows that every probability measure p_N assigning $\mathsf{Ret}_n(p^*_S, \epsilon, x)$ unit probability causes the scientist to accrue almost n-retractions in complexity class n. For ease of notation, when the path $\langle K_j \rangle_{j \leq n+1}$

and the number $x \in (0,1]$ are held fixed and obvious from context, we will write $\mathsf{Ret}_n(p^*_S, \epsilon)$.

Proposition 3.2.7. Let p^*_S be a measure on $\mathcal{A}^*_S$ that is convergent in probability. Let n be a natural number, and suppose that $\langle K_j \rangle_{j \leq n+1}$ is a path of length $n+1$ in K, such that $K_0 \in \min_K(\emptyset)$. Let $\epsilon, x \in (0,1]$ be such that $x < 1 - (1 - \frac{\epsilon}{n})^{1/n}$. Then for every p_N on A_N such that $p_N(\mathsf{Ret}_n(p^*_S, \epsilon, x)) = 1$, it follows that:

$$\overline{\rho}_n(p^*_S, p_N) > n - \epsilon$$

Proof: Let $\epsilon, x \in (0,1]$ be such that $x < 1 - (1 - \frac{\epsilon}{n})^{1/n}$. One can check this ensures that $(1-x)^n > 1 - \frac{\epsilon}{n}$

First, we prove that $\mathsf{Ret}_n(p^*_S, \epsilon)$ is non-empty. Consider the world $w^0 = K_0^\infty$ that is constantly K_0. As p^*_S is convergent in probability, there is some $m_{0,\epsilon}$ such that for all $m \geq m_{0,\epsilon}$:

$$p^*_S([M(w^0 \upharpoonright m) = T_{K_0}]) > 1 - x$$

Now define a world w^1 such that

$$w^1_j = \begin{cases} K_0 \text{ if } j \leq m_{0,\epsilon} \\ K_1 \text{ if } j > m_{0,\epsilon} \end{cases}$$

Then, again as p^*_S is convergent in probability, there is some $m_{1,\epsilon}$ such that for all $m \geq m_{1,\epsilon}$:

$$p^*_S([M(w \upharpoonright m) = T_{K_1}]) > 1 - x$$

Now define a world w^1 such that

$$w^2_j = \begin{cases} K_0 \text{ if } j \leq m_{0,\epsilon} \\ K_1 \text{ if } m_{0,\epsilon} < j \leq m_{1,\epsilon} \\ K_2 \text{ if } j > m_{1,\epsilon} \end{cases}$$

Continuing in this way, we can define $m_{0,\epsilon} < m_{1,\epsilon} < \ldots < m_{n,\epsilon} \in \mathbb{N}$ and world w^n such that for all $j \leq n$:

$$p^*_S([M(w^n \upharpoonright m_{j,\epsilon}) = T_{K_j}]) > 1 - x$$

It follows that $w^n \in \mathsf{Ret}_{p^*_S, \epsilon}$, and so $\mathsf{Ret}_{p^*_S, \epsilon}$ is non-empty as claimed.

Now we show that for all measures p_N on A_N such that $p_N(\mathsf{Ret}_n(p^*_S, \epsilon)) = 1$:

$$\overline{\rho}(p^*_S, p_N) > n - \epsilon$$

Let $w \in \mathsf{Ret}_n(p^*_S, \epsilon)$. By definition of $\mathsf{Ret}_n(p^*_S, \epsilon)$, it follows that there are natural numbers $m_{0,w} < m_1 \ldots < m_{n,w}$ such that $p^*_S([M(w \upharpoonright m_{j,w}) = T_{K_j}]) > 1 - x$ for all $j \leq n$. It follows that:

$$\begin{aligned} p^*_S(\bigcap_{j \leq n} [M(w \upharpoonright m_{j,w}) = T_{K_j}]) &\geq \Pi_{j \leq n} \, p^*_S([M(w^n \upharpoonright m_{j,w}) = T_{K_j}]) \\ &> (1-x)^n \\ &> 1 - \frac{\epsilon}{n} \end{aligned}$$

Here, the last inequality follows from the choice of x. Define:

$R_{n,w} := \{M \in A^*_S : \exists q_1 < q_2 \ldots < q_n \in \mathbb{N} \forall j \leq (m_{j,w} < q_j \leq m_{j+1,w} \,\&\, r(M, w, q_j) = 1)\}$

That is, $R_{n,w}$ is the set of methods that retract at least once between $m_{j,w}$ and $m_{j+1,w}$ in w for all $j \leq n$. Notice that $\bigcap_{j \leq n}[M(w \upharpoonright m_{j,w}) = T_{K_j}] \subseteq R_{n,w}$. It immediately follows that:

$$p^*_S(\bigcap_{j \leq n}[M(w \upharpoonright m_{j,w} = T_{K_j}]) \leq p^*_S(R_{n,w})$$

Finally, by definition of $R_{n,w}$, for all $M \in R_{n,w}$, It follows that:

$$\sum_{j=0}^{\infty} \rho(M, w, j) \geq n$$

All that remains is to compute the quantity $\overline{\rho}_n(p^*_S, p_N)$:

$$\begin{aligned}
\overline{\rho}_n(p^*_S, p_N) &= \int_{\mathsf{Comp}_n} \int_{A^*_S} \sum_{j=0}^{\infty} \rho(M, w, j) \cdot p_N(dw) \cdot p^*_S(dM) \\
&\geq \int_{\mathsf{Ret}_n(p^*_S, \epsilon)} \int_{R_{n,w}} \sum_{j=0}^{\infty} \rho(M, w, j) \cdot p^*_S(dM) \cdot p_N(dw) \\
&\geq \int_{\mathsf{Ret}_n(p^*_S, \epsilon)} n \cdot \left(\int_{R_{n,w}} p^*_S(dM) \right) \cdot p_N(dw) \\
&= \int_{\mathsf{Ret}_n(p^*_S, \epsilon)} n \cdot p^*_S(R_{n,w}) \cdot p_N(dw) \\
&\geq \int_{\mathsf{Ret}_n(p^*_S, \epsilon)} n \cdot p^*_S(\bigcap_{j \leq n}[M(w \upharpoonright m_{j,w} = T_{K_j}])) \cdot p_N(dw) \\
&> \int_{\mathsf{Ret}_n(p^*_S, \epsilon)} n \cdot (1 - \frac{\epsilon}{n}) \cdot p_N(dw) \\
&\geq \int_{\mathsf{Ret}_n(p^*_S, \epsilon)} (n - \epsilon) \cdot p_N(dw) \\
&= n - \epsilon
\end{aligned}$$

□

Theorem 3.2.1. Suppose K contains an infinite path. Then the game G^* has finitely-additive mixed strategy equilibria.

Proof: The proof requires first introducing a considerable number of definitions and notational conventions. First, given any set R, define $cof(R) \subset 2^R$ be the set of co-finite subsets of R. For any positive natural number n, define $\mathbb{N}^{[n]}$ to be the set of all increasing n-tuples of natural numbers. In symbols, for all $n \in \mathbb{N}$:

$$\mathbb{N}^{[n]} = \{s \in \mathbb{N}^n : \forall i < n - 1(s_i < s_{i+1})\}$$

Note that when $n = 0$ the set $\mathbb{N}^{[n]}$ is the empty sequence $\langle\rangle$. For any subset $S \subseteq \mathbb{N}^{[n]}$ and any i such that $i < n$, define:

$$\begin{aligned}
S_i &= \{s \upharpoonright i \ : \ s \in S\} \\
\mathbb{N}(s, S, i) &= \{m \in \mathbb{N} : (s \upharpoonright i) * m \in S_{i+1}\} \\
\mathsf{CofExt}_n &= \{S \subseteq \mathbb{N}^{[n]} \ : \ \exists S' \subseteq \mathbb{N}^{[n]} \forall s \in S' \forall i < n(\mathbb{N}(s, S, i) \in cof(\mathbb{N}))\}
\end{aligned}$$

Here, CofExt_n is a mnemonic for "co-finitely extendable."

By assumption, K contains an infinite path. Denote this path by $\langle K_n \rangle_{n \in \mathbb{N}}$. Now, given an increasing n-tuple $s \in \mathbb{N}^{[n]}$, define w^s to be the world in Comp_n such that

$$w^s_j = \begin{cases} K_0 \text{ if } j < s_0 \\ K_i \text{ if } \exists i < n(s_{i-1} \leq s_i) \\ K_n \text{ if } j \geq s_{n-1} \end{cases}$$

In other words, w^s is the world that begins by presenting the effect set K_0 until the s_0^{th} stage of inquiry, at which point it presents K_1 and continues to present K_1 until the s_1^{th} stage of inquiry, at which point it presents K_2, and so on. When s is the empty sequence $\langle\rangle$, the world w^s is the constant sequence $(K_0)^\infty$. Given $s \in \mathbb{N}^{[n]}$ and $i \in \mathbb{N}$, recall $s * i$ is the result of concatenating i to the end of the sequence s (i.e. $s * i = \langle s_0, s_1, \ldots, s_{m-1}, i\rangle$). Notice that $s * i$ is an element of $\mathbb{N}^{[n+1]}$ if and only if $s \in \mathbb{N}^{[n]}$ and $i > s_{n-1}$. Accordingly, for each positive natural number n, $s \in \mathbb{N}^n$, and $I \subseteq \mathbb{N}$, define:

$$\begin{aligned}
s * I &= \{s * i : i \in I \text{ and } i > s_{n-1}\} \\
W^{s*I} &= \{w^{s*i} \in W_K \ : \ i \in I\}
\end{aligned}$$

Notice that W^{s*I} is always a subset of the $(n+1)^{st}$ complexity class when $s \in \mathbb{N}^n$ (though it may be empty if there is no $i \in I$ such that $i > s_n$).

For each positive natural number n and each $s \in \mathbb{N}^{[n-1]}$, define:

$$\begin{aligned}
C_n[s] &= W^{s*\mathbb{N}} \\
C_n &= \{w^s : s \in \mathbb{N}^{[n]}\} = \bigcup_{s \in \mathbb{N}^{[n-1]}} C_n[s] \\
Y_n &= \{\{w^s \ : \ s \in S\} \subseteq \mathsf{Comp}_n \ : \ S \in \mathsf{CofExt}_n\} \\
Z_n &= \{C_n \setminus U \ : \ U \in Y_n\} \\
B_n &= Y_n \cup Z_n
\end{aligned}$$

When $n = 0$, define $C_0 = Y_0 = \{(K_0)^\infty\}$, and let $Z_0 = \{\emptyset\}$ so that $B_0 = \{(K_0)^\infty, \emptyset\}$. We will need the following two lemmas, whose proofs are postponed for exposition.

Lemma 3.2.6. For all natural numbers $n \in \mathbb{N}$, the set B_n is an algebra on C_n.[15]

[15] See Appendix for definition of algebra. In set theory and logic, the measure-theoretic definition of algebra offered in the appendix is equivalent to the definition of a Boolean algebra on the power set of some set Ω, where meets and joins correspond to the set-theoretic operations of intersection and union.

Lemma 3.2.7. Let $\langle p^n \rangle_{n\in\mathbb{N}}$ be a sequence of finitely additive probability measures on same underlying measurable space $(\Omega, \mathcal{F})$, and let $\langle r_n \rangle_{n\in\mathbb{N}}$ be a sequence of real numbers such that $\sum_{n\in\mathbb{N}} r_n = 1$. Then $p := \sum_{n\in\mathbb{N}} r_n \cdot p^n$ is a probability measure on $(\Omega, \mathcal{F})$.

Because K contains an infinite branch, the algebra B_n is non-empty for each $n \in \mathbb{N}$. Let $p^n : B_n \to [0,1]$ be a probability measure on B_n such that $p^n(U) = 1$ for all $U \in Y_n$ (and, hence, $p^n(U) = 0$ for all $U \in Z_n$). Notice that p^n is purely finitely additive for all natural numbers n. Let B be the smallest algebra containing $\cup_{n\in\mathbb{N}} B_n$. One can see that, because each of the B_n's is itself an algebra and the B_n's are disjoint, B is simply the result of closing $\cup_{n\in\mathbb{N}} B_n$ under finite unions. Hence, every event $U \in B$ can be represented as a finite union $U = U_{i_1} \cup U_{i_2} \ldots \cup U_{i_n}$ for some set of natural numbers $\{i_1, \ldots, i_n\}$ such that $U_{i_j} \in B_{i_j}$. For all $n \in \mathbb{N}$, extend p^n to a probability measure $\overline{p}^n$ on B by defining:

$$\overline{p}^n(U) = \begin{cases} p^n(U) \text{ if } U \in B_n \\ 0 \text{ otherwise} \end{cases}$$

Define a function $g : \mathbb{N} \to [0,1]$ as follows:

$$g(n) = \begin{cases} \frac{1}{8} & \text{if } n \in \{0,1\} \\ \frac{1}{2^{d+1}} & \text{if } n = 2^d \text{ for some } d \in \mathbb{N}, \\ \frac{1}{2^{d+1}(2^d-1)} & \text{if } 2^d < n < 2^{d+1} \text{ for some } d \in \mathbb{N} \end{cases}$$

It's easy to check that g is actually a countably additive probability measure on the power set of natural numbers. Now, for any $U = U_{i_1} \cup U_{i_2} \ldots U_{i_n}$ in B, define:

$$p(U) = g(i_1) \cdot \overline{p}^{i_1}(U_{i_1}) + g(i_2) \cdot \overline{p}^{i_2}(U_{i_2}) + \ldots + g(i_n) \cdot \overline{p}^{i_n}(U_{i_n})$$

Then, by Lemma 3.2.7, p is a probability measure on B. By Lemma 3.2.6 and Theorem 4.1.2 in the appendix, the function p extends to a finitely additive probability measure p_N on A_N. Notice that, for all $n \in \mathbb{N}$, the measure $p_N \upharpoonright B_n$ is equal to $g(n) \cdot p^n$, as each of the algebras B_n are disjoint.

The following lemma is the major step to proving the theorem:

Lemma 3.2.8. Let p^*_S be a measure on A^*_S that converges in probability. Then for all natural numbers $n \in \mathbb{N}$:

$$\overline{\rho}_n(p^*_S, p_N) \geq n \cdot g(n)$$

Proof: When $n = 0$, the proof is trivial as retractions are always greater than or equal to zero. So suppose $n > 0$.

First, we introduce one more piece of notation. Note that because p^*_S is convergent in probability, there exists a function $\gamma(p^*_S, -, -) : W_K \times (0,1] \to \mathbb{N}$ such that:

$$\gamma(p^*_S, w, \epsilon) = \min\{m \in \mathbb{N} \ : \ p^*_S([M(w \upharpoonright m')] = T_w) > 1 - \epsilon \text{ for all } m' \geq m\}$$

In general, if $\mathcal{P}(\mathcal{A}^*_S)_{cp}$ is the set of probability measures on $\mathcal{A}^*_S$ that converge in probability, one can consider γ as a function with domain $\mathcal{P}(\mathcal{A}^*_S)_{cp} \times W_K \times (0,1]$. Importantly, notice that for any triple $\langle p^*_S, w, \epsilon \rangle$, the set

$$\Gamma(p^*_S, w, \epsilon) := \mathbb{N} \setminus \{0, 1, \ldots, \gamma(p^*_S, w, \epsilon)\}$$

is co-finite in $\mathbb{N}$. Now:

$$\begin{aligned} \overline{\rho}_n(p^*_S, p_N) &= \int_{\mathsf{Comp}_n} \int_{A^*_S} \sum_{m=0}^{\infty} \rho(M, w, m) \cdot p^*_S(dM) \cdot p_N(dw) \\ &= g(n) \cdot \int_{\mathsf{Comp}_n} \int_{A^*_S} \sum_{m=0}^{\infty} \rho(M, w, m) \cdot p^*_S(dM) \cdot p^n(dw) \end{aligned}$$

where the last equality follows by the definition of p_N. Hence, it suffices to show that:

$$\int_{\mathsf{Comp}_n} \int_{A^*_S} \sum_{m=0}^{\infty} \rho(M, w, m) \cdot p^*_S(dM) \cdot p^n(d\eta) \geq n.$$

Let $\epsilon \in [0,1)$ and $x \in (0,1]$ be such that $(1-x)^{n+1} > 1 - \frac{\epsilon}{n+1}$ and define $\mathsf{Ret}_n(p^*_S, \langle K_j \rangle_{j \leq n} \epsilon, x)$ as in Proposition 3.2.7. Recall that $\mathsf{Ret}_n(p^*_S, \langle K_j \rangle_{j \leq n}, \epsilon, x)$ is the set of worlds in complexity class n where the mixed strategy p^*_S produces T_{K_0} with probability greater than $1-x$ at some point of inquiry m_1, then produces T_{K_1} with probability greater than $1-x$ at some later point of inquiry $m_1 > m_0$, and so on. Again, we write $\mathsf{Ret}_n(p^*_S, \epsilon)$ dropping the path parameter, as it is held fixed, and x because it is a function of ϵ.

It suffices to show that $p^n(\mathsf{Ret}_n(p^*_S, \epsilon)) = 1$, for then:

$$\begin{aligned} \int_{\mathsf{Comp}_n} \int_{A^*_S} \sum_{m=0}^{\infty} \rho(M, w, m) \cdot p^*_S(dM) \cdot p^n(d\eta) &\geq \int_{\mathsf{Ret}_n(p^*_S, \epsilon, x)} \int_{A^*_S} \sum_{m=0}^{\infty} \rho(M, w, m) \cdot p^*_S(dM) \cdot p^n(d\eta) \\ &> n - \epsilon \end{aligned}$$

where the last inequality follows by Proposition 3.2.7. As ϵ was chosen arbitrarily, the result would follow. But to show $p^n(\mathsf{Ret}_n(p^*_S, \epsilon)) = 1$, by the definition of p^n, it suffices to show that $U \subseteq \mathsf{Ret}_n(p^*_S, \epsilon)$ for some $U \in Y_n$, as $p^n(U) = 1$ for all $U \in Y_n$.

The proof proceeds by induction on n, and the base case is trivial. So suppose there is $U \in Y_n$ such that $U \subseteq \mathsf{Ret}_n(p^*_S, \epsilon)$. We want to find some $V \in Y_{n+1}$ such that $V \subseteq \mathsf{Ret}_{n+1}(p^*_S, \epsilon)$. By definition of Y_n, there exists $S_U \in \mathsf{CofExt}_n$ such that

$$\{w^s \in W_K \ : \ s \in S_U\} \subseteq U.$$

Recall $\Gamma(p^*_S, w^t, x) = \mathbb{N} \setminus \{0, 1 \ldots, \gamma(p^*_S, w^t, x)\}$. With that in mind, define $S_V = \bigcup_{s \in S_U} s * \Gamma(p^*_S, w^s, x)$ and define

$$V = \{w^s \in W_K \ : \ s \in S_U\}.$$

By definition of CofExt_n, we have that $\mathbb{N}(s, S_U, i) \in cof(\mathbb{N})$ for all $s \in S_U$ and all $i < n$. But notice that $(S_V)_i = (S_U)_i$ for all $i \leq n$, and hence, $\mathbb{N}(s, S_V, i) =$

$\mathbb{N}(s, S_U, i)$ for all $i < n$. It follows that $\mathbb{N}(s, S_V, i) \in cof(\mathbb{N})$ for all $s \in S_V$ and all $i < n$. Moreover, $\mathbb{N}(s, S_V, n)$ is co-finite for all $s \in S_V$ because (i) $\Gamma(p_S^*, w, x)$ is co-finite for any world w, and (ii) $\mathbb{N}(s, S_V, n) = \Gamma(p_S^*, w^s, x)$. Thus, we've show that $\mathbb{N}(s, S_V, i) \in cof(\mathbb{N})$ for all $s \in S_V$ and all $i \leq n$, from which it follows that $S_V \in \mathsf{CofExt}_{n+1}$.

To show that $V \subseteq \mathsf{Ret}_{n+1}(p_S^*, \epsilon)$, recall that by inductive hypothesis $U \subseteq \mathsf{Ret}_n(p_S^*, \epsilon)$, and so for all $w \in U$, there are $s_0 < s_1 \ldots < s_n$ such that

$$p_S^*([M(w \restriction s_j) = T_{K_j}]) > 1 - x.$$

for all $j \leq n$. Now every world $w \in V$ is of the form w^{s*t} where $w^s \in U \subseteq \mathsf{Ret}_n(p_S^*, \epsilon)$ and $t > \gamma(p_S^*, w^s, x)$. Thus:

$$w^s \restriction j = w^{s*t} \restriction j.$$

for all $j \leq \gamma(p_S^*, w^s, x) < t$. It follows that

$$p_S^*([M(w^s \restriction j) = A]) = p_S^*([M(w^{s*t} \restriction j) = A])$$

for all $j \leq \gamma(p_S^*, w^s, x) < t$ and all $A \in \mathsf{Ans}$. Thus, for all $j \leq n-1$, it follows that $p_S^*([M(w^{s*t} \restriction s_j) = T_{K_j}]) = p_S^*([M(w^s \restriction s_j) = T_{K_j}]) > 1 - x$. Moreover, by definition of $\gamma(p_S^*, w^s, x)$, it follows that

$$p_S^*([M(w^{s*t} \restriction s_n] = T_{K_n}) = p_S^*([M(w^s \restriction s_n) = T_{K_n}]) > 1 - x.$$

Finally, because p_S^* converges in probability, there is some $s_{n+1} \in \mathbb{N}$ such that $p_S^*([M(w^{s*t} \restriction s_{n+1}) = T_{K_{n+1}}]) > 1 - x$. Therefore, the sequence $s_0, s_1, \ldots, s_{n+1}$ witness the fact that $w^{s*t} \in Ret_{n+1}(p_S^*, \epsilon)$. So we've shown that $V \subseteq \mathsf{Ret}_{n+1}(p_S^*, \epsilon)$ as desired.

(End Proof of Lemma 3.2.8) □

Now to finish the proof of Theorem 3.2.1, we use the above show that, for any mixed strategy p_S^* for the scientist that is convergent in probability, $\overline{\epsilon}(p_S^*, p_N) = \overline{\rho}(p_S^*, p_N) = \infty$. Therefore, because $p_N(W_K) = 1$, the mixed strategy p_N is optimal for Nature. It follows that, for any mixed strategy p_S^* for the scientist that is convergent in probability, (p_S^*, p_N) is a Nash equilibrium. To show $\overline{\epsilon}(p_S^*, p_N) = \overline{\rho}(p_S^*, p_N) = \infty$, note:

$$\begin{aligned}
\overline{\rho}(p_S^*, p_N) &= \int_{A_N} \int_{A*_S} \sum_{m=0}^{\infty} \rho(M, w, m) \cdot p_S^*(dM) \cdot p_N(dw) \\
&= \sum_{n=0}^{\infty} \int_{\mathsf{Comp}_n} \int_{A*_S} \sum_{m=0}^{\infty} \rho(M, w, m) \cdot p_S^*(dM) \cdot p_N(dw) \\
&\geq \sum_{n=0}^{\infty} n \cdot g(n) && \text{by the previous lemma} \\
&> \sum_{n=0}^{\infty} 2^n \cdot g(2^n) \\
&= \sum_{n=0}^{\infty} 2^n \cdot \frac{1}{2^{n+1}} && \text{by definition of } g \\
&= \sum_{n \in \mathbb{N}} \frac{1}{2} \\
&= \infty
\end{aligned}$$

Because each of the forced retractions above also constitutes an error, one can use a similar argument to that above to show that $\overline{\epsilon}(p_S^*, p_N) = \infty$ as well. Thus, Nature can do no better than p_N as it minimizes her convergence costs c_N, and maximizes the number of errors and retractions the scientist might accrue. Hence, for any mixed strategy p_S^* for the scientist, the pair (p_S^*, p_N) constitutes a Nash equilibrium. All that remains is to provide proofs of Lemmas 3.2.6 and 3.2.7.

Proof of Lemma 3.2.6: B_0 is clearly an algebra. So assume $n > 0$. First, note that $C_n = \{w^s : s \in \mathbb{N}^{[n]}\}$. Hence, as $\mathbb{N}^{[n]} \in \mathsf{CofExt}_n$, it follows that $C_n \in Y_n$, and therefore, $C_n \in B_n$. Next, notice that B_n is closed under complements (relative to C_n) by construction, as every element is either in Y_n or Z_n, and the elements in Y_n are complements of those in Z_n by construction (and vice versa). Finally, we must show that B_n is closed under finite unions. To so so, we first show that CofExt_n is closed under finite unions and intersections.

By construction, CofExt_n is trivially closed under finite unions. To show CofExt_n is closed under finite intersections, one proceeds by induction on n. The base case, when $n = 1$, follows from the fact that co-finite subsets of $\mathbb{N}$ are closed under intersection. For the inductive step, let $S, S' \in \mathsf{CofExt}_{n+1}$. Then:

$$S \cap S' = \{s * \mathbb{N}(s, S \cap S', n+1) \ : \ s \in S_n \cap S'_n\}$$

But $\mathbb{N}(s, S \cap S', n+1) = \mathbb{N}(s, S, n+1) \cap \mathbb{N}(s, S', n+1)$ is co-finite because co-finite subsets of $\mathbb{N}$ are closed under intersection. Moreover, $S_n \cap S'_n \in \mathsf{CofExt}_n$ by inductive hypothesis. This shows that $S \cap S' \in \mathsf{CofExt}_{n+1}$ because one can check that every $U \in \mathsf{CofExt}_{n+1}$ is of the form:

$$U = \{u * F_U(u) \ : \ u \in U' \in \mathsf{CofExt}_n \text{ and } F_U : U' \to cof(\mathbb{N})\}$$

Now we show that B_n is closed under finite unions. Let $U, V \in B_n$. The proof breaks into two cases:

Case 1: Suppose either U or V (or both) is an element of Y_n. Without loss of generality, assume $U \in Y_n$. We claim that $U \cup V \in Y_n$. To do so, note that, by definition of Y_n, there exists $S_U \in \mathsf{CofExt}_n$ such that $\{w^s : s \in S_U\} \subseteq U$. But then $\{w^s : s \in S_U\} \subseteq U \cup V$, and it follows that $U \cup V \in Y_n$.

Case 2: Suppose both $U, V \in Z_n$. We claim that $U \cup V \in Z_n$. By Demorgan's laws, it suffices to show that Y_n is closed under intersection. But this follows immediately from the fact that CofExt^n is closed under finite intersections.

(End Proof of Lemma 3.2.6) □

.

Proof of Lemma 3.2.7: Let $\langle p^n \rangle_{n \in \mathbb{N}}$ be a sequence of finitely additive probability measures on same underlying measurable space $(\Omega, \mathcal{F})$, and let $\langle r_n \rangle_{n \in \mathbb{N}}$ be a sequence of real numbers such that $\sum_{n \in \mathbb{N}} r_n = 1$. Define $p := \sum_{n \in \mathbb{N}} r_n \cdot p^n$. We want to show p is a probability measure on $(\Omega, \mathcal{F})$. First, note that

$$p(\Omega) = \sum_{n=0}^{\infty} r_n \cdot p^n(\Omega) = \sum_{n=0}^{\infty} r_n = 1$$

Next, for any finite collection of disjoint events $\{Q_1, \ldots, Q_m\}$, one has:

$$\begin{aligned} p(\bigcup_{i \leq m} Q_i) &= \sum_{n=0}^{\infty} r_n \cdot p^n(\bigcup_{i \leq m} Q_i) \\ &= \sum_{n=0}^{\infty} r_n \cdot \left(\sum_{i \leq m} p^n(Q_i) \right) \\ &= \sum_{i \leq m} \sum_{n=0}^{\infty} r_n \cdot p^n(Q_i) \\ &= \sum_{i \leq m} p(Q_i) \end{aligned}$$

(End Proof of Lemma and Theorem 3.2.1) □

Although the above proof only constructs one equilibrium in G^* (when K contains an infinite path), it provides a recipe for constructing uncountably many more. The mixed strategy p_N for Nature in the above proof is a weighted average of probability measures of the form p^n, where p^n is defined on an algebra B_n that is a subset of complexity class n. The function $g : \mathbb{N} \to [0, 1]$, which weighted the amount of probability assigned to B_n, is a "St. Petersburg"-like distribution, in that it assigned appropriate probabilities to various worlds so that the expected cost to the scientist is infinite. One can simply modify the function g to obtain various other St. Petersburg-like distributions, and the remainder of the proof remains unchanged.

3.2.3 G^*_*: Actions as Methods and Forcing Patterns

In the last section, I showed that, if Nature's set of pure strategies is identified with the set of worlds and the scientist's set of pure strategies consists of methods, then the canonical representation of the KGS model as a game has finitely-additive equilibria, but it does not have a countably additive equilibrium under certain conditions. But there is an asymmetry in the above game that puts Nature at a disadvantage. As the scientist's strategies consist of functions from effects to answers, there is a sense in which the scientist can *respond* to Nature's moves, even though it's a one-shot game. What if one permits Nature to do the same?

In this section, I provide a third way of representing the KGS model as a strategic game that eliminates this asymmetry. Here, Nature's set of pure strategies consists of functions that map finite answer sequences provided by the scientist to effect sets (i.e. $L : \mathsf{Ans}^{<\omega} \rightarrow 2^E$). It is shown in this game there are, in fact, infinitely many countably-additive, mixed strategy Nash equilibria.

Let the set of pure strategies for the scientist A^*_S be the set of functions of the form $M : (2^E)^{<\omega} \rightarrow \mathsf{Ans}$. Similarly, let the set of pure strategies for Nature A^*_n be the collection of functions of the form $L : \mathsf{Ans}^{<\omega} \rightarrow 2^E$ such that $L(\alpha) \subseteq f(\alpha'$ for all $\alpha, \alpha' \in \mathsf{Ans}^{<\omega}$ where α is an initial segment of α' (in symbols, $\alpha \leq \alpha'$).

Then, for any pure strategy for Nature $L : \mathsf{Ans}^{<\omega} \rightarrow 2^E$ and any strategy $M : (2^E)^{<\omega} \rightarrow \mathsf{Ans}$ for the scientist, define by simultaneous recursion:

$$\begin{aligned}
\eta_0 &= L(\emptyset) \\
\alpha_0 &= M(\langle \eta_0 \rangle) \\
\eta_{n+1} = &= L(\langle \alpha_0, \ldots, \alpha_n \rangle) \\
\alpha_{n+1} = &= M(\langle \eta_0, \ldots, \eta_{n+1} \rangle)
\end{aligned}$$

Then the sequences $\eta^{M,L} = (\eta_n)_{n \in \mathbb{N}}$ and $\alpha^{M,L} = (\alpha_n)_{n \in \mathbb{N}}$ are members of $(2^E)^\omega$ and Ans^ω respectively. Hence, one can use the same techniques in Section 4.2.1 to define the total number of errors and retractions committed by the method M in response to Nature's strategy L as follows. namely, define:

$$(M, L) \preceq_S (M', L') \Leftrightarrow (\alpha^{M,L}, \eta^{M,L}) \preceq_S (\alpha^{M',L'}, \eta^{M',L'})$$

where the ordering on the right side of the biconditional is defined as in equation (I) in Section 4.2.1.

One can extend this ordering to mixed strategies, but to do so, one first needs to define the notion of a mixed strategy. Define the events:

$$\begin{aligned}
[M_\eta = \alpha \upharpoonright m] &= \{M \in A^*_S \;:\; M(\eta \upharpoonright j) = \alpha \upharpoonright j \text{ for all } j \leq m\} \\
[L_\alpha = \eta \upharpoonright m] &= \{L \in A^*_N \;:\; L(\alpha \upharpoonright j) = \eta \upharpoonright j \text{ for all } j \leq m\} \\
\mathcal{A}^*_S &= \sigma(\{[M_\eta = \alpha \upharpoonright m] \;:\; M \in A^*_S, \eta \in A_N, \alpha \in A_S\}) \\
\mathcal{A}^*_N &= \sigma(\{[L_\alpha = \eta \upharpoonright m] \;:\; L \in A^*_N, \eta \in A_N, \alpha \in A_S\})
\end{aligned}$$

Notice that $\mathcal{A}_S^*$ is defined identically as in the previous section; the definition is repeated as a reminder. Then mixed strategies for the scientist and Nature are probability measures p_S^* and p_N^* on $\mathcal{A}_S^*$ and $\mathcal{A}_N^*$ respectively. Then one can define:

$$\begin{aligned}
\overline{\epsilon}(p_S^*, p_N^*) &= \int_{A_N^*}\int_{A_S^*} \epsilon(\alpha^{M,L}, \eta^{M,L}) \cdot p_S^*(dM) \cdot p_N^*(dL) \\
\overline{\rho}(p_S^*, p_N^*) &= \int_{A_N^*}\int_{A_S^*} \rho(\alpha^{M,L}, \eta^{M,L}) \cdot p_S^*(dM) \cdot p_N^*(dL) \\
\overline{c}_S(p_S^*, p_N^*) &= \int_{A_N^*}\int_{A_S^*} c_S(\alpha^{M,L}, \eta^{M,L}) \cdot p_S^*(dM) \cdot p_N^*(dL) \\
\overline{c}_N(p_N^*, p_N^*) &= \int_{A_N^*}\int_{A_S^*} c_N(\eta^{M,L}) \cdot p_S^*(dM) \cdot p_N^*(dL)
\end{aligned}$$

Notice that, because Nature is capable of responding to the scientist's answers at successive stages of inquiry in G_*^*, the cost $c_N(p_N^*, p_S^*)$ above depends upon both Nature's strategy and the scientist's strategy. As in G^*, the definition of the preference relation for the scientist (and Nature) requires considering methods that "converge in probability." Because the game has changed, however, one must alter the definition of convergent in probability ever so slightly. Say a pure strategy $L \in A_N^*$ for Nature is **convergent** if for all methods $M \in A_S^*$, the sequence $\eta^{M,L}$ is a world (i.e. $\eta^{M,L} \in W_K$). If L is a convergent strategy for Nature, say a pure strategy M for the scientist is **convergent in** L if there exists an $n \in \mathbb{N}$ such that $M(\eta^{M,L} \restriction n') = T_{\eta^{M,L}}$ for all $n' \geq n$. Say M is **convergent** (simpliciter) if M converges in L for all convergent L. Finally, say a mixed strategy p_S^* for the scientist is **convergent in probability** if for all convergent strategies L for Nature, one has:

$$\lim_{n\to\infty} p_S^*([M(\eta^{M,L} \restriction n) = T_{\eta^{M,L}}]) = 1$$

Given the definition of convergent in probability, retractions, and errors for pairs of actions A_S^* and A_N^*, one can then define the preference relation for the scientist in the same way as in Equation III of the previous section. This completes the third formalization of the KGS model as a game. Call the game G_*^*.

An additional similarity between G^* and G_*^* is that both admit an alternative definition to the preference relation that is equivalent when p_S^* and p_N^* are both countably additive. The alternative definition is described and made possible by the following propositions:

Proposition 3.2.8. The function $\varphi_*^* : A_S^* \times A_N^* \to A_S \times A_N$ defined by $\varphi_*^*(M, L) = (\alpha_{M,L}, \eta_{M,L})$ is $\sigma(\mathcal{A}_S^* \times \mathcal{A}_N^*)/\sigma(\mathcal{A}_S \times \mathcal{A}_N)$-measurable.

Proof: The proof is nearly identical to that of Proposition 3.2.5.

□

Proposition 3.2.8 motivates the following approach to evaluating mixed strategies. Let p_S^* and p_N^* be probability measures on $\mathcal{A}_S^*$ and $\mathcal{A}_N^*$ respectively. Let p^* be the unique induced product measure on $\sigma(\mathcal{A}_S^* \times \mathcal{A}_N^*)$ (Recall, we are assuming that both p_S^* and p_N^* are both countably additive for the moment). Let p be the measure induced on $\sigma(\mathcal{A}_S \times \mathcal{A}_N)$ by the mapping φ_*^*. That is, for all $X \in \sigma(\mathcal{A}_S \times \mathcal{A}_N)$, define:

$$p(X) = p^*((\varphi_*^*)^{-1}(X))$$

The function p is a probability measure because φ_*^* is measurable, in light of proposition 3.2.8. Finally, for $X \in \mathcal{A}_S$ and $Y \in \mathcal{A}_N$, define:

$$\begin{array}{rcl} p_S(X) & = & p(X \times A_N) \\ p_N(Y) & = & p(A_S \times Y) \end{array}$$

Then one can use the preference relations defined in Equation II to order mixed strategies in this new game as follows:

$$\begin{array}{ll} (p_S^*, p_N^*) \preceq_S (q_S^*, q_N^*) & \Leftrightarrow (p_S, p_N) \preceq_S (q_S, q_N) \\ (p_S^*, p_N^*) \succeq_N (q_S^*, q_N^*) & \Leftrightarrow (p_S, p_N) \preceq_S (q_S, q_N) \end{array}$$

In order to begin analyzing whether G_*^* has a solution, it is necessary to build a catalog of known measurable events. A short lemma will be helpful in doing so:

Then by analogous reasoning to that in Lemma 3.2.1, one obtains the following facts:

Proposition 3.2.9. The following events are measurable:

1. For any $A \in \mathsf{Ans}$ and any $n \in \mathbb{N}$, the event $[M_{\eta,n} = A] = \{M \in A_S^* : M(\eta \upharpoonright m) = A\}$ is $\mathcal{A}_S^*$-measurable. For any $E_0 \subset E$, the event $[f_{\alpha,m} = E_0] = \{f \in A_N^* \ : \ f(\alpha \upharpoonright m) = E_0\}$ is $\mathcal{A}_N^*$-measurable

2. For any $M \in A_S^*$, the singleton $\{M\}$ is $\mathcal{A}_S^*$-measurable, and analogously, the singleton $\{f\}$ is $\mathcal{A}_N^*$-measurable for any $f : \mathsf{Ans}^{<\omega} \rightarrow 2^E$. In particular, one can identify any world w with a function L_w such that $L_w(\alpha \upharpoonright n) = w_n$ for all $\alpha \in \mathsf{Ans}^{<\omega}$ and all $n \in \mathbb{N}$. So one obtains that any world $w \in W$ is $\mathcal{A}_N^*$-measurable.

3. For any $\eta \in A_N^*$, the set $\mathcal{M}_c^\eta = \{M \in A_S^* : M \text{ converges in } \eta\}$ is $\mathcal{A}_S^*$-measurable, and analogously, the set $\mathcal{F}_c^M = \{f \in A_N^* \ : \ \eta_{M,f} \in W\}$ is $\mathcal{A}_N^*$-measurable.

4. The set $\mathcal{M}_c = \{M \in A_S^* : \text{ for all } \eta \in W, M \text{ converges in } \eta\}$ is $\mathcal{A}_S^*$-measurable. Similarly, $\mathcal{F}_c = \{f \in A_N^* : \text{ for all } M \in \mathcal{M}_c, \eta_{M,f} \in W\}$ is $\mathcal{A}_N^*$-measurable.

Unlike the first two representations of the KGS model as a game, one need not resort to finite mixed strategies in order to obtain Nash equilibria when K contains a infinite path. Here, we present an intuitive argument that Nash equilibria exist in the above game even when K is infinite, and we make the argument more precise below. Assume that Nature has an infinite collection of dice, and suppose that each die has 2^n sides labeled with numbers 1 to 2^n. Furthermore, assume that, for each die with 2^n sides, the probability that the die will land on the highest labeled side 2^n is .99 and the probability it will land on any other side is $\frac{1}{100\cdot(2^n-1)}$. Nature can then achieve an infinite number of expected errors and retractions, while still converging, by employing the following mixed strategy. Nature flips a coin until it lands heads. If the coin lands heads on the n^{th} toss, she then rolls the die with 2^n sides. Suppose that the die lands on side k. Nature then picks an arbitrary sequence of theories $\langle T_1, T_2, \ldots, T_k\rangle$ such that $T_i \subseteq T_{i+1}$ for all $i \leq k$. She then presents theory T_1 at the outset of inquiry. If the scientist never returns T_1, then Nature wins by presenting T_1 indefinitely. Otherwise, the scientist eventually returns T_1, and then Nature switches to T_2. Nature repeats this process until she reaches theory T_k, at which point she stops returning different effect sets. Suppose that, in response to this mixed strategy, the scientist employs any mixed strategy that assigns probability one to convergent methods (i.e $p^*_S(\mathcal{M}_c) = 1$).

If Nature and the scientist follow these respective strategies, then the scientist will accrue an infinite number of expected retractions and errors. Nature can do no better, moreover, as she achieves the maximum possible gains in the game. Why? It is proven below that the scientist achieves a convergence cost of zero in this game by assigning probability one to convergent methods. Hence, by employing the above strategy, Nature achieves the best possible outcome in the game, as she minimizes her cost of non-convergence (namely, by making it zero) and maximizes the error and retraction costs to the scientist (by making them infinite). The scientist, in turn, can do no better because, if he were to reduce his expected number of errors and retractions, he would do so on pains of increasing the probability that he does not converge to the true theory, thus increasing his cost of non-convergence. Hence, we've found (an infinite) number of Nash equilibria.

We now make this argument more precise. Suppose K contains an infinite path $\langle K_n\rangle_{n\in\mathbb{N}}$. For each $n \in \mathbb{N}$ define $\mu_n : \{K_n\}_{n\in\mathbb{N}} \to \{K_n\}_{n\in\mathbb{N}}$ by:

$$\mu_n(S) = \begin{cases} K_{j+1} \text{ if } S = K_j \text{ and } j < n \\ K_{n+1} \text{ otherwise} \end{cases}$$

Again, for each n, define $L_n \in A^*_N$ as follows. For every $\alpha \in \mathsf{Ans}^\omega$ and any $m \in \mathbb{N}$:

$$\begin{aligned} L_n(\emptyset) &= L(\alpha_0) = K_0 \\ L_n(\alpha \upharpoonright m+1) &= \begin{cases} L_n(\alpha \upharpoonright m) \text{ if } \alpha_m \neq T_{L_n(\alpha\upharpoonright m)} \\ \mu_n(L_n(\alpha \upharpoonright m)) \text{ otherwise} \end{cases} \end{aligned}$$

Here, L_n corresponds to a strategy for Nature that involves playing K_0 until the scientist returns T_{K_0}, then playing K_1 until the scientist returns T_{K_1}, and so on until returning K_{n+1} for eternity. The following facts are easily proven:

Lemma 3.2.9.

1. L_n is convergent for all n.
2. Suppose M is convergent. The $\rho(M, L_n) := \rho(\alpha^{M,L_n}) \geq n$.
3. Suppose p_S^* is convergent in probability. Then $\overline{\rho}(p_S^*, L_n) \geq n$ for all $n \in \mathbb{N}$.

The above lemma can be used to prove the following.

Theorem 3.2.2. Suppose K contains an infinite path. Then there exist infinitely many countably additive mixed strategy Nash equilibria in G_*^*.

Proof: Like the proof of Theorem 3.2.1, one defines a "St. Petersburg"-like distribution p_N^* on the L_n's and shows that, for any mixed strategy for the scientist p_S^* that converges in probability, the $\overline{\rho}(p_S^*, p_N^*) = \overline{\epsilon}(p_S^*, p_N^*) = \infty$. Because there are infinitely many such St. Petersburg-like distributions, the result follows. We give an example below.

Let $g : \mathbb{N} \to [0, 1]$ be the probability measure defined on the power set of the natural numbers like the one defined in Theorem 3.2.1. Recall:

$$g(n) = \begin{cases} \frac{1}{8} & \text{if } n \in \{0, 1\} \\ \frac{1}{2^{d+1}} & \text{if } n = 2^d \text{ for some } d \in \mathbb{N}, \\ \frac{1}{2^{d+1}(2^d-1)} & \text{if } 2^d < n < 2^{d+1} \text{ for some } d \in \mathbb{N} \end{cases}$$

Let p_N^* be the unique countably additive measure such that $p_N^*(L_n) = g(n)$. Then by the previous lemma:

$$\begin{aligned} \overline{\rho}(p_S^*, p_N^*) &= \int_{A_N^*} \int_{A_S^*} \rho(\alpha^{M,L}) \cdot p_S^*(dM) \cdot p_N^*(dM) \\ &= \sum_{n=0}^{\infty} g(n) \cdot \left[\int_{A_S^*} \rho(\alpha^{M,L_n}) \cdot p_S^*(dM) \right] \qquad \text{definition of } p_N^* \\ &\geq \sum_{n=0}^{\infty} g(n) \cdot n \qquad \text{by Lemma 3.2.9} \\ &= \infty \end{aligned}$$

□

3.3 A Brief Discussion of the Three Representations and the New Efficiency Conjecture

Although each of the three representations of the KGS model as a game have equilibria, in no such games are the scientist's half of the equilibria constituted uniquely by Ockham strategies. This raises the question: why do the game theoretic results above not provide an alternative proof of the Efficiency Theorems?

Roughly, there are two hurdles to proving a more general Efficiency Theorem within the game-theoretic framework. Recall in the description of the KGS model, one does not compare the worst-case costs of two methods over *all* worlds because such costs are, in general, infinite. Instead, one compares the costs of two methods across complexity classes, where such costs are often bounded. This suggests the conjecture that, if the preference relations above in G^*, and G^*_* are refined to consider comparisons of errors and retractions within complexity classes, only mixtures of Ockham methods (that converge in probability) for the scientist will be in equilibria with Nature's finitely additive mixed strategies.

The major stumbling block to proving this conjecture, however, is to provide a more general definition of simplicity. Why? If Nature "zeros out" particular complexity classes by playing the complexity class with probability zero, then the scientist, it seems, ought never to guess a theory of that complexity, even if such a theory is simplest according to the definitions of the KGS model. Note this isn't a challenge to the defense of Ockham's razor provided by the Efficiency Theorems, for if Nature's mixed strategy represents the scientist's prior distribution on worlds (or fooling strategies in G^*_*), then the scientist hardly ought to consider a theory with zero probability "simplest." In other words, the KGS model ought to be revised so that the definition of simplicity is a function, in some way or another, of the scientist's prior distribution on worlds. Yet it is not clear how the definition of simplicity ought to reflect one's prior distribution, and so this very general conjecture remains open.

Chapter 4

Appendices

4.1 Measure Theory and Probability

4.1.1 Measures and Measure Spaces

Let Ω be any set, and $\mathcal{F}$ be a subset of 2^{Ω}, the power set of Ω. $\mathcal{F}$ is called a *σ-algebra* if (i) $\Omega \in \mathcal{F}$, (ii) $S \in \mathcal{F}$ implies that S^c, the complement of S is in $\mathcal{F}$, and (iii) If $\langle S_n \rangle_{n \in \mathbb{N}}$ is a countable sequence of events such that $S_n \in \mathcal{F}$ for all $n \in \mathbb{N}$, then $\bigcup_{n \in \mathbb{N}} S_n \in \mathcal{F}$. In other words, a σ-algebra on a set Ω is a collection of subsets of Ω that contains Ω itself and is closed under countable unions and complements. By Demorgan's laws, σ-algebras are also closed under countable intersections. If one relaxes condition (iii) so that $\mathcal{F}$ need only be closed under finite unions, then $\mathcal{F}$ is called an *algebra* (simpliciter). A pair $\langle \Omega, \mathcal{F} \rangle$ where $\mathcal{F}$ is a σ-algebra on Ω is called a *measurable space,* and elements of $\mathcal{F}$ are called *measurable sets.*

Lemma 4.1.1. If $\langle \mathcal{F}_i \rangle_{i \in I}$ is a family of σ-algebras on some set Ω, then $\bigcap_{i \in I} \mathcal{F}_i$ is a σ-algebra on Ω.

Let $B \subseteq 2^{\Omega}$ and $\Sigma(B) = \{\mathcal{F} \subseteq 2^{\Omega} \ : \ \mathcal{F} \text{ is a } \sigma- \text{ algebra and } B \subseteq 2^{\Omega}\}$. Then by the above lemma, $\sigma(B) = \bigcap_{\mathcal{F} \in \pm(\mathcal{B})} \mathcal{F}$ is a σ-algebra on Ω. A particularly important σ-algebra is the *Borel algebra* $\mathbb{B}(\mathbb{R}^n)$, which the one generated by open sets in Euclidean space $\mathbb{R}^n$ (i.e. $\mathbb{B}(\mathbb{R}^n) = \sigma(\mathcal{O})$ where $\mathcal{O} = \{\mathcal{S} \subseteq \mathbb{R}^{\setminus} :$ $\mathcal{S}$ is open $\}$). In general, if Ω is a set and τ is a topology on Ω, then $\sigma(\tau)$ is called the *Borel Algebra with respect to τ*.

If $\langle \Omega, \mathcal{F} \rangle$ is a measurable space, then a function $p : \mathcal{F} \to \mathbb{R} \cup \{\infty\}$ is called a *countably additive measure* if for every countable collection $\{S_n\}_{n \in \mathbb{N}}$ of disjoint sets $S_n \in \mathcal{F}$, it follows that $p(\bigcup_{n \in \mathbb{N}} E_n) = \sum_{n \in \mathbb{N}} p(E_n)$. Similarly, a function $p : \mathcal{F} \to \mathbb{R} \cup \{\infty\}$ is called a *finitely additive measure* if $p(\bigcup_{i \leq n} E_i) = \sum_{i \leq n} p(E_i)$ for any finite collection of disjoint measurable sets $\{E_1, \ldots, E_n\}$. Notice that every countably additive measure is finitely additive, but not vice versa. For example, let p be a measure on $2^{\mathbb{N}}$ that assigns probability zero to each natural number and probability one to the set of all natural numbers. For this reason, call measures that are finitely additive but not countably additive *purely finitely additive*. A *measure space* is a triple $\langle \Omega, \mathcal{F}, p\rangle$ such that $\langle \Omega, \mathcal{F}\rangle$ is a measurable space and p is a finitely additive measure. If p is countably additive, then $\langle \Omega, \mathcal{F}, p\rangle$ is called a *countably additive measure space*.

A measure p is said to be *σ-finite* if there exists a countable set of measurable sets $\{S_n\}_{n\in\mathbb{N}}$ such that $\Omega = \bigcup_{n\in\mathbb{N}} S_n$ and each S_n has finite measure (i.e $p(S_n) < \infty$ for all $n \in \mathbb{N}$). A *probability space* is a measure space in which $p(\Omega) = 1$. All probability measures are trivially σ-finite, as $p(\Omega) = 1$, and so the constant countable sequence $\langle \Omega, \Omega, \ldots\rangle$ yields the desired witness to σ-finiteness.

It is often difficult to define a measure on the entirety of some σ-algebra $\mathcal{F}$ on a set Ω. However, if one has a measure p on a family of sets $B \subseteq 2^\Omega$ and $\mathcal{F} = \sigma(B)$, then, under certain conditions specified by the following theorem, p can be extended to a measure on the entirety of $\mathcal{F}$:

Theorem 4.1.1 (Caratheodory)**.** Let B be an algebra on a set Ω, and suppose $p : B \to \mathbb{R} \cup \{\infty\}$ is a countably-additive measure on B. Then there exists a countably-additive measure $\overline{p} : \sigma(B) \to \mathbb{R} \cup \{\infty\}$ extending p (i.e $\overline{p} \restriction B = p$). If p is, in addition, σ-finite, then $\overline{p}$ is unique and σ-finite.

An analogous result also holds for finitely additive measures.

Theorem 4.1.2. Let B be an algebra on a set Ω, and suppose $p : B \to \mathbb{R}\cup\{\infty\}$ is a finitely-additive measure on B. Then there exists a finitely-additive measure $\overline{p} : \sigma(B) \to \mathbb{R} \cup \{\infty\}$ extending p (i.e $\overline{p} \restriction B = p$).

The difference between the two theorems is that extensions of a finitely additive measure will in general not be unique, even when the finite measure is bounded (and hence, σ-finite). See Swartz (1994) for a proof of the latter theorem. A particularly important countably additive measure space is $\langle \mathbb{R}^n, \mathbb{B}(\mathbb{R}^n), \lambda\rangle$, where λ is the *Lebesgue measure*, which is unique countably-additive extension of the function that assigns every open ball its volume. The existence and uniqueness of λ follows from Caratheodory's Theorem.

Another important application of Caratheodory's Theorem is the construction of measures on product spaces. Let $\langle \Omega, \mathcal{F}, p\rangle$ and $\langle \Omega', \mathcal{F}', p'\rangle$ be two measure spaces. Define:

$$\mathcal{F} \otimes \mathcal{F}' = \sigma(\{E \times E' : E \in \mathcal{F}, E' \in \mathcal{F}'\})$$

and consider the measurable space $\langle \Omega\times\Omega', \mathcal{F}\otimes\mathcal{F}'\rangle$. For every pair of sets $E \in \mathcal{F}$ and $E' \in \mathcal{F}'$, define a function $p \times p'(E \times E') = p(E) \cdot p(E')$. Then if p and p' are countably additive and σ-finite, then by Caratheodory's theorem, the set function $p\times p'$ extends to a unique countably additive σ-finite measure $\overline{p \times p'}$ on $\mathcal{F} \otimes \mathcal{F}'$. As a particular example, when p and p' are both probability measures, then $\overline{p \times p'}$ is a probability measure (as $\overline{p \times p'}(\Omega \times \Omega') = p \times p'(\Omega \times \Omega') = p(\Omega) \cdot p'(\Omega') = 1$). When either p or p' are purely finitely additive, Theorem 4.1.2 guarantees the existence of a measure on $\mathcal{F} \otimes \mathcal{F}'$, but uniqueness is not guaranteed.

4.1.2 Measurable Maps and Random Variables

Let $\langle \Omega, \mathcal{F}, p\rangle$ and $\langle \Omega', \mathcal{F}', p'\rangle$ be two measure spaces and $X : \Omega \rightarrow \Omega'$ be a function. Then X is said to be $\mathcal{F}/\mathcal{F}'$*-measurable* if $X^{-1}(S) \in \mathcal{F}$ for all $S \in \Omega'$. When the measurable spaces are unambiguous, one just says that X is measurable without mentioning the underlying sets and σ-fields. If $\langle \Omega, \mathcal{F}, p\rangle$ is a probability space, then X is called an *abstract random variable.* If $\langle \Omega, \mathcal{F}, p\rangle$ is a probability space and If $\langle \Omega', \mathcal{F}', p'\rangle$ is the measure space $\langle \mathbb{R}, \mathbb{B}(\mathbb{R}), \lambda$, then X is called a *random variable* (simpliciter). If X is a measurable mapping from the measure space $\langle \Omega, \mathcal{F}, p\rangle$ to some other measurable space $\langle \Omega', \mathcal{F}'$, then one can define a measure p' on $\langle \Omega', \mathcal{F}'$ as follows:

$$p'(S) = p(X^{-1}(S))$$

for all $S \in \mathcal{F}'$. In this case, p' is called the *image measure* or *measure induced by* X. When X is a random variable, then p' is called the distribution of X.

4.1.3 Integration

Let $(\Omega, \mathcal{F}, p)$ be a measure space and $g : \Omega \rightarrow \mathbb{R}$ a measurable function. Then g is called *simple* if there exist finitely many real numbers $r_1, r_2, \ldots, r_n \in \mathbb{R}$ and finitely many events $Q_1, \ldots, Q_n \in \mathcal{F}$ such that $g = \sum_{i=1}^{n} r_i \cdot 1_{Q_i}$, where 1_{Q_i} is the characteristic function of the event Q_i. The the integral $\int f dp$ of an arbitrary $\mathcal{F}/\mathbb{B}(\mathbb{R})$-measurable function f is defined as follows. For simple functions $g = \sum_{i=1}^{n} r_i \cdot 1_{Q_i}$, one defines $\int g dp = \sum_{i=1}^{n} r_i \cdot p(Q_i)$. Then for any arbitrary measurable function f, the integral $\int f dp$ is then defined to be $\sup\{\int g dp \ : \ g \text{ is simple and } g(\omega) \leq f(\omega) \text{ for all } \omega \in \Omega\}$.

If $(\Omega_1, \mathcal{F}_1, p_1)$ and $(\Omega_2, \mathcal{F}_2, p_2)$ are countably-additive measure spaces, then Fubini's theorem provides a direct connection between integration over the product space

Theorem 4.1.3 (Fubini-Tonelli)**.** Suppose $(\Omega_1, \mathcal{F}_1, p_1)$ and $(\Omega_2, \mathcal{F}_2, p_2)$ are countably-additive, σ-finite measure spaces. Let $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, p_1 \times p_2)$ be the product space, and $f : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}^+$ be $\mathcal{F}_1 \otimes \mathcal{F}_2/\mathbb{B}(\mathbb{R})$-measurable. Then:

$$\int_{\Omega_1 \times \Omega_2} f \; d\overline{p_1 \times p_2} = \int_{\Omega_1} \int_{\Omega_2} f \; dp_2 \cdot dp_1 = \int_{\Omega_2} \int_{\Omega_1} f \; dp_1 \cdot dp_2$$

Fubini's theorem fails for merely finitely-additive measures, and so specification of order of integration is important when the measures under consideration are merely finitely-additive.

4.2 Directed Acyclic Graphs and Bayesian Networks

A *directed acyclic graph* (DAG) is an ordered pair $G = (V, A)$ where V is a finite set of *vertices* and $A \subset V \times V$ is an anti-reflexive, anti-symmetric relation representing the graph's *arrows* (often called *edges*). If $(v, v') \in A$, then we say v is a *parent* of v', and that v' is a *child* of v. Often, we will write $v \rightarrow v'$ is in A or $v \rightarrow v'$ appears in G if the ordered pair $(v, v') \in A$. For brevity, let $PA_G(v)$ denote the set of parents of v in G, and let $Ch_G(v)$ denote its children. An *undirected path* (or simply, path) π is a sequence $\pi = \langle v_1, v_2, \ldots, v_n \rangle$ such that $v_i \neq v_j$ for $i \neq j$ and either $(v_i, v_{i+1}) \in A$ or $(v_{i+1}, v_i) \in E$ for all $1 \leq i \leq n$. If $\pi_1 = \langle v_1, v_2, \ldots, v_n \rangle$ and $\pi_2 = \langle w_1, w_2, \ldots, w_m \rangle$ are paths, then we let $\pi_1 \frown \pi_2 = \langle v_1, v_2, \ldots, v_n, w_1, w_2 \ldots, w_m \rangle$ denote the concatenation of the two paths. Notice that the concatenation $\pi_1 \frown \pi_2$ of two paths need not be a path itself, as some variable v may be on both π_1 and π_2. Abusing notation, we will sometimes write $\{v_1, v_2, \ldots v_n\} \subseteq \pi$ if the variables $v_1, v_2, \ldots v_n$ appear on the path π. Finally, if $\pi_1 = \langle v_1, v_2, \ldots, v_n \rangle$ is a path, and $\pi_2 = \langle v_1, v_2, \ldots, v_r \rangle$ where $r \leq n$ is an initial segment (not necessarily proper) of π_1, then we say π_2 is a *subpath* of π_1 and we write $\pi_2 \sqsubseteq \pi_1$.

If $\pi = \langle v_1, v_2, \ldots, v_n \rangle$ is such that $(v_i, v_{i+1}) \in A$ for all i, then we say π is a *directed path* from v_1 to V_n. In this case, we also say that v_1 is an *ancestor* of v_n and that v_n is a *descendant* of v_1. Let $Desc_G(v)$ denote the set of descendants of v, and similarly let $Anc_G(v)$ denote its ancestors. Finally, if $v_1 \rightarrow v_3 \leftarrow v_2$ appears in a graph G, then we say v_3 is a *collider* with respect to v_1 and v_2. If v_2 is a collider with respect to v_1 and v_3 and, in addition, there is no edge between v_1 and v_2, then we say v_3 is an *unshielded collider* with respect to v_1 and v_2.

A Bayesian network (or Bayes net, for short) is a pair $\mathbb{B} = (G, p_G)$ where (i) the set of vertices V in the graph $G = (V, A)$ are random variables over a common measurable space $\langle \Omega, \mathcal{F}, p \rangle$ and (ii) p_G is the joint distribution on Euclidean space induced by the random variables V. Although Bayes nets are arbitrary ordered pairs of the form $\mathbb{B} = (G, p_G)$, this paper only considers Bayes nets that represent causal relationships amongst variables. As such, I assume that Bayes nets satisfy two conditions, called the Causal Markov Condition and the Faithfullness Condition.

The *Causal Markov Condition* (CMC) states that any random variable $v \in V$ is probabilistically independent of its non-descendants conditional on its parents in the graph G. Symbolically, let $v, v' \in V$ be such that $v' \notin Desc_G(v)$, then CMC states that

$$p_G(v, v'|PA_G(v)) = p_G(v|PA_G(v))p_G(v'|PA_G(v)).$$

In this case, say that p_G is *Markov* for the graph G. For any subset $S \subseteq V$, write $v \amalg v'|V_0$ to abbreviate the statement v is independent of v' conditional on V_0. In general, define cic_V to be the set of *conditional independence constraints* of the form $v \amalg v'|S$, where $v, v' \in V$ and $V_0 \subseteq V \setminus \{v, v'\}$.

The Causal Markov Condition can be written in a slightly more perspicuous way as follows. Let V be a finite set of random variables and p_G a joint probability distribution over V. Let I_p denote the set of conditional independence statements of the form $v \amalg v'|S$ that hold with respect to p. Finally, let $G = (V, A)$ be a DAG and $\mathcal{MG}$ denote the set of conditional independence constraints of the form:

$$v \amalg v' | PA_G(v)$$

where $v, v' \in V$ and $v' \notin Desc_v$. Then we say $\mathbb{B} = (G, p)$ satisfies the CMC if and only if p is Markov for G if and only if $\mathcal{MG} \subseteq I_p$.

The second assumption here about Bayes nets is the faithfulness condition, which states that if two variables $v_1, v_2 \in V$ are independent conditional on $S \subset V$, then this independence is logically entailed by the Markov condition. Equivalently, let $\mathbb{B} = (G, p_G)$ be a Bayes Net where p_G is Markov for G. Then we say $\mathbb{B}$ satisfies the faithfulness condition and that p_G is *faithful* for G if for all $\mathbb{B}' = (G, p'_G)$ where p'_G is Markov for G, we have $I_{p_G} \subseteq I_{p'_G}$. More perspicuously, let $G = (V, A)$ be a DAG, and define

$$I_G = \bigcap \{Ip : p \text{ is a joint distribution over } V \text{ and } \mathcal{MG} \subseteq I_p\}.$$

Then $\mathbb{B} = (G, p)$ is faithful and p is faithful for G if and only if $I_p = I_G$. For brevity, say that I_G is the set of conditional independencies *implied by the graph* G.

Although it is standard to characterize a DAG by the set of conditional *independencies* that it implies, it is sometimes helpful to have notation to speak of probabilistic *dependencies.* Accordingly, abbreviate the negation of a statement $v \amalg v' | V_0$ in $\textsc{cic}_V$ by $D(v, v'|V_0)$. That is, $D(v, v'|V_0)$ is the assertion that v and v' are **not** probabilistically independent given V_0. Let $\textsc{cdc}_V$ be the set of all *conditional dependence constraints.* Then because every DAG $G = (V, E)$ implies some set of conditional independence constraints I_G in $\textsc{cic}_V$, one can define D_G to be negations of conditional independence constraints in the set $\textsc{cic}_V \setminus I_G$, and say that D_G is the set of conditional dependence constraints implied by the graph G.

For any given set of variables V and joint distribution P, define the *Markov Equivalence Class* to be the equivalence class of all DAGs representing the same conditional independencies on V. Symbolically, let $G = (V, A), G' = (V, A')$ be two DAGs over the same set of variables, and write $G \equiv G'$ if and only if $I_G = I_{G'}$. That is, $G \equiv G'$ if and only if G and G' imply the same set of conditional independencies. Clearly, $\equiv$ is an equivalence relation. Then let $\overline{G} = \{G' : G' \text{ is a DAG and } G \equiv G'\}$. Markov Equivalence classes are important because, supposing the true causal graph is G, one cannot expect to distinguish between any two members of $\overline{G}$ using conditional independence information alone.

The graphical representations of Markov equivalence classes are called *patterns.* If $\overline{G}$ and $\overline{H}$ are Markov equivalence classes, then we will use the lower case letters g and h to denote their respective patterns. A pattern is much like a graph, except that an arrow A in a pattern g may be undirected if there are two graphs G and G' in $\overline{G}$ such that A has opposite orientations in G and G' respectively.

Using graph-theoretic notions only, the following theorem provides conditions under which two graphs belong to the same pattern.

Theorem 4.2.1 (Verma and Pearl)**.** Let $G = (V, A)$ and $G' = (V, A')$ be two DAGs. Then the following are equivalent:

- $G \equiv G'$
- G and G' have the same adjacencies and unshielded colliders.

By the above theorem, therefore, one can know when two Bayes Nets $B = (G, p)$ and $B' = (G', p')$ (over a common set of variables) imply the same conditional independencies by looking only at their respective graphs. That is, when investigating the conditional independence statements that hold for a Bayes net $B = (G, p)$, the CMC and faithfulness condition allow one to ignore the probability distribution p and pay attention solely to the graph G. Furthermore, the following definitions and theorem allow us to ascertain when *any particular* conditional independence statement $v \amalg v'|V_0$ holds for a Bayes net $B = (G, p)$ solely by looking at its graph.

Definition 4.2.1. Let $G = (V, E)$, $v, v' \in V$, π an undirected path between v and v', and $V_0 \subseteq V - \{v, v'\}$. Then a vertex v'' is **active on** π **relative to** U just in case either

1. v'' is not a collider on π and $v'' \notin V_0$
2. v'' is a collider on π and either (i) $v'' \in V_0$ or (ii) there is $w \in Desc_{v''} \cap V_0$ (or both).

Then say a path π between v and v' is active relative to V_0 just in case every variable on π is active.

Definition 4.2.2. Let $G = (V, E)$, $v, v' \in V$, and $V_0 \subseteq V - \{v, v'\}$. Then v and v' are **d-separated given** V_0 if and only there is no undirected path π between v and v' such that π is active relative to V_0.

Then the following theorem shows that d-separation and conditional independence are equivalent for Bayes nets satisfying the CMC and faithfulness condition.

Theorem 4.2.2 (Verma and Pearl)**.** Let $B = (G, p)$ be a Bayes net satisfying the CMC and faithfulness condition. Let $G = (V, E)$, $v, v' \in V$, and $V_0 \subseteq V - \{v, v'\}$. Then $v \amalg v'|V_0$ if and only if v and v' are d-separated given U.

A useful and important consequence of Verma and Pearl's theorem is the following proposition:

Proposition 4.2.1. Suppose either $v \to v'$ or $v' \to v$ appears in G. Then $D(v, v'|V_0) \in D_G$ for all $V_0 \subseteq \{v, v'\}$. Equivalently, $v \amalg v'|V_0 \notin I_G$ for all $V_0 \subseteq \{v, v'\}$.

Finally, given two Bayes nets $B_1 = (G_1, p_G)$ and $B_2 = (G_2, p_G)$ satisfying the CMC and faithfulness condition, one can often characterize the relationships between the sets of conditional independence constraints their graphs imply and their respective graphical structures. Suppose that $v \to v'$ is an edge in G. Say that $v \to v'$ is **covered** if and only if

$$PA_G(v') \setminus \{v\} = PA_G(v).$$

A **covered edge reversal** involves flipping a covered edge to obtain a new G'. Then:

Theorem 4.2.3 (Chickering 2002)**.** Let G_1 and G_2 be two DAGs. Then:

1. $I_{G_1} = I_{G_2}$ if and only if G_1 can be obtained from G_2 by a finite sequence of covered edge reversals.

2. $I_{G_1} \subseteq I_{G_1}$ if and only if G_1 can be obtained from G_2 by a finite sequence of edg additions and covered edge reversals.

Bibliography

Aristotle. Posterior Analytics. *Complete Works: Volumes I and II.* Ed. Jonathan Barnes. Princeton University Press, 1971.

Laurence Bonjour. *The Structure of Empirical Knowledge.* (Cambridge: Harvard University Press, 1985). pp.183.

Alan Baker. Quantitative Parsimony and Explanatory Power. *British Journal for the Philosophy of Science.* Vol. 54. 2003. pp. 245-259.

Alan Baker. "Simplicity." *Stanford Encyclopedia of Philosophy.* Available at ⟨ plato.stanford.edu. ⟩. 2004.

Alan Baker. Occam's Razor in Science: A Case Study from Biogeography. *Biology and Philosophy.* Vol. 22. 2007. pp. 193215.

Claude Berge, Melvin Dresher, Albert William Tucker, Philip Wolfe. *Contributions to the Theory of Games.* Princeton University Press, 1957.

Prasanta S. Bandyopadhayay, Robert J. Boik, Prasun Basu. "The Curve Fitting Problem: A Bayesian Approach." *Philosophy of Science.* Vol. 63, No. 3. Supplement. *Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association.* Part I: Contributed Papers. (Sep., 1996), pp. S264-S272.

Mario Bunge. "The Weight of Simplicity in the Construction and Assaying of Scientific Theories." *Philosophy of Science.* Vol. 28, No. 2. 1961. pp. 120-149.

Mario Bunge. "The Complexity of Simplicity." *The Journal of Philosophy.* Vol. 59, No. 5. 1962. pp. 113-135.

Nicholas Copernicus. *On the Revolutions of Heavenly Spheres.* Translated by Charles G. Wallis. Prometheus Books, 1995.

John Etchemendy. *The Concept of Logical Consequence.* Stanford, California: CSLI Publications. 1999.

Arthur Fine. "Natural Ontological Attitude." In *Philosophy of Science.* Ed. Papineau. Oxford University Press, USA, 1996.

Peter Fishburn. “On the Foundations of Game Theory: The Case of Non-Archimedean Utilities.” *International Journal of Game Theory.* Vol. 2, 1972. pp. 65-71.

R.A. Fisher. “On the Mathematical Foundations of Theoretical Statistics.” *Philosophical Transactions of the Royal Society.* A 222. pp. 309-368.

Richard Foley. “What’s to Be Said for Simplicity?” *Philosophical Issues.* Vol. 3. Science and Knowledge, 1993. pp. 209-224.

Malcom Forster. “The New Science of Simplicity. ” *Simplicity, Inference and Modelling.* Eds. Arnold Zellner, Hugo. A. Keuzenkamp, and Michael McAleer. (Cambridge: Cambridge University Press, 2001). pp. 83-119.

Malcom Forster and Elliott Sober. “How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions.” *The British Journal for the Philosophy of Science.* Vol. 45. 1994. pp. 1 - 35.

Michael Friedman. “Explanation and Scientific Understanding.” *The Journal of Philosophy¿* Vol. 71, No. 1, 1974. pp. 5-19

Michael Friedman. *Foundations of Spacetime Theories: Relativistic Physics and Philosophy of Science.* Princeton University Press, 1983.

Galileo. *Dialogue Concerning Two World Systems: Ptolemaic and Copernican.* Translated by Stillman Drake. Second Edition. University of California Press. 1953.

Nelson Goodman. An Improvement in the Theory of Simplicity. *The Journal of Symbolic Logic* Vol. 14, No. 4, 1950. pp. 228-229.

Nelson Goodman. “New Notes on Simplicity.” *The Journal of Symbolic Logic.* Vol. 17, No. 3, 1952. pp. 189-191.

Nelson Goodman. “Axiomatic Measurement of Simplicity.” *The Journal of Philosophy.* Vol. 52, No. 24, 1955. pp. 709-722.

Nelson Goodman. “The Test of Simplicity.” *Science.* Vol. 128, No. 3331. 1958. pp. 1064-1069.

Nelson Goodman. “Recent Developments in the Theory of Simplicity.” *Philosophy and Phenomenological Research.* Vol. 19, No. 4. 1959, pp. 429-446.

Nelson Goodman. “Safety, Strength, Simplicity.” *Philosophy of Science.* Vol. 28, No. 2, 1961. pp. 150-151.

Clark Glymour. *Theory and Evidence.* Princeton University Press, 1980.

Gilbert Harman and Sanjeev Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory* MIT Press, 2007.

David Heath and William Sudderth. “On a theorem of de Finetti, Oddsmaking and Game Theory.” *Annals of Mathematical Statistics.* Vol. 43. No.6. 1972. pp. 2072-2077.

Carl Hempel and Paul Oppenheim. “Studies in the Logic of Explanation.” *Philosophy of Science.* Vol. 15, 1948. pp. 135-175.

Hempel, C. (1966) *Philosophy of Natural Science* (Englewood-Cliffs: Prentice Hall).

Christopher Hitchcock and Elliot Sober. “Prediction Versus Accommodation and the Risk of Overfitting.” British Journal for the Philosophy of Science. Vol. 55. 2004.

Harold Jeffreys. *Theory of Probability.* Oxford: Clarendon Press, 1961.

Kevin Kelly. *The Logic of Reliable Inquiry.* Oxford University Press, 1996.

Kevin Kelly. ”Ockhams Razor, Truth, and Information.” In *Handbook of the Philosophy of Information.* Ed. Johan van Behthem and Peter Adriaans. *Forthcoming.*

Kevin Kelly. ”Simplicity, Truth, and Probability.” Unpublished Draft, 2009.

Kevin T. Kelly. “A Topological Theory of Empirical Simplicity and its Connection to the Truth.” Unpublished Draft, 2009.

Kevin Kelly and Clark Glymour. “Why Probability Does Not Capture the Logic of Scientific Justification. *Contemporary Debates in the Philosophy of Science*, Ed. Christopher Hitchcock. Oxford: Blackwell, 2004 pp. 94-114.

Kevin Kelly and Conor Mayo-Wilson. “Causal Discovery, Retractions, and Their Minimization.” Unpublished Draft.

Kevin Kelly and Conor Mayo-Wilson. “Ockham Efficiency Theorem for Empirical Methods Conceived as Empirically-Driven, Countable-State Stochastic Processes.” Unpublished Draft.

Kevin Kelly and Oliver Schulte. “Church’s Thesis and Hume’s Problem,” in *Logic and Scientific Methods.* Ed. M. L. Dalla Chiara. Dordrecht: Kluwer, 1997, pp. 383-398.

John G. Kemeny. “A Logical Measure Function.” *The Journal of Symbolic Logic.* Vol. 18, No. 4, 1953. pp. 289-308.

John G. Kemeny. “The Use of Simplicity in Induction.” *The Philosophical Review.* Vol. 62, No. 3, 1953. pp. 391-408.

John G. Kemeny. “Two Measures of Complexity.” *The Journal of Philosophy.* Vol. 52, No. 24, 1955. pp. 722-733.

John G. Kemeny. “A New Approach to Semantics: Part I” *The Journal of Symbolic Logic,* Vol. 21, No. 1, 1956. pp. 1-27.

John G. Kemeny. “A New Approach to Semantics: Part II” *The Journal of Symbolic Logic,* Vol. 21, No. 1, 1956. pp. 149-161.

Philip Kitcher. “Explanation, Conjunction, and Unification.” *The Journal of Philosophy,* Vol. 73, No. 8, 1976. pp. 207-212.

Thomas Kuhn. *The Essential Tension: Selected Studies in Scientific Tradition in Change* University of Chicago Press, 1977.

Henry E. Kyburg, Jr. “A Modest Proposal Concerning Simplicity.” *The Philosophical Review.* Vol. 70, No. 3. 1961. pp. 390-395.

Larry Laudan. “A Confutation of Convergent Realism.” *Philosophy of Science,* Vol. 48, No. 1, 1981 pp. 19-49.

Isaac Levi. *Gambling with Truth: An Essay on Induction and the Aims of Science.* MIT Press, 1973.

David Lewis. *Counterfactuals.* Oxford: Basil Blackwell, 1973. pp. 87.

Deborah Mayo. *Error and the Growth of Experiment Knowledge.* University of Chicago, 1996.

Deborah Mayo and Aris Spanos. “Severe Testing as a Basic Concept in a NeymanPearson Philosophy of Induction.” *British Journal for the Philosophy of Science.* Vol. 57, 2006. pp. 323357.

Margaret Morrison. “Unifying Scientific Theories: Physical Concepts and Mathematical Structures.” Cambridge University Press, 2000.

Wayne Myrvold. “A Bayesian Account of the Virtue of Unification.” *Philosophy of Science¿* Vol. 70, 2003. pp. 399423.

Leonard Koellner Nash. *The Nature of the Natural Sciences.* Boston: Little, Brown. 1963.

Isaac Newton. *The Mathematical Principles of Natural Philosophy.* New York: Citadel Press. 1964.

Daniel Nolan. “Quantitative Parsimony.” British Journal for the Philosophy of Science. Vol. 48. 1997. pp. 329-343.

Ariel Osborne and Martin Rubinstein. *A Course in Game Theory.* MIT Press, 1994.

Karl Popper. *The Logic of Scientific Discovery.* (translation of Logik der Forschung). Hutchinson, London, 1959.

Ptolemy. *The Almagest.* In *Great Books of The Western World.* Ed. Robert Maynard Hutchins. 1958.

Hilary Putnam (1973). “Reductionism and the Nature of Psychology.” *Cognition* Vol. 2. pp. 131-146.

W.V.O Quine. “Simpler Theories of A Complex World.” *Synthese* Vol. 15. 1963. pp. 103-110.

Hans Reichenbach. *Experience and Prediction.* University of Chicago Press, 1938.

Joseph Kadane, Mark Schervish, Teddy Seidenfeld. *Rethinking the Foundations of Statistics.* Cambridge University Press, 1999.

Note on Noncooperative Convex Games. *Pacific Journal of Mathematics.* Vol. 5, Suppl. 1, 1955. pp. 807-815.

Leonard Savage. *The Foundations of Statistics.* Second Edition. Dover Publications, 1972.

Rene Schilling. *Measures, Integrals, and Martingales.* Cambridge University Press, Cambridge. 2005.

George Schlesinger. “Dynamic Simplicity.” *The Philosophical Review,* Vol. 70, No. 4, 1961. pp. 485-499.

Oliver Schulte. “The Logic of Reliable and Efficient Inquiry,” *The Journal of Philosophical Logic*, Vol. 28, 1999. pp. 399-438.

Oliver Schulte. “Means-Ends Epistemology,” *The British Journal for the Philosophy of Science*. Vol. 50, 1999. pp. 1-31.

Oliver Schulte, Wei Luo, W. and Griner. “Mind Change Optimal Learning of Bayes Net Structure.” Unpublished Manuscript. 2007.

Simon, H. Science Seeks Parsimony, not Simplicity: Searching for Pattern in Phenomena. (In A. Zellner, H. Keuzenkamp, and M. McAleer (Eds.) *Simplicity, Inference and Modelling.* Cambridge: Cambridge University Press. 2001. pp. 83-119

Elliot Sober. *Simplicity.* Oxford: Clarendon Press, 1975.

Elliot Sober. ‘*Reconstructing the Past: Parsimony, Evolution, and Inference.* Cambridge: MIT Press, 1988.

Elliot Sober. “What is the Problem of Simplicity?” In *Simplicity, Inference and Modelling.* Eds. Arnold Zellner, Hugo. A. Keuzenkamp, and Michael McAleer. Cambridge: Cambridge University Press, 2001. pp. 13-32.

Elliot Sober. “Parsimony” Forthcoming In *The Philosophy of Science - An Encyclopedia*. Ed. S. Sarkar. Routledge, 2005.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Second Edition. Cambridge: MIT Press, 2000.

Aris Spanos. "Parametric versus Non-Parametric Inference: Statistical Models and Simplicity." In *Simplicity, Inference and Modelling.* Eds. Arnold Zellner, Hugo. A. Keuzenkamp, and Michael McAleer. Cambridge: Cambridge University Press, 2001. pp. 83-119.

Patrick Suppes. "Nelson Goodman on the Concept of Logical Simplicity." *Philosophy of Science.* Vol. 23, No. 2, 1956. pp. 153-159.

Lars Svenonius. "Definability and Simplicity." *Journal of Symbolic Logic.* Vol. 20, 1955. pp. 235-250.

Charles Swartz. *Measure, Integration and Function Spaces.* World Scientific, 1994.

Peter Turney. "The Curve Fitting Problem: A Solution." *British Journal for the Philosophy of Science.* Vol. 41. 1990. pp. 509-530.

Peter Turney. "A Theory of Cross-Validation Error." *The Journal of Theoretical and Experimental Artificial Intelligence.* Vol. 6. 1994. pp. 361-392.

Bas van Fraassen. *The Scientific Image.* Oxford University Press, 1980.

Vladamir Vapnik. *The Nature of Statistical Learning Theory.* New York: Springer, 2000.

Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications.* (New York: Springer Verlag, 1997).

Ming Li and Paul Vitanyi. *Simplicity, Information, Kolmogorov Complexity, and Prediction. In Simplicity, Inference and Modelling.* Eds. Arnold Zellner, Hugo A. Keuzenkamp, and Michael McAleer. (Cambridge: Cambridge University Press, 2001).

William Whewell. *Philosophy of the Inductive Sciences.* Volumes I and II. Johnson Repint Corporation, 1966.

Jiji Zhang. "Underdetermination of Statistical Quantities by Statistical Data." Unpublished. 2007.

Dorothy Wrinch and Harold Jeffreys. "On Some Aspects of the Theory of Probability." *Philosophical Magazine* Vol. 38. 1919, pp. 715-731.

Dorothy Wrinch and Harold Jeffreys. "On Certain Fundamental Principles of Scientific Inquiry." *Philosophical Magazine.* Vol. 42. 1921. 369-390,

Dorothy Wrinch and Harold Jeffreys. "On Certain Fundamental Principles of Scientific Inquiry." *Philosophical Magazine.* Vol. 45. 1923. 368-374.