Notre Dame Law Review

Volume 79 | Issue 2 Article 2

2-1-2004

Scientific Activism and Restraint: The Interplay of Statistics, Judgment, and Procedure in Environmental Law

David E. Adelman

SCIENTIFIC ACTIVISM AND RESTRAINT: THE INTERPLAY OF STATISTICS, JUDGMENT, AND PROCEDURE IN ENVIRONMENTAL LAW

*David E. Adelman**

INTRODUCTION

Americans have a love-hate relationship with science.[1] We receive minute-by-minute updates of the Dow Jones Industrial Average and the NASDAQ, read news accounts of recent medical studies on the risks from diseases, and track scientific reports estimating the annual increase in average global temperature from greenhouse gases. All of these numbers involve elaborate forms of statistical analysis and testing, and we rely to a great extent on the good faith and presumed objectivity of the scientists and economists involved to ensure that the numbers are accurate. As the recent accounting scandals and space shuttle disaster in the United States have demonstrated, however, defining statistical measures and ensuring accuracy are far from trivial tasks—even in areas many people assume are readily amenable to

* Associate Professor, James E. Rogers College of Law, University of Arizona; B.A., Reed College, 1988; Ph.D., Stanford University, 1993; J.D., Stanford Law School, 1996. For helpful comments, the author thanks Graeme Austin, John Barton, Ellen Bublick, Robert Glennon, Stephen Goldberg, Greg Mandel, Toni Massaro, Jamie Ratner, Carol Rose, Dalia Tsuk, Elliot Weiss, and David Wexler; special thanks to David Kaye, Mark Kelman, and Ted Schneyer.

1 Recent surveys suggest that many Americans have conflicting attitudes towards science. A 2002 report published by the National Science Foundation (NSF) found that "[i]n general, Americans express highly favorable attitudes toward [science and technology]." NAT'L SCIENCE BOARD, SCIENCE AND ENGINEERING INDICATORS—2002, at 7-12 (2002), *available at* http://www.nsf.gov/sbe/srs/seind02/pdfstart.htm. At the same time, many Americans believe that "people would do better by living a simpler life without so much technology" (44%) and that "technological discoveries will eventually destroy the Earth" (about 30%). *Id.* at 7-13. Even more striking, approximately 50% of the people surveyed agreed that "[w]e depend too much on science and not enough on faith." *Id.* Moreover, for hotly contested technologies, such as genetic engineering and nuclear power, public confidence in science is much lower. *Id.* at 7-16 to 7-17, 7-21.

meaningful quantification.[2] These kinds of failures have led Americans to question the reliability and value of such scientific methods and to question the role of science in society more generally.[3]

Similar schisms exist over how science is used in setting environmental policy. For most critics of environmental regulation, broad reliance on science is viewed as progress towards increased rationality and objectivity.[4] For many environmentalists, however, the presumed authority of science has engendered far greater concern and opposition. Environmentalists argue that science is too uncertain to resolve regulatory choices, that it is inaccessible to the general public, and

---

2 *See* Kurt Eichenwald, *Pushing Accounting Rules to the Edge of the Envelope*, N.Y. TIMES, Dec. 31, 2002, at C1 (commenting that "quarterly financial reports are far from the precise things that investors dream them to be"); Michael H. Granof & Stephen A. Zeff, *Generally Accepted Accounting Abuses*, N.Y. TIMES, June 28, 2002, at A27 (observing that despite what investors might like to believe, "accounting rules are hardly objective[;] [t]hey are open to interpretation—and, of course, manipulation—and there is often more than one reasonable way to measure expenses or revenues"); John Schwartz, *Computer Program That Analyzed Shuttle Damage Was Misused, Engineer Says*, N.Y. TIMES, Aug. 25, 2003, at A9 (concluding that a computer program "helped NASA mistakenly decide that the shuttle Columbia had not been deeply harmed"); John Schwartz & Matthew L. Wald, *Echoes of Challenger: Shuttle Panel Considers Longstanding Flaws in NASA's System*, N.Y. TIMES, Apr. 13, 2003, at A27 (noting that "the seemingly hard numbers of the Boeing analysis [of the impact of falling debris] appeared at the time to trump the gut feelings of the engineers" and may have contributed to the loss of the space shuttle).

3 NAT'L SCIENCE BOARD., *supra* note 1, at 7-16 to 7-17, 7-21 to 7-23 (noting that public support for space exploration dropped dramatically following the Challenger space shuttle accident and that recent negative publicity regarding genetic engineering has contributed to its diminishing level of public acceptance).

4 *See, e.g.*, BJORN LOMBORG, THE SKEPTICAL ENVIRONMENTALIST: MEASURING THE REAL STATE OF THE WORLD 348 (2001) (urging the need for a "strong regulation system" based on sound science); Stephen Breyer, *The Interdependence of Science and Law*, 280 SCIENCE 537, 537–38 (1998) (observing that "there is an increasingly important need for law to reflect sound science"); Frank B. Cross, *The Subtle Vices Behind Environmental Values*, 8 DUKE ENVTL. L. & POL'Y F. 151, 151 (1997) ("Reliance on science is broadly consistent with liberty and democracy. These values of the scientific method are far more valid than some of the values underlying public risk perceptions."); John D. Graham, *Legislative Approaches to Achieving More Protection Against Risk at Less Cost*, 1997 U. CHI. LEGAL F. 13, 41–43 (discussing, as current Administrator of the Office of Information and Regulatory Affairs at the Office of Management and Budget, the merits of a sound science approach to environmental regulation); William Reilly, *Taking Aim Toward 2000: Rethinking The Nation's Environmental Agenda*, 21 ENVTL. L. 1359, 1362 (1991) (commenting, as the former Administrator of the Environmental Protection Agency, that "sound science is our most reliable anchor in a turbulent sea of environmental policy and regulation").

that it is often used to mask questions of social values.[5] Underlying much of this concern is a general skepticism of scientific expertise and opposition to the antidemocratic overtones of delegating decisionmaking authority to a scientific elite. Yet, in spite of these conflicts, science continues to dominate environmental decisionmaking. The "best available science," for example, is the established standard for regulatory decisions under many environmental statutes,[6] and science is used in political and legal battles to justify, defend, and challenge environmental laws. Moreover, scientific methods, such as cost-benefit analysis and risk assessment, have become the common metrics of environmental regulation.

The controversy over the role of science in environmental policymaking is at base about the proper scope of scientific discretion. The debate first ignited around risk assessment methods used to derive regulatory standards for industrial chemicals.[7] Risk assessment created a furor because uncertainties in risk estimates are often very

5 *See, e.g.*, Adam Babich, *Too Much Science in Environmental Law*, 28 COLUM. J. ENVTL. L. 119, 126 (2003) (arguing that "current scientific theories about risk make a poor starting point for regulatory standard setting"); Devra Lee Davis, *The "Shotgun Wedding" of Science and Law: Risk Assessment and Judicial Review*, 10 COLUM. J. ENVTL. L. 67 (1985) (describing the misuse and limitations of risk assessment in the context of judicial review); Sheila Jasanoff & Dorothy Nelkin, *Science, Technology, and the Limits of Judicial Competence*, 214 SCIENCE 1211, 1213 (1981) (commenting that scientific uncertainty can mistakenly "lead both scientists and regulators to recommend inaction" and often may "encourage litigants to translate questions of social value into technical discourse"); Howard Latin, *Good Science, Bad Regulation, and Toxic Risk Assessment*, 5 YALE J. ON REG. 89, 90 (1988) ("challeng[ing] the conventional view that scientific perspectives should dominate the risk-assessment process"); Thómas O. McGarity, *Substantive and Procedural Discretion in Administrative Resolution of Science Policy Questions: Regulating Carcinogens in EPA and OSHA*, 67 GEO. L.J. 729, 781 (1979) (arguing that federal regulation of carcinogens cannot be dictated by science, but must instead be resolved using a results-oriented approach); Wendy E. Wagner, *Congress, Science, and Environmental Policy*, 1999 U. ILL. L. REV. 181, 181 (asserting that "Congress has put too much emphasis on scientific data—operating under the mistaken belief that science, alone, can provide the solutions to environmental problems"); Wendy E. Wagner, *The Science Charade in Toxic Risk Regulation*, 95 COLUM. L. REV. 1613, 1629 (1995) [hereinafter Wagner, *Toxic Risk Regulation*] (arguing that "[a]gency scientists and bureaucrats engage in a 'science charade' by failing first to identify the major interstices left by science in the standard-setting process and second to reveal the policy choices they made to fill each trans-scientific gap").

6 *See* Endangered Species Act of 1973, 16 U.S.C. § 1533(b)(1)(A) (2000) ("best scientific and commercial data available"); Clean Water Act of 1977, 33 U.S.C. § 1314(a)(1) ("reflecting the latest scientific knowledge"); Safe Drinking Water Act, 42 U.S.C. § 300g-1(b)(3)(A) ("best available, peer-reviewed science"); Clean Air Act, 42 U.S.C. § 7408(a)(2) ("reflect the latest scientific knowledge").

7 *See supra* note 5.

large at the low exposure levels most relevant to regulatory standards. Opposition also was heightened because, unlike many science-policy disputes, the problem is easily understood: The harm from most toxic chemicals at low exposure levels falls below the detection limits of existing testing methods. As a result, scientists must make judgments, or educated guesses, about the behavior of a chemical's toxicity at low exposure levels to fill this gap in the data. Environmentalists have cried foul on the ground that such judgments are matters of social value that lie outside the jurisdiction of scientific expertise, and objected to risk methods because they transform a question of policy into obscure technical details.[8] Regulatory critics are also hostile towards broad scientific discretion, but they view risk assessment methods as the solution, not the problem.[9] Their ire is directed at fuzzy qualitative science, which they describe pejoratively as "junk science" and consider to be presumptively suspect. For regulatory critics, risk assessment methods are by definition legitimate because they are quantitative and inductive.[10]

The positions on both poles of this debate are one-sided. Most significantly, they fail to recognize that good science encompasses logically deductive and inductive methods, as well as broad scientific values, such as simplicity, breadth, and consistency.[11] In other words, science involves a mix of qualitative judgments (based on broad principles) and quantitative data. This Article deviates from the standard debate insofar as it focuses on the underlying scientific methods them-

8 *See* Jasanoff & Nelkin, *supra* note 5, at 1213; Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1629.

9 *See generally* RISKS, COSTS, AND LIVES SAVED: GETTING BETTER RESULTS FROM REGULATION (Robert W. Hahn ed., 1996) (criticizing the current system where regulatory policy is set based on political pragmatism and highlighting key issues underlying risk assessment); Graham, *supra* note 4, at 14 (calling for Congress to pass a statute requiring the Executive Office of the President to create an "explicit, rigorous process of priority setting" among and within the various risk protection agencies).

10 This general veneration of risk assessment methods should not be read to imply that regulatory critics are not critical of certain approaches to risk assessment, such as the use of overly conservative assumptions. *See, e.g.*, Albert L. Nichols & Richard J. Zeckhauser, *The Perils of Prudence: How Conservative Risk Assessments Distort Regulation*, 10 REGULATION 13, 17–19 (Nov./Dec. 1986) (describing how regulatory policy is distorted by the use of conservative risk assessments); W. Kip Viscusi, *Regulating the Regulators*, 63 U. CHI. L. REV. 1423, 1437 (1996) (explaining how agencies use conservative risk estimates that "institutionalize an irrational form of economic behavior").

11 Thomas S. Kuhn considered both rigorous methods, which he referred to as normal science, and "idiosyncratic factors" based on these scientific values to be essential ingredients of scientific progress. THOMAS S. KUHN, THE ESSENTIAL TENSION 329–31 (1977).

selves, not purported misapplications of them. The Article examines the role of scientific discretion in environmental law by evaluating how scientific judgment is shaped by scientific methods—particularly statistical techniques. Three stages of scientific judgment will be examined: (1) reducing experimental information to quantitative results; (2) drawing inferences from discrete scientific studies; and (3) integrating the results of multiple experimental studies for purposes of setting environmental policies.

Statistical methods are of central importance to scientific judgment and to this Article. I will argue that statistics should be understood as a formal system for structuring how expert judgment is integrated into scientific determinations. Two theories of statistical inference will be described: "frequentism," which is based on methods for controlling and minimizing error rates in statistical models; and "Bayesianism," which is founded on a theorem for combining experimental results and scientific judgments that satisfies certain principles of logical coherence.[12] The two theories differ in their basic analytical approach. Frequentism uses objective standards of "statistical significance" to ensure that statistical methods stringently test scientific hypotheses, whereas Bayesian methods derive the probability that a hypothesis is true based on subjective scientific judgments and the available data. These differences are pivotal because they result in divergent approaches to conducting and interpreting statistical analyses and because they generate results that often differ substantially.[13] The important elements to appreciate initially are the basic contrasts between the two theories: frequentism is defined by measures of statistical significance, objective frequencies, and rigorous testing; Bayesianism is based upon direct probabilities, subjective judgments, and logical coherence. These differences cause the two methods to incorporate expert judgments into scientific assessments very differently.

Describing and explaining scientific methods, as the preceding paragraph evidences, presents its own set of challenges. I hope to minimize these impediments by drawing on parallels that exist between legal and scientific methods (some of which are interesting in their own right). The basic premise of this approach is that scientific and judicial judgment bear many similarities in how they are struc-

12 *See* IAN HACKING, AN INTRODUCTION TO PROBABILITY AND INDUCTIVE LOGIC 127–28, 171–73, 190 (2001); M.S. Bartlett, *Probability and Chance in the Theory of Statistics*, 141 PROC. ROYAL SOC'Y LONDON 518, 528–29 (1933); *see also infra* Part IV.A.

13 The more common frequentist significance testing employs statistical model error rates to evaluate whether a scientific hypothesis is refuted by observed data. Bayesian analyses, in contrast, use experimental data to estimate the degree of belief, often referred to as epistemic probability, a hypothesis warrants.

tured. Important parallels exist at three levels: the interpretive theories that delimit judicial review and scientific judgment, the allocation of burdens of proof in judicial and scientific judgments, and the regulatory models and methodological rules that structure judicial and scientific judgments, respectively. These parallel features map directly onto the three stages of scientific judgment described above—interpretive principles onto stage one, burdens of proof onto stage two, and procedural rules and regulatory models onto stage three. The similarities between the legal and scientific frameworks will be drawn upon at each stage, as demonstrated below.

Stage 1. Interpretive theories of judicial review mirror those used in science. Judicial review is practiced according to interpretive principles that range from strict formalism to antiformalism.[14] Stated simply, formalism limits judges to the plain language of a statute or the Constitution; antiformalism denies that such a plain meaning can be discerned in most cases and presumes judges will invoke principles beyond the specific statutory or constitutional provisions in question.[15] Scientific positivism and relativism parallel these two branches of legal thought. In its most basic form, positivism views scientific facts as speaking for themselves and scientific methods merely as means for collecting and interpreting the plain meaning of data.[16] Relativism rejects this passive model of science, holding instead that scientific facts are neither univocal nor free-floating—where judges require external principles to resolve the meaning of a statutory provision, scientists require theories to make sense of experimental data.[17] Judges and scientists, in this light, are both subject to being labeled activists and condemned for either their formalist or antiformalist methods. For example, environmentalists' objections to risk methods are in essence objections to a form of "scientific activism" in which scientists (improperly) extrapolate beyond a strict reading of the limited facts available. The legal and scientific debates both raise challenging questions about the proper scope of expert judgment and discretion.

---

14 Examples of strict formalism include texualism and originalism; examples of antiformalism include legal pragmatism and rights based theories. *See* JOHN HART ELY, DEMOCRACY AND DISTRUST 1–3 (1980); Paul Brest, *The Fundamental Rights Controversy: The Essential Contradictions of Normative Constitutional Scholarship*, 90 YALE L.J. 1063, 1064–65 (1981); Thomas C. Grey, *Judicial Review and Legal Pragmatism*, 38 WAKE FOREST L. REV. 473, 478 (2003).

15 *See* Brest, *supra* note 14, at 1064–65.

16 *See infra* Part II.B.

17 *See infra* Part II.B.

Stage 2. Burdens of proof in environmental law and science often overlap because environmental standards are generally based on scientific evidence. This connection is borne out by the fact that the controversy over regulatory burdens of proof has often focused on statistics. The other important element of the debate is the Precautionary Principle, which environmentalists have long used to argue that regulated industries should bear the burden of proving that their products and activities are safe.[18] Environmentalists object to traditional statistical methods (i.e., frequentist) because they improperly place the scientific burden of proof on proponents of environmental regulation by implicitly starting with a baseline assumption that no harm exists.[19] Environmentalists consequently consider the bias of frequentist methods to be a central impediment to their decades-long effort to reform methods of scientific inference for purposes of environmental standard setting. These objections raise important questions about how statistical methods are used by scientists and the relationship between standards of statistical significance and legal burdens of persuasion.

Stage 3. It is a simple truism that legal rules and scientific methods shape how scientific standards are established in environmental law. More significant, however, is the observation that the two central models found in law and science are structurally similar—in both contexts, one model is procedurally oriented and the other expertise based. For legal academics, risk assessment methods are the example of choice in the contest between the two models. For example, Supreme Court Justice Stephen Breyer and professor Wendy Wagner have both used toxics risk assessment to illuminate the virtues of each approach.[20] Breyer argues for an expert model based on establishing a politically insulated scientific elite within the government to resolve difficult scientific problems.[21] Wagner advocates a procedural approach that is designed to ensure that matters of scientific fact and environmental policy are made transparent and clearly distinguished from each other.[22] The central difference between their proposals is that where Wagner seeks to separate the science from the policy, Breyer lumps expert judgment together with empirical observations.

Scientists also have an obvious interest in controlling how scientific judgments are made and used in environmental policymaking.

18 *See infra* Part III (describing the Precautionary Principle).

19 *See infra* Part III.

20 STEPHEN BREYER, BREAKING THE VICIOUS CIRCLE: TOWARD EFFECTIVE RISK REGULATION 42–50 (1993); Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1613–17.

21 BREYER, *supra* note 20, at 59–61, 80–81.

22 Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1701–03, 1711–20.

The focus of the scientific debate, however, is on statistical methods—one side propounding a frequentist approach, the other a Bayesian.[23] Frequentists occupy the proceduralist camp with their objective testing methods and attention to analytical transparency.[24] Bayesians, in the other, embrace an expert model that integrates scientific judgments directly into statistical analyses.[25] Among scientists, climate change research has become a flash point for this methodological standoff. For example, Professor Stephen Schneider, a prominent climate scientist and environmentalist, has argued tirelessly on behalf of Bayesian methods because he believes they ensure that the judgments of scientific experts are adequately considered by policymakers.[26] Schneider's campaign has met with significant opposition. Frequentist critics claim that Bayesian methods will further politicize environmental science because they incorporate subjective judgments, which naturally reflect a scientist's personal biases.[27] Policymakers will, as a result, have to contend with a plethora of Bayesian estimates that vary from scientist to scientist, rather than a single frequentist analysis of the data. The choice between frequentist and Bayesian methods therefore implicates the debate over regulatory models used in environmental policymaking, and vice versa.

This Article examines the role of scientific discretion in environmental law and describes the interplay between scientific judgment,

23 *See, e.g.*, COLIN HOWSON & PETER URBACH, SCIENTIFIC REASONING: THE BAYESIAN APPROACH (1989) (explaining the analytical underpinnings of Bayesian methods and arguing that it is a superior method); DEBORAH G. MAYO, ERROR AND THE GROWTH OF EXPERIMENTAL KNOWLEDGE (1996) (providing a standard book on the analytical underpinnings of frequentist methods); Brian Dennis, *Should Ecologists Become Bayesians?*, 6 ECOLOGICAL APPLICATIONS 1095 (1996) (analyzing the weaknesses and limitations of Bayesian methods); Aaron M. Ellison, *An Introduction to Bayesian Inference for Ecological Research and Environmental Decision-Making*, 6 ECOLOGICAL APPLICATIONS 1036 (1996) (describing the virtues of Bayesian analysis relative to traditional frequentist methods); Jim Giles, *When Doubt is a Sure Thing*, 418 NATURE 476 (2002) (reporting on the dispute over Bayesian and frequentist methods among climate scientists); David Malakoff, *Bayes Offers a 'New' Way to Make Sense of Numbers*, 286 SCIENCE 1460 (1999) (reporting on the rise of Bayesian methods by scientists); Stephen H. Schneider, *What Is 'Dangerous' Climate Change?*, 411 NATURE 17 (2001) (advocating the use of Bayesian methods in climate change policy); Lara J. Wolfson et al., *Bayesian Environmental Policy Decisions: Two Case Studies*, 6 ECOLOGICAL APPLICATIONS 1056 (1996) (demonstrating how Bayesian methods can be used in environmental policymaking).

24 MAYO, *supra* note 23, at 10.

25 *See supra* note 12 and accompanying text.

26 *See* Giles, *supra* note 23, at 476–77; Schneider, *supra* note 23, at 18.

27 *See* Giles, *supra* note 23, at 477–78. A scientist from the private sector, for example, will presumably derive very different Bayesian estimates of pollutant levels in a stream than a scientist from an environmental group.

statistics, and procedure in environmental policy. Part I of the Article sets the stage with a brief introduction to statistics. The core sections of the Article, Parts II through IV, follow the three-stage framework described above. Part II discusses the legal and scientific debates over quantitative methods in environmental policy and challenges the dominant theories of science found in legal scholarship. I reject the prevailing theories in favor of an experimentally grounded approach that acknowledges the central role of qualitative judgments in science and the difficult balancing that is required to protect the integrity of science while ensuring transparency and political accountability. Part III addresses the debate over burdens of proof in environmental law, and uses several rationales for the Precautionary Principle to evaluate common misconceptions and concerns about the use of frequentist methods in environmental science. I show that statistical tests are more flexible than most people appreciate and propose a solution to environmentalists' concerns—"equivalence testing"—that reverses the benign-until-proven-guilty presumption of traditional frequentist methods.[28] Finally, Part IV evaluates the competing virtues of Bayesian and frequentist methods for structuring how scientific judgments are made and analyzes their parallels with the legal models Breyer and Wagner propose. Both statistical theories are found to have practical limitations and to present important interpretive challenges. I conclude by describing an alternative to the two legal models that alleviates the shortcomings of the statistical methods by integrating them with standard legal procedures. This approach has two additional benefits: it enhances the range of options available to guide scientific judgment in environmental policy, and it promotes a deeper appreciation on the part of lawyers and policymakers for the limits and strengths of scientific methods.

I. INTRODUCTION TO BAYESIAN AND FREQUENTIST METHODS OF STATISTICAL INFERENCE

Statistics is often mistakenly viewed as a collection of related techniques that lack any substantive content.[29] Statistics in fact consists of

28 *See* Roger L. Berger & Jason C. Hsu,. *Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets*, 11 STAT. SCI. 283, 283–84 (1996); Graham B. McBride, *Equivalence Tests Can Enhance Environmental Science Management*, 41 AUSTL. & N.Z. J. STAT. 19, 20 (1999).

29 Statistics is no more a collection of descriptive techniques void of theoretical content than legal procedures are independent of substantive objectives. Yet, as early as the Progressive Era, statisticians sought to separate their technical work from its potential political implications. *See* THEODORE M. PORTER, THE RISE OF STATISTICAL THINKING 1820–1900, at 23–39 (1986). "Partly as a defensive move, and partly to reas-

certain mathematical theorems and models of scientific inference that are premised on substantive beliefs about nature.[30] It also functions in two distinct modes. First, statistics encompasses a collection of mathematical techniques (e.g., means, medians, probability functions) that are used to analyze observed propensities in experimental systems, such as the likelihood of rolling double sixes with a set of dice.[31] In this mode, statistics is used to evaluate the results of multiple observations, such as calculating the mean concentration of a pollutant in a river from multiple test sites. Second, statistics is used to make probability estimates for scientific inference, which are most commonly associated with traditional methods for determining

---

sure interested political leaders that their support of statistics would not embarrass them, the statist[icians] adopted the position that they were concerned exclusively with facts." *Id.* at 35. Neither the logic nor theory of statistics, however, supported statisticians' denials that statistical methods entailed substantive assumptions or values. *Id.*

30 Randall Collins, *Statistics Versus Words*, 2 SOC. THEORY 329, 331 (1984). As one commentator has put it:

> [*A*]*ll* principles of theory evaluation [i.e., experimental testing] make some substantive assumptions about the structure of the world we live in *and* about us as thinking, sentient beings. The difference between procedural and substantive methodological rules is thus entirely a matter of degree and of context. And as soon as we acknowledge that point, it becomes clear that the cogency of any methodological principle is, at least in part, hostage to the vicissitudes of our future interactions with the natural world. But that is just another way of saying that methodologies and theories of knowledge are precisely that, viz., *theories.*

LARRY LAUDAN, BEYOND POSITIVISM AND RELATIVISIM 171 (1996).

31 The simple roll of a fair die involves a stochastic system that is governed purely by chance or random process. In complex real world settings, however, the concept of chance often reflects both our state of knowledge and a characteristic of reality. *See* JOHN EARMAN, BAYES OR BUST? 54 (1992). Chance can arise when we do not, and perhaps cannot as a practical or epistemological matter, know the starting conditions or sequence of actions that caused an event. If while hiking I am hit by a falling tree branch, at least two independent chains of causation precipitated the event: the actions that led me to be hiking in the particular place at the time the branch fell and the events that led to the collapse of the branch. In this example, two independent causal chains (my decision to go hiking and the branch failure) converged to cause the chance event. Importantly, the system of interactions in this example is not completely random; in fact, much predictive order exists, for example, in the tree branch's failure, even if not all of the necessary information is available. If it were possible to reconstruct each of these causal chains, one could fully explain the causes of the event. Accordingly, chance is not limited to "the absence of causality," but often incorporates our ignorance of causality based on what we choose, or are able, to observe or test. Collins, *supra* note 30, at 332, 350.

whether an experimental result is statistically significant.[32] In this second mode, statistical methods are used to determine whether or not certain data support a particular scientific hypothesis, such as a theory about the health risks from specific airborne pollutants; they do not function as a direct summary of the trends observed in the data, as in a median value or experimental uncertainty estimate.

As competing statistical theories, frequentist and Bayesian methods differ in several import respects. First, frequentist methods span both the data summary and scientific inference modes, whereas Bayesian methods are limited to scientific inference. Second, while both theories are derived from the same basic axioms,[33] they adopt divergent approaches to defining probability and deriving statistical inferences. Frequentists define probability *objectively* as the "long-run frequencies" in a population.[34] The frequency, for example, that samples from a body of water exceed a regulatory limit or the incidence rate of a genetic defect in a population are representative of such properties.[35] Bayesians, in contrast, define probability as the level of confidence an individual has about an event or thing based on their *subjective* judgment.[36] Bayesian methods, for instance, could be used to combine expert judgments and observational data to determine the "conditional probability" that air emissions from a power plant are

32 *See* IAN HACKING, THE EMERGENCE OF PROBABILITY 1, 11–16 (1975) (explaining that probability is, on the one hand "statistical, concerning itself with stochastic laws of chance processes," but on the other hand, "is epistemological, dedicated to assessing reasonable degrees of belief in propositions quite devoid of statistical background").

33 HACKING, *supra* note 12, at 135. Three central axioms, often referred to as Kolmogoroff's axioms, form the basic logical framework of all probability theory. The framework is structured such that for an event E on a trial of kind K, the probability of E, $P(E)$, obeys the following axioms: (1) $0 \leq P(E) \leq 1$; (2) $P(\Omega) = 1$, where Ω is a sure outcome; and (3) $P(E \cup F) = P(E) + P(F)$, where E and F are mutually exclusive (i.e., do not overlap)—this axiom generalizes as $P(\Sigma E_n) = \Sigma P(E_n)$ if each E_n is mutually exclusive of the others. The expression $P(E \cup F)$ is the probability of events E and F both occurring. IAN HACKING, LOGIC OF STATISTICAL INFERENCE 18–19 (1965). Acceptance of these axioms is virtually universal and no attempt will be made to justify them here.

34 *See* HACKING, *supra* note 12, at 144–45, 190–91; HACKING, *supra* note 33, at 2.

35 In the first case, the body of water contains a specific concentration (i.e., long-run frequency) of the chemical. In the second case, a general population exists, all humans, that has a (presumably) stable subpopulation with the genetic defect, and the long-run frequency is the subpopulation's size divided by the total population's size. A frequentist would take multiple samples of the water and study sample human populations to obtain estimates of these long-run frequencies, and use these data as a basis for statistical inference.

36 HACKING, *supra* note 12, at 131–32, 140–44.

harmful to human health.[37] Third, frequentist probabilities do not vary from person to person, whereas Bayesian probabilities may—even if based on the same data. Scientists could, for example, derive very different Bayesian estimates for the air pollution levels in a city using the same monitoring data.

Bayesian and frequentist methods of statistical inference are mathematically distinct in the following respect: the Bayesian "perspective fixes on one fundamental logical property of the probability rules—Bayes's [theorem]. The frequency perspective fixes on another fundamental logical property of the probability rules—laws of large numbers."[38] In the former, Bayes's theorem operates as a logical algorithm for incorporating evidence into a probability estimate.[39] For example, if you believe at the start of the season that your favorite baseball team has a 30% chance of winning the World Series, Bayes's theorem provides a logically consistent means of revising this estimate during the course of the baseball season as your team's record evolves. In the latter, the approximations made possible by the mathematical law of large numbers are integral to the derivation of frequentist statistical testing methods and to their analytic power even when the available data are limited.[40] The central point to grasp here is that Bayesian

37 "Conditional probability" signifies that the probability estimate is conditioned (i.e., based on or relative to) an initial judgment about the system being studied and the available data. DAVID HOWIE, INTERPRETING PROBABILITY 30–31 (2002). Bayesians focus on the data, not statistical samples of abstract populations. Bayesians begin with a mathematical estimate of the statistical results, referred to as a "prior probability," and use observational data to refine the starting predictions and to quantify directly the probability (e.g., 60%) of the violation occurring. *Id.*

38 *See* HACKING, *supra* note 12, at 190. The probability rules referred to here are the basic axioms of probability theory, from which Bayes's theorem is derived directly, and the law of large numbers refers to a critical mathematical simplification (valid for large samples) that was instrumental in deriving the normal distribution.

39 As a number of legal academics have urged, Bayes's theorem could be used to assist juries in weighing and integrating the evidence in a case. *See, e.g.*, MICHAEL O. FINKELSTEIN, QUANTITATIVE METHODS IN LAW 87–98, 103–04, 289–310 (1978); PROBABILITY AND INFERENCE IN THE LAW OF EVIDENCE (P. Tillers & E.D. Green eds., 1988); David L. Faigman & A.J. Baglioni Jr., *Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence*, 12 LAW & HUM. BEHAV. 1, 16 (1988) (asserting that "an expert's Bayesian formulation will not overwhelm the average trier of fact," and courts should focus less on the potential of overwhelming the jury and more on impressing the relevance of techniques). *But see* Ronald J. Allen & Brian Leiter, *Naturalized Epistemology and the Law of Evidence*, 87 VA. L. REV. 1491, 1493 (2001) ("employ[ing] the naturalized epistemology approach to criticize existing theories of different evidentiary rules, including Bayesianism").

40 HACKING, *supra* note 12, at 190, 196–98.

and frequentist methods are based on distinct mathematical theorems, which dictate how they are applied.

The two statistical theories also reflect opposing philosophies for scientific inference: Bayesian methods are inductive while frequentist methods are hypothetical-deductive. Take the example of assessing the likelihood of global warming during the next decade. A Bayesian would use Bayes's theorem to combine the available information, including the judgments of scientific experts, and derive inductively the probability that global warming will occur (e.g., it is 60% likely). A frequentist, in contrast, would start with a "null hypothesis" that global warming will not occur and then conduct an experiment to test whether this null hypothesis is consistent with the collected data. If the experimental data are inconsistent with the null hypothesis, the result is simply characterized as "statistically significant." Frequentist methods consequently do not quantify directly the likelihood of global warming; they function instead as a means for testing (i.e., falsifying) hypotheses. The more rigorous the statistical testing, the greater the confidence a scientist using frequentist methods will have in a hypothesis if it withstands such testing.

The historical development of statistics sheds further light on the differences between the Bayesian and frequentist approaches and the different roles statistical methods play in science. Statistics was used in two distinct ways during its early development: (1) as a method for combining experimental observations, and (2) as a means for drawing scientific inferences.[41] The theoretical unification of these roles took about 150 years and involved the work of some of the most brilliant mathematicians of this period.[42] The discussion that follows begins by describing work in the 1700s and early 1800s on the statistical analysis of experimental observations and then shifts to the late 1800s and early 1900s to consider the work of Sir Ronald Fisher and Sir Harold Jeffreys on statistical inference. Their views embody critical elements of frequentism and Bayesianism, respectively, and reveal how these distinctive traditions differ.

Astronomers working before the mid-1700s did not combine measurements, except where precisely replicated measurements were averaged together.[43] At the time, the virtue of obtaining multiple measurements under a range of conditions was completely misunderstood—scientists believed that combining observations would cause

41 STEPHEN M. STIGLER, THE HISTORY OF STATISTICS: THE MEASUREMENT OF UNCERTAINTY BEFORE 1900, at 4 (1986).

42 *Id.*

43 *Id.* at 30.

errors to multiply, not cancel.[44] However by 1805, the Method of Least Squares revealed that a data set's arithmetic mean represented "the center around which the results of observations arrange themselves, so that the deviations from that center are as small as possible."[45] The center of gravity identified by the mean value of multiple observations emboldened scientists to combine measurements, but did not allow experimental error to be estimated.[46] Experimental error estimation had to wait for the derivation of the "normal distribution," or "bell-shaped curve," which scientists recognized as modeling observations for which experimental errors are random.[47] Once this connection was made, the normal distribution became the standard mathematical model for calculating experimental error.[48]

Development of the normal distribution represented a turning point in statistical theory. The archetype for a random, or stochastic, system became a simple gas consisting of a large number of atoms randomly interacting such that their individual velocities over time are not causally connected.[49] The conditions of this simple atomic model—independent elements identically distributed—also define these terms.[50] Two aspects of an atomistic model, and thus the normal distribution, are particularly distinctive and powerful. First, the model provides a "zero point" against which to assess a system's order because the elemental components of an atomistic model are assumed to be completely uncorrelated.[51] Scientists realized that this property is crucial, as it can be exploited to assess causal relations in

---

44 *Id.* at 4.

45 *Id.* at 14–15.

46 *Id.* at 61, 139–40. Interestingly, some astronomers resisted statistical methods for calculating errors on the basis that it preempted their expertise in judging their own measurements. HOWIE, *supra* note 37, at 19.

47 STIGLER, *supra* note 41, at 141–45.

48 *Id.* at 157–58.

49 Collins, *supra* note 30, at 334–35. Randomness can also be defined more formally as (1) not being able to gamble a system; or (2) relative to an ideal computer, any program sufficient to generate the sequence would be at least as long as the sequence. HACKING, *supra* note 12, at 145.

50 *See* HACKING, *supra* note 33, at 20–21; Collins, *supra* note 30, at 333–35. In this context, "identically distributed" simply means that the velocities and magnitudes are symmetric about the mean value for the system.

51 Causal connections between elements of a system exist along a continuum, which ranges from full causal interdependence (e.g., carbon atoms in a diamond crystal) to complete causal independence (e.g., helium atoms in a balloon). Model frequency-type systems lie at the complete causal independence end of the scale, which is, in effect, an "absolute zero" for causal relations within a system.

experimental systems.[52] For example, if a scientist wishes to determine whether a chemical is harmful, she employs an atomistic model of her experimental conditions as a baseline against which to assess whether the chemical has a discernible effect.[53] Second, the model is accurately represented by two simple parameters, making it exceptionally easy to interpret.[54]

The synthesis of combining experimental observations and evaluating experimental error established the basic mathematical foundation for modern statistics. Calculating the error associated with observations, however, was just the start. Scientists soon realized that the normal distribution could be used to model a broad range of physical and social systems.[55] Social scientists were actually the first to grasp the significance of these methods beyond their use in experimental error analysis.[56] At the time, statistical methods were employed to identify regularities in social statistics, culminating in the use of the normal curve to construct an idealized "average man."[57] Economists also used the new methods to fulfill their aspirations for a "social mechanics" comparable to Newtonian mechanics in physics.[58] Statistical theory enabled them to translate a physical atomistic model into an economic theory in which utility became as fundamental to economic theory as energy was to theoretical physics.

---

52 It is nevertheless essential to keep in mind that "a statistical model is not simply a basis against which to test some other theory; it provides a model of the phenomena itself." Collins, *supra* note 30, at 348.

53 A corollary principle necessary for this (frequentist) approach to be valid is that only causally connected relations are observed over repeated observations, or in large systems, because causally unconnected relations average out (i.e., cancel). *See* PORTER, *supra* note 29, at 97.

54 HACKING, *supra* note 33, at 72–73 (noting that the two parameters are the mean, or average value, and variance of the distribution).

55 Because scientists through much of the nineteenth century believed that the physical world was deterministic, it was decades before physical science began to use statistics for purposes other than error analysis. HOWIE, *supra* note 37, at 41–42, 200. According to this perspective, determinism obviated the need for probabilistic models because the natural sciences would generate precise descriptions of natural phenomena. *Id.*

56 *See* HOWIE, *supra* note 37, at 37; PORTER, *supra* note 29, at 111–13 (noting it was ultimately the regularity of social statistics that inspired physical scientists to adopt statistical methods).

57 *See* STIGLER, *supra* note 41, at 169–71, 201, 215, 221.

58 *See* PORTER, *supra* note 29, at 257. By reifying an atomistic model of society, their approach abstracted frequentist theory by treating individual people atomistically, that is, as analogues of atoms in a simple gas system, and by defining utility as an analogue of energy in this simple physical model. *See id.* at 256–57 (noting that Edgeworth referred to utility as the "invisible energy of pleasure"); *see also* JOHN M. KEYNES, A TREATISE ON PROBABILITY 249 (1921).

The work of social scientists during the nineteenth century led to more sophisticated techniques, such as statistical correlation and regression, that transformed the experimental methods of the social sciences and soon found their way back into the physical sciences.[59] These subsequent developments in statistics generalized the approach taken in astronomy:

> Observations and statistics agree in being quantities grouped about a Mean; they differ, in that the Mean of observations is real, of statistics is fictitious. The mean of observations is a cause, as it were the source from which diverging errors emanate. The mean of statistics is a description, a representative quantity which, if we must in practice put one quantity for many, minimizes the error unavoidably attending such practice . . . . In short observations are different copies of one original; statistics are different originals affording one "generic portrait."[60]

The median height among a collection of people and mean concentration of an airborne chemical are representative of a statistical generic portrait or summary. The rising importance of statistical methods also exposed important differences in how probability was interpreted, precipitating the permanent division in statistics between the frequentist and Bayesian schools. Fisher's and Jeffreys's work and advocacy enhanced the debate over the proper interpretation of probability.

Ronald Fisher was instrumental in developing the formal methods for frequentist statistical inference and experimental design. According to Fisher, "science was a matter of random statistical aggregates, and the data representative of a population."[61] Fisher's view of science was deeply informed by his work in Mendelian genetics, which scientists have aptly characterized as nature's "perfect gambling machine."[62] Population genetics became the central metaphor of Fisher's work: Just as a human population contains many genetic subpopulations, so too is the universe made up of innumerable populations or classes of things, which experiments randomly "sample" to

---

59 *See* STIGLER, *supra* note 41, at 358–61.

60 *Id.* at 309.

61 HOWIE, *supra* note 37, at 164.

62 *Id.* at 61 ("Mendelism was unique in involving a chance mechanism that generated with exact and fixed probability one of a set of clearly-defined outcomes. Genetic probabilities could thus be *treated* as inherent to the world rather than reflecting incomplete knowledge.") (citing G.A. Barnard, *Reply to S.L. Zakell*, 4 STAT. SCI. 258, 259–60 (1989)).

determine their properties.[63] For Fisher, statistical inference involved obtaining a statistical sample of a population, such as sediment sampling points in a river, from which the fixed (i.e., objective) frequencies of the population were inferred.[64] Fisher's test for statistical significance provides a measure of the fidelity between an experimental sample statistic, such as a mean sediment contaminant level, and the corresponding parameter in the real world population, here the mean sediment contaminant level of every point in the river.[65] The great strength of Fisher's work was that his statistical tests were both simple to apply and valid for even relatively small experimental samples.

Harold Jeffreys was a physicist with an astonishing talent for developing mathematically tractable models for complex systems in geophysics, meteorology, and astrophysics.[66] As in astronomy, Jeffreys's work was observational, not experimental (i.e., not based on carefully controlled and replicated studies), "and the relevant data [were] often scarce and of variable quality."[67] These constraints led Jeffreys to integrate observations from diverse fields. "His inferences often tied geological and archeological results to astronomical observations and even considerations from atomic physics or the classical theory of waves. A frequency interpretation was simply unavailable: such data could not be regarded as from a specified [frequentist] population."[68]

As a result, Jeffreys viewed probability as a characteristic of imperfect knowledge and thus relative to the available information.[69] The nature of Jeffreys's scientific work also made it natural for him to integrate his scientific judgments with empirical observations. For Jeffreys, Bayesian analysis provided the formal theoretical framework to

63 *Id.* at 63. Under this theory, the statistical frequencies measured in an experiment do not represent the "credibility" of the result; they are the relative frequencies of the sample. *Id.*

64 *Id.* at 70, 74. Determining, for example, the average frequency, on a daily basis, of rain in a specific region draws a sample from an essentially infinite population consisting of days. Fisher's work was particularly remarkable insofar as it allowed scientists to infer general laws based on relatively small sample sizes. *See id.* at 71.

65 *Id.* at 63. Just as one would predict intuitively, the larger the sample size and better controlled the experiment, the better the sample statistic will approximate the parameter in the population being tested. *Id.* at 71.

66 *See id.* at 5, 82, 84.

67 *Id.* at 113.

68 *Id.* at 169.

69 *See id.* at 4, 8, 89. One of Jeffreys's seminal discoveries was that the Earth's core is molten. In this context, it made no sense to determine the long-run probability that this was true. Either it was or it was not molten—positing a hypothetical population of Earths from which to draw samples would have been absurd. *Id* at 8.

combine his scientific judgments with the available observational data to make scientific inferences.

The differences in Fisher's and Jeffreys's approaches demonstrate the close connection between substantive science and statistics.[70] Fisher began with a probabilistic, or stochastic, model of the universe analogized from Mendelian population genetics. He deduced from this model an objectivist approach to statistical inference based on rigorous testing protocols, under which confidence in a hypothesis is strengthened if it passes such tests.[71] Jeffreys's experience as a physicist, in contrast, led him to Bayesian analysis, which provided a logically coherent means of combining expert judgment and diverse empirical observations to derive conditional probability estimates for his physical theories.[72] The two approaches treat new information very differently. For Fisher, new data reduce statistical error rates by providing a larger sample that is more representative of the true population, and thus a stronger test for statistical significance. For Jeffreys, new information does not reduce statistical error rates, it alters—either up or down—one's degree of belief (i.e., the probability) that a hypothesis is true.

The work of Fisher and Jeffreys reveals how the frequentist and Bayesian approaches to probability are premised on two distinct conceptual foundations. Frequentism adopts a world view in which abstract populations are the building blocks of the universe. Under this framework, experimental science is simply a process of obtaining "random samples from a population of fixed distribution," much as one might take multiple samples of a gumball machine to estimate the relative abundance of the different flavors it contains.[73] As Fisher's work suggests, this model of reality was generalized from his experimental work in Mendelian genetics, not proven.[74] Bayesians reject Fisher's unproven presumption that the world is divided into abstract populations and opt instead for the logical coherence grounded in

70 The work of Fisher and Jeffreys also illustrates just how much "[e]ach interpretation of probability . . . [is] suited to a particular sort of [scientific] inquiry." *Id.* at 9.

71 Because Jeffreys viewed "populations" as "unobserved data," he rejected Fisher's form of statistical inference because it enabled "a hypothesis that may be true [to] be rejected because it has not predicted observable results that have not occured." *Id.* at 155 (quoting H. JEFFREYS, THEORY OF PROBABILITY 357 (2d ed. 1948)).

72 Jeffreys's Bayesian approach was pointedly criticized for creating a "wretched hybrid" of objective frequencies and subjective judgment (i.e., data, experience, and psychological factors) when what scientists wanted to compare were mathematical models to objective data—in short, experimental observations must be separated from scientific judgment. *See id.* at 161–62.

73 *Id.* at 37.

74 *See id.* at 107.

Bayes's theorem. Bayesians are not, however, able to avoid this metaphysical uncertainty, as it re-enters their analysis in the form of the scientific judgments on which Bayesian analyses are based. These points are crucial to appreciating the relationship between probability and judgment in the two theories. Probability and scientific judgment are distinct for frequentists because they treat probability as an objective property that is used to justify scientific judgments. Probability and scientific judgment are merged for Bayesians because they treat probability as a subjective property that incorporates subjective judgments directly into probability estimates. The two theories consequently incorporate scientific judgments in very different ways.

II. SCIENTIFIC JUDGMENT AND QUANTIFICATION IN ENVIRONMENTAL POLICY

Environmentalists object to statistical methods for two central reasons. First, statistical variables distort environmental policies by superimposing overly simplistic generic portraits of complex problems.[75] To give just one example, environmental risks of industrial chemicals are often reduced to assessments of human carcinogenicity, without any consideration of harms to wildlife, ecosystems, or non-cancer human health risks.[76] This type of distortion arises inexorably from statistical quantification, which in all but the most trivial systems sacrifices representation accuracy for analytic clarity and tractability.[77] Second, traditional frequentist methods of statistical infer-

75 *See* Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329, 1372–76 (1971) [hereinafter Tribe, *Trial by Mathematics*]; Laurence H. Tribe, *Ways Not to Think About Plastic Trees: New Foundations for Environmental Law*, 83 YALE L.J. 1315, 1329–32 (1974) [hereinafter Tribe, *Plastic Trees*].

76 Critics also charge that simplifying assumptions, such as functional linearity and continuity, are more misleading than enlightening and therefore should be critically re-evaluated. *See* Laurence H. Tribe, *Policy Science: Analysis or Ideology*, 2 PHIL. & PUB. AFF. 66, 73–74, 87–88, 92–93 (1972); Tribe, *Plastic Trees*, *supra* note 75, at 1331–32. A great deal of attention also has been directed at the values implicit in various simplifying assumptions in cost-benefit analysis, such as the use of economic discounting and risk metrics in environmental policy. *See* Frank Ackerman & Lisa Heinzerling, *Pricing the Priceless: Cost-Benefit Analysis of Environmental Protection*, 150 U. PA. L. REV. 1553, 1578–81 (2002); Lisa Heinzerling, *Regulatory Costs of Mythic Proportions*, 107 YALE L.J. 1981, 2060–64 (1998); Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1616.

77 In addition to determining what factors a statistic should reflect (e.g., all risks or some set of well known risks), scientists must determine (1) the level at which information will be aggregated and (2) the scale used to quantify the data. The first question raises an interpretive problem, namely, that "statistical uniformity [may be] mostly superficial, attained only when one considers a great mass and smears together the variety of the phenomena." PORTER, *supra* note 29, at 183–84. A globally aver-

ence are biased because they implicitly place the scientific burden of proof on those seeking to demonstrate risks to the environment or human health. Although these objections are distinct, statistical inference is dependent on quantification as a precondition for its application. Methods of quantification (e.g., scaling or variable type) and any underlying scientific theories frame all statistical analyses, making statistical inference dependent on these methodological and theoretical judgments. This Part focuses on the uncertainties inherent in, and approaches to, developing quantitative measures for statistical analyses.

Environmental policy is riven by disputes between environmentalists and regulatory opponents over the use of quantitative methods—although, neither is consistent in their critique.[78] For example, environmental risks are typically formulated by critics of regulation as objective properties that can be readily aggregated for an entire population, just as economists aggregate individual utilities. In opposition, environmentalists argue that such overly reductive approaches to risk distort reality by disregarding localized conditions, such as isolated areas of high risk that disproportionately affect certain communities, and by excluding important values commonly overlooked in risk models.[79] Yet, many simplifying assumptions also are made in ar-

---

aged temperature, for example, actually obscures regional variability and limits the range of causal relations that can be resolved. Statistical uniformity in such cases is an artifact of the system's size, and is thus not accurately representative of the quantity being analyzed. *Id.* at 184–85. Statistical scaling is determined either by the observation technique or by the nature of the hypothesis, such that some statistics fall into strict categories while others extend along a continuum. For example, survival or death from lead exposure is categorical (an individual is either dead or alive, not somewhere in between) whereas cognitive losses from lead exposure occur along a continuum ranging from extreme to nominal impairment. In this simple example, a scientist would therefore use different statistical scales depending on whether she was concerned about mortality or cognitive harm to children.

78 Justice Stephen Breyer's book, *Breaking the Vicious Circle*, illustrates this point perfectly. In chapter one, he bases his analysis on a great deal of cost-benefit data, which he largely takes as uncontested; in chapter two, he describes in detail the problems and limitations of existing risk assessment methods, which he describes as highly uncertain and subject to numerous simplifying assumptions. BREYER, *supra* note 20, at 3–51. However, as professor Lisa Heinzerling has argued persuasively, the cost-benefit methods Breyer relies on in his first chapter are equally susceptible to the sorts of uncertainties and simplifying assumptions Breyer takes pains to elucidate in the second chapter of his book. Heinzerling, *supra* note 76, at 2064–69.

79 *See supra* note 76. Risk models, for example, assume that risk varies continuously and linearly, such that a marginal increase in a low-level risk is treated as equivalent to a marginal increase in a high-level risk. For some people, risks above a certain level will be intolerable and they would not weigh marginal increases in risk above this level equally with risks below it. This is analogous to, though the inverse of,

eas of environmental science that environmentalists use to justify regulatory programs.[80] Climate change models, for example, ignore localized conditions to enable statistical data to be interpretable and predictions to be made.[81] Each side of the debate thus has its favored areas of science where quantification, despite numerous simplifying assumptions, remains relatively uncontested, as well as its favorite examples of "pseudo-science" permeated by "trans-scientific" judgments (i.e., technical questions that cannot be fully resolved by science).[82] Predictably, the line between presumptively legitimate methods and so-called junk science is drawn such that the interests of the particular advocate are supported by good science.

Science and trans-science cannot be distinguished based on the presence or absence of qualitative judgments—at most, this is a question of degree rather than kind. Indeed, if science were defined by such analytical purity, virtually nothing could be characterized as scientific. As discussed below, philosophers such as Sir Karl Popper exposed the fallacy that science can be reduced to either logically inductive or deductive methods.[83] Yet, conservative proponents of the natural sciences frequently describe them in a purely positivist mode

---

economists' claims that the marginal value of money diminishes with a person's wealth.

80 Environmental assessments use broad based indices, such as global carrying capacity and growth limits, that are subject to numerous judgments and uncertainties. Moreover, some of the most celebrated environmental books have used quantitative estimates to dramatize the impacts of population growth and the limits of environmental sustainability. *See, e.g.*, PAUL R. ERHLICH, THE POPULATION BOMB (rev. ed. 1978) (documenting the tensions between population growth and environmental unsustainability); DONELLA H. MEADOWS ET AL., BEYOND THE LIMITS (1992) (updating the *Limits to Growth* analysis and making an effort to describe a sustainable future); DONELLA H. MEADOWS ET AL., THE LIMITS TO GROWTH (1972) (using a systems based approach to model human impacts on the global environment and to assess the limits of human and economic growth); *see also* WORLD WILDLIFE FUND, LIVING PLANET REPORT (2002), *available at* http://www.wwf.org.uk/filelibrary/pdf/livingplanet2002.pdf (giving a contemporary analysis of global impacts using a variety of quantitative indices).

81 *See supra* note 77.

82 *See* KENNETH R. FOSTER & PETER W. HUBER, JUDGING SCIENCE 55–58 (1997) ("[Trans-science] concerns questions that are scientific in Popper's sense but are not resolvable in practice."); Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1619 nn.21–22. The term "trans-science" was first coined by the physicist Alvin Weinberg to mean "questions which can be asked of science and yet *which cannot be answered by science.*" Alvin M. Weinberg, *Science and Trans-Science*, 10 MINERVA 209, 209 (1972).

83 *See infra* Part II.B; RICHARD H. GASKINS, BURDENS OF PROOF IN MODERN DISCOURSE 152–53 (1992); Naomi Oreskes et al., *Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences*, 263 SCIENCE 641, 642 (1994); Weinberg, *supra* note 82, at 209.

that portrays science as purely a matter of logic.[84] Similarly, it is this mechanical picture of science that many environmentalists critical of quantitative methods have in mind when they challenge the purported bias of scientific findings.[85] Both perspectives are misleading. Selectively rejecting scientific methods based on a positivist caricature of science is a strawman tactic used by regulatory critics to discredit certain types of science, and by environmentalists to bind scientific experts to a rigid model of science. All areas of science necessarily draw on a variety of quantitative and qualitative methods, making scientific positivism little more than a malleable fiction that is wielded effectively in an offensive mode but has little constructive value.

At the other extreme, the relativists' claim that science is suffused by subjective judgments that negate the separation of scientific facts and values is equally problematic.[86] This perspective ignores important low-level methodological strategies, such as using independent and diverse experimental methods to test scientific hypotheses, that prevent or limit the intrusion of values into science. A bipolar debate over the proper role of science in environmental policy therefore misses a great deal of what gives science its normative authority. Too few people in the legal community adequately appreciate that much useful science falls between the positivist and relativist poles. The sections that follow describe typical approaches to scientific uncertainty and quantification within environmental law, present the formal philosophical arguments on which these approaches are based, and identify a middle ground for understanding how quantitative methods evolve and succeed in the face of broader theoretical uncertainties. In this light, scientific authority derives from its prospective power to

84 *See, e.g.*, SHEILA JASANOFF, THE FIFTH BRANCH: SCIENCE ADVISORS AS POLICYMAKERS 1 (1990) (describing how commentators perceive science advisors as "critical intelligence" in a regulatory system that is otherwise too vulnerable to politics); Graham, *supra* note 4, at 41–42 (describing sound science as "objective, weight-of-the-evidence evaluation that is peer reviewed"); Dorothy Nelkin, *The Political Impact of Technical Expertise*, 5 SOC. STUD. SCI. 35, 36 (1975) (describing the idealized role of technical expertise in policymaking). This "scientist" perspective dates back to the Theodore Roosevelt conservation era, which was based on using trained impartial experts, drawing on objective scientific facts, to make resource management decisions. SAMUEL P. HAYS, CONSERVATION AND THE GOSPEL OF EFFICIENCY: THE PROGRESSIVE CONSERVATION MOVEMENT 1890–1920, at 2–3 (1959).

85 *See, e.g.*, PROTECTING PUBLIC HEALTH & THE ENVIRONMENT 71 (Carolyn Raffensperger & Joel Tickner eds., 1999) (articulating a simple positivist view of science).

86 As one critic has put it, the relativist "move from the alleged failure of [a sampling of] methodological rules to the presumption that all methodologies are hopeless is to engage in just the sort of naive inductivism about which [they are] otherwise so abusive." LAUDAN, *supra* note 30, at 104.

generate accepted truths, as opposed to some claim to analytical determinacy, and treating science in purely relativist terms (i.e., it is all political) threatens the potential growth and success of environmental science.

A. The Debate Within the Legal Community over Quantification

The rhetorical power and limitations inherent in statistical methods of quantification clearly warrant the legal community's attention.[87] The basic problem, according to environmentalists, is that the process of identifying and naming categories can in fact never be wholly neutral (e.g., defining risk solely in terms of human cancer deaths), as all facts embody certain substantive beliefs and values.[88] The mixing of facts and values negates the neutrality of science because simplifying assumptions employed to construct quantitative measures obscure certain "perspectives and possibilities," such as hotly contested environmental values like protecting endangered species or pristine wilderness.[89] Further, this blending of science and policy inexorably leads scientists to inject their own ideological predilections into their analysis (e.g., industry scientists make self-serving assump-

87 These distortions are perhaps more potent than one might initially assume when applied in a legal or policy setting: The use of statistical methods, for example in risk assessment, is necessarily based on a host of simplifying assumptions that amount to a highly simplified and ultimately inaccurate version of reality. Nevertheless, because of the perceived solidity and utility of a numerical formulation, we conflate this caricature with conditions in the real world, disregarding the acknowledged limitations of the theory employed. *See* MARK KELMAN, A GUIDE TO CRITICAL LEGAL STUDIES 291 (1987). Thus, once a policymaker has risk estimates before her, they are treated as gospel in the absence of competing estimates or a sophisticated understanding of the underlying methods. In so doing, policymakers transform a highly simplified theory, or hypothesis, about the world into a method of obtaining information that is descriptive of it; theory seamlessly becomes an established method for observing facts.

88 *See* Tribe, *supra* note 76, at 75–77.

89 *See id.*; Ackerman & Heinzerling, *supra* note 76, at 1578–81; Heinzerling, *supra* note 76, at 2060–64; Tribe, *Plastic Trees*, *supra* note 75, at 1319. A common example is the so-called "dwarfing of soft variables," which occurs when unquantified factors (soft variables) are discounted, or ignored, relative to quantified factors in an analysis. *See* Heinzerling, *supra* note 76, at 2060–61; Tribe, *Trial by Mathematics*, *supra* note 75, at 1361–65. If, for instance, an agency quantifies certain risks associated with a proposal, such as human health risks from cancer, but neglects others, such as ecological impacts, the quantified risks will be given disproportionate weight in setting policy. For some of these commentators, the cryptic nature of quantitative methods aggravates these deficiencies because "numbers disguise the numbers' true meaning and inhibit useful and informed discussion about the matters in question." Heinzerling, *supra* note 76, at 2042.

tions that chemicals are harmless at very low exposure levels).[90] Similar to the legal debates over formalist approaches to judicial review, the objectivity of science is challenged because of such individual biases or decisionmakers' interest in cloaking questions of policy in technical jargon.[91]

The favored example of policy masquerading as science in environmental law is chemical risk assessment. In her article, *The Science Charade in Toxic Risk Regulation*, Professor Wendy Wagner identifies many so-called trans-scientific judgments made when interpreting the results of animal studies for purposes of determining a chemical's toxicity:[92]

> Extrapolating [the high-dose] results to potential effects of low levels of the substance on humans then presents the next two trans-scientific junctures, which are often collapsed into one. First . . . an extrapolatory model must be selected that will predict low-dose effects on animals based solely on high-dose data. Although there are several scientifically plausible extrapolatory models . . . the choice of one model over another cannot be resolved by science and thus must be determined by policy factors. This policy choice will have significant implications for the level ultimately chosen as adequate to protect public health. Second . . . since the similarities between animals and humans with regard to their sensitivity to carcinogens

---

90 *See* Tribe, *Plastic Trees*, *supra* note 75, at 1332. Many industry scientists, for example, subscribe to the view that toxic chemicals often have a threshold below which they have no effect. Most environmentalists reject this theory as unproven and likely unprovable given existing methods. Both sides of this debate arguably make judgments about the underlying science that are informed by their individual prejudices and interests.

91 *See generally* BREYER, *supra* note 20, at 47–49 (explaining that scientists from government regulatory agencies can frame uncertain scientific evidence to fit political agendas); Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1628–31 (describing how government agency scientists fill "trans-scientific gaps" in the standard setting process by making policy choices and by failing to reveal those choices). The same kinds of allegations have been made regarding the role of judges in reviewing statutes or the Constitution:

> It's because everybody down deep knows [that judges can get away with inserting their own value judgments into opinions] that few come right out and argue for the judge's own values as a source of constitutional judgment. Instead the search purports to be objective and value-neutral; the reference is to something "out there" waiting to be discovered, whether it be natural law or some supposed value consensus of historical America, today's America, or the America that is yet to be.

ELY, *supra* note 14, at 48.

92 Wagner defines trans-science broadly to include "significant splits in the scientific community that are identified by scientists as major controversies over 'scientific judgment.'" Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1620 n.22.

> are largely unknown and incapable of being studied directly, a policy choice must again be made.[93]

Justice Stephen Breyer makes many similar observations, although from a different perspective, in his book, *Breaking the Vicious Circle*:

> The more frequently used animal studies are often more uncertain [than human studies]. The investigator applies a high dose of a supposed carcinogen to the animals; if they develop a higher than average number of tumors, the analyst tries to extrapolate backward to low doses in humans. What assumptions shall be made in doing so? What extrapolation model should be used? Risk analysts tend to use, for both animal and epidemiological studies, a linear model, which extrapolates backward on a straight line. . . . Critics argue that to use such mathematical models is like saying "If ten thousand men will drown in ten thousand feet of water, then one man will drown in one foot of water," or "If dropping ten bottles off a ten-foot wall breaks all ten, then dropping ten bottles off a one-foot wall will break one."[94]

Both authors raise compelling points about the intermixing of science and policy.[95] Yet, they come to different conclusions about the problems with federal policymaking. Breyer considers the "disciplinary canon" to be often far too weighted towards "erring on the safe side,"[96] whereas Wagner claims that trends in chemical regulation reveal that the system is biased against less studied chemicals and often neglects chemicals posing greater potential risks.[97]

Despite their mutual concern about administrative efficiency and scientific accuracy, Breyer and Wagner propose conflicting solutions for ensuring that regulatory decisionmaking is effective when the available science is indeterminate.[98] Consistent with his faith in scientific expertise, Breyer recommends establishing a distinct civil service career path in health and environmental policy and calls for the "crea-

---

93 *Id.* at 1626.

94 BREYER, *supra* note 20, at 44 (citations omitted). The critics Breyer refers to are extrapolating linearly down to zero.

95 I question whether the analogies Breyer offers at the end of the excerpt are appropriate. Evidence exists that certain chemicals, even at very low concentrations, are toxic, whereas the two examples Breyer provides are designed to be absurd because we apparently have no reason to fear any harm at the lower magnitudes.

96 BREYER, *supra* note 20, at 43.

97 Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1682–84 (objecting to the EPA's focus on human epidemiological data to the exclusion of other studies based on animal data or in vitro studies).

98 BREYER, *supra* note 20, at 48–51, 55–56; Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1673–74, 1677–88.

tion of a small, centralized administrative group, charged with a rationalizing mission," which would have prestige, independence, and substantial authority.[99] Wagner is more focused on political accountability and process, which leads her to propose procedural amendments to the Administrative Procedure Act.[100] Wagner's reforms have two elements: First, the amendments would "expressly require the agencies to separate science from policy and task the courts with the responsibility for ensuring that the agency does so in an accurate and accessible way;" second, if an agency passes this first test, a reviewing "court should accord the content of the policy and science decisions great deference."[101] In short, Wagner relies on a different group of experts—judges—to provide a procedural check on scientific experts engaged in environmental policymaking.

Both proposals have serious drawbacks. Much of the case for Breyer's proposal turns on the political isolation and objectivity of the elite administrative group he recommends. Even setting aside concerns about political accountability,[102] insulating such a group from politics seems wildly unlikely. One need only consider the political storms surrounding nominations of federal judges, who may be politically insulated once appointed but must first traverse a political minefield.[103] Given the broad authority Breyer intends to delegate to this elite group, it is virtually inconceivable that a similar vetting process would not arise. Federal scientific advisory boards offer another point of comparison.[104] Despite their (generally) low profile, highly technical orientation, and limited tenures, scientific advisory boards have not been immune to politicization—especially where environmental issues are concerned. In 1983, for example, the Reagan Administration was known to have had a "hit list" for "green" scientists

99 BREYER, *supra* note 20, at 59–61.

100 5 U.S.C. §§ 551–559 (2000).

101 Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1711–12 (footnote omitted); *see also* McGarity, *supra* note 5, at 746–47 (describing a proposal that shares many similarities with Wagner's).

102 Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1674–77.

103 *See* Stephen Murdoch, *The Politics of Judicial Confirmation*, WASH. LAW., Sept. 2002, at 22.

104 *See, e.g.*, JASANOFF, *supra* note 84, at 1–2 (describing scientific advisory committees as "offer[ing] a flexible, low cost, means for government officals to consult with knowledgeable and up-to-date practitioners in relevant scientific and technical fields"); BRUCE L.R. SMITH, THE ADVISERS: SCIENTISTS IN THE POLICY PROCESS 1 (1992) (noting the existence of about 1000 different committees and dividing them into four categories: peer review panels, program advisory committees, ad hoc factfinding or investigating committees, and standing committees).

on the EPA's Scientific Advisory Board.[105] More recently, the current Bush Administration has been accused by prominent members of the scientific community of creating "an epidemic of politics" in its appointments to scientific advisory boards.[106] These realities make political insulation chimerical where powerful interests are affected, and they seriously undermine Breyer's approach.

Wagner all but concedes the major flaw in her proposal toward the end of her article: "Although the capability of the judiciary to review science-policy delineations is the limiting factor to success of the reform, it is conceivable that this obstacle can be overcome."[107] Wagner acknowledges this difficulty earlier, concluding that "distinguishing between questions resolvable by science and those that must remain trans-scientific requires familiarity with the current capabilities and limitations of scientific experimentation."[108] I am far less sanguine than Wagner about relying on judicial review. First, scientific literacy is only a necessary condition; it provides no assurance that the courts' procedural distinctions, which also are vulnerable to value judgments, will not be shot through with political considerations. Second, an analogue of her proposal exists in the National Environmental Policy Act (NEPA)[109] and suggests further obstacles. Like Wagner's amendments, NEPA is designed to promote better informed decisions by federal agencies and to enhance public participation.[110] Moreover, while NEPA's regulations do not precisely contemplate Wagner's distinction, their mandate that the incompleteness or un-

105 JASANOFF, *supra* note 84, at 89.

106 The Bush Administration has been accused of "stacking" several high-profile committees, most notably the CDC's Advisory Committee on Childhood Lead Poisoning and the FDA's Reproductive Health Drugs Advisory Committee. *See* Ceci Connolly, *Hill Group Faults HHS for Ideology*, WASH. POST, Oct. 22, 2002, at A25; Maureen Dowd, Editorial, *Tribulation Worketh Patience*, N.Y. TIMES, Oct. 9, 2002, at A27; Dan Ferber, *Critics See a Tilt in a CDC Science Panel*, 297 SCIENCE 1456 (2002); Ellen Goodman, Editorial, *Religious Profiling?*, WASH. POST, Oct. 19, 2002, at A23; Donald Kennedy, *An Epidemic of Politics*, 299 SCIENCE 625 (2003); Sheryl Gay Stolberg, *Bush's Science Advisers Drawing Criticism*, N.Y. TIMES, Oct. 10, 2002, at A27; Rick Weiss, *HHS Seeks Science Advice to Match Bush Views*, WASH. POST, Sept. 17, 2002, at A1. This is also a political issue, with several Democrats now objecting to the Administration's practices. *See* David Malakoff, *Democrats Accuse Bush of Letting Politics Distort Science*, 301 SCIENCE 901 (2003).

107 Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1719.

108 *Id.* at 1627.

109 42 U.S.C. §§ 4321–4370(f) (2000).

110 *See* National Environmental Policy Act of 1969, 42 U.S.C. §§ 4321, 4332; *see also* 40 C.F.R. §§ 1500.1, 1501.2 (2002). Implementing regulations for NEPA stress the importance of effectively communicating with the public and addressing directly questions regarding "incomplete or unavailable information." *Id.* § 1502.22.

availability of information be disclosed is very similar.[111] Experience with NEPA is not promising with regard to either streamlining the regulatory process or improving transparency: NEPA is the preeminent statute for delaying agency action, and the environmental impact statements that NEPA mandates are notoriously arcane.[112] The same complex science, politics, and bureaucratic tendencies are likely to overwhelm Wagner's reforms as well.

Breyer and Wagner both make sweeping generalizations from the uncertainties of risk assessment methods without considering science more broadly.[113] This narrow scope weakens their arguments and is compounded by their apparently static view of environmental science.[114] Fixating on trans-science also causes them to ignore that science is itself a process and to blithely reject the potential for science to progress.[115] Breyer and Wagner both fall into this trap by failing to address the tradeoffs that exist between maintaining scientific integrity and safeguarding democratic principles, and this oversight undermines the merits of both their arguments in the process. Part II.B. discusses the origins of scientific uncertainty and theories of scientific progress in order to clarify the problems Breyer and Wagner raise and to explain how scientific judgment is used in environmental science.

---

111 Council on Environmental Quality Environmental Impact Statement, 40 C.F.R. § 1502.22. This section requires a statement to be included in an environmental impact statement regarding all incomplete or unavailable "information relevant to reasonably foreseeable significant impacts." *Id.*

112 WILLIAM H. RODGERS, ENVIRONMENTAL LAW § 9.1, at 817–18 (2d ed. 1994) (describing an "avalanche of litigation" and injunctions); Bradley C. Karkkainen, *Toward A Smarter NEPA: Monitoring and Managing Governments' Environmental Performance*, 102 COLUM. L. REV. 903, 915–16 (2002) (noting the highly technical nature of environmental impact statements); William C. Sullivan et al., *Assessing the Impact of Environmental Impact Statements on Citizens*, 16 ENVTL. IMPACT ASSESSMENT REV. 171, 174–75, 177–79 (1996) (observing that citizens generally do not comprehend the information presented in a typical environmental impact statement).

113 Breyer claims the example of "cancer-causing substances" has "illustrative power," but nowhere seeks to substantiate this claim. BREYER, *supra* note 20, at 3.

114 In fact, trans-science is frequently defined as immune to scientific investigation. *See* McGarity, *supra* note 5, at 733–34 ("[B]y definition, scientific experimentation is incapable of resolving trans-scientific issues."); *see also* FOSTER & HUBER, *supra* note 82, at 55 ("[Trans-science] concerns questions that are scientific in Popper's sense but are not resolvable in practice.").

115 The concept of trans-science is further complicated by environmentalists' claims that scientific facts and values cannot be separated. Yet, if science, as opposed to trans-science, is to have any content, it seems to demand that science be capable of establishing facts that are independent of value judgments. Reliance on the science/trans-science distinction therefore seems to require qualification of the fact-value argument environmentalists often make.

Part II.C. presents a methodologically grounded view of science that offers a different account of how quantitative methods succeed.

B. The Problem of Underdetermination in Science

Breyer and Wagner focus on a single instance of a much broader question—how to obtain reliable scientific information and theories. The aim of this discussion is to demonstrate that no single scientific theory (or statistical technique) is adequate to the task; all are distorting insofar as they utilize an imperfect framework for evaluating scientific problems. This incompleteness, however, should not be construed to imply that science is hopelessly indeterminate; to the contrary, methodological strategies exist that overcome these shortcomings. The discussion that follows briefly examines the treatment of scientific uncertainty by Karl Popper and Thomas Kuhn. As already noted in the Introduction, the discussion will draw on arguments made regarding judicial review to ground the scientific theory in familiar legal territory.

Classical positivists or empiricists, such as Francis Bacon, portrayed science as consisting purely of gathering positive facts from which scientific theories are mechanically inferred.[116] They believed that knowledge is obtained passively, through nature imprinting itself on inert minds, and that scientific uncertainty is simply an absence of facts.[117] This passive model was challenged by the interpretive Kantian school, which asserted that all experience and understanding are actively mediated through an innate conceptual framework. In the last century, the work of Karl Popper incorporated the Kantian activist model.[118] Popper believed that reliable scientific inference is best achieved through a combination of imaginative hypothesis testing and a rigorous war of attrition, or falsification, based on hard-headed critical analysis.[119]

Popper's work has been immensely influential in science and law.[120] A central tenet of Popper's work is that scientific inference is

116 *See* ALAN MUSGRAVE, COMMON SENSE, SCIENCE AND SCEPTICISM 48–54 (1993); DOROTHY ROSS, THE ORIGINS OF AMERICAN SOCIAL SCIENCE 17–18 (1991).

117 Imre Lakatos, *Falsification and the Methodology of Scientific Research Programmes, in* CRITICISM AND THE GROWTH OF KNOWLEDGE 91, 104 (Imre Lakatos & Alan Musgrave eds., 1970).

118 KARL R. POPPER, CONJECTURES AND REFUTATIONS 93–96, 190–91, 383–84 (1963).

119 *Id.* at 26–30.

120 Popper, in particular, was favorably cited by the Supreme Court in the seminal 1993 *Daubert* opinion. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 593 (1993). Popper and Carl Hempel were the only two philosophers cited by the

never purely instrumental—facts and theories combine to alter an individual's conceptual framework, as opposed to merely augmenting their breadth of knowledge.[121] Popper's position is analogous to the commonplace view that a judge's perspective is necessarily influenced by her experience (facts), which, in turn, will affect the principles (theories) she considers when interpreting a given constitutional provision (fact). Moreover, because theories are inextricably linked to facts, theories may operate in an observational mode to generate facts for hypothesis testing (e.g., theories about how compounds absorb light are used in experiments designed to test theories about atmospheric pollution) or may be the subject of hypothesis testing themselves.

Popper began with two central arguments: (1) "all theories are not only equally unprovable but also equally improbable,"[122] and (2) "no conclusive *disproof* of a theory can ever be produced."[123] This latter argument is critically important because it necessitates the rejection of naive logical positivism, which is premised on the possibility of falsifying, as opposed to verifying, scientific theories. According to Popper, those who wait for conclusive disproof before eliminating a theory will have to wait forever, and "will never benefit from experience."[124] Naive logical positivism therefore fails for the same reason that classical empiricism does, namely, all factual observations are

---

Supreme Court. *Id.* According to the Court, "[s]cientific methodology today is based on generating hypotheses and testing them to see if they can be falsified; indeed, this methodology is what distinguishes science from other fields of human inquiry." *Id.* (quoting Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation*, 86 NW. U. L. REV. 643, 645 (1992)). The Supreme Court then went on to quote Popper directly: "[T]he criterion of the scientific status of a theory is its falsifiability, or refutability, or testability." *Id.* (quoting POPPER, *supra* note 118, at 37).

121 POPPER, *supra* note 118, at 108–14; KARL R. POPPER, THE LOGIC OF SCIENTIFIC DISCOVERY 37–38, 59 n.1 (1959).

122 Lakatos, *supra* note 117, at 95; *see also* POPPER, *supra* note 121, at 27–30, 254–62. The argument goes as follows: First, if theories cannot be proven by facts, but only falsified, then all theories are fallible. Second, if all facts are constructed from fallible theories, then all facts themselves are also fallible. Third, if all facts used to negate theories are fallible, then the experimental negation of any theory is itself also fallible. This series of syllogisms may be summed up as follows: "Scientific theories are not only equally unprovable, and equally improbable, they are also equally undisprovable." Lakatos, *supra* note 117, at 103. Moreover, theories become more resistant to empirical challenges as they mature. *Id.* at 101–02.

123 POPPER, *supra* note 121, at 50, 80–81 (emphasis added); *see also* POPPER, *supra* note 118, at 278–79 (noting that "actual tests are never conclusive and always tentative").

124 POPPER, *supra* note 121, at 50.

"theory impregnated" and thus do not exist outside a theoretical framework.[125] Scientists, for example, cannot conduct an unmediated measurement of a chemical's toxicity because theories about exposure levels and pathways, animal models, and experimental controls are required to make sense of the data and to direct how an experiment is conducted. Accordingly, just as judges often must appeal to substantive principles to interpret a constitutional provision, so too must scientists use substantive theories to resolve the meaning of their data; both are theory-impregnated in this sense.

Popper's second argument can be restated as follows: no single theory "forbid[s] any observable state of affairs."[126] An example will help to illustrate this point. Assume that an astronomer has observed a planet revolving around a distant star and is able to map the planet's orbit based on her observations.[127] Assume further that the observed orbit deviates significantly from the orbit one would expect to find by performing classical (i.e., Newtonian) calculations. Does this deviation lead the astronomer to reject Newton's theory? No. The astronomer will instead construct a hypothesis (e.g., an unknown second planet is altering the first planet's orbit) to explain the orbital deviation and then conduct further experiments to test the validity of this secondary hypothesis. Popper argued that a chain of such observations followed by "correcting" or "auxiliary" hypotheses could be constructed to explain away any deviant result, and that this process could go on indefinitely.[128] Any collection of facts is therefore going to be theoretically underdetermined because it can be accounted for by numerous competing theories.

Underdetermination creates a fundamental dilemma: If theories can neither be proved nor disproved conclusively, on what grounds may scientists adopt any given scientific theory? Popper attempted to resolve this problem by recourse to a "survival-of-the-fittest" model for science, under which hypothesis selection occurs through a process of falsifying theories against each other rather than against isolated facts.[129] For each test, one theory is employed as the (tentatively) accepted "background knowledge" that supplies the facts employed to

125 POPPER, *supra* note 118, at 188–89. Some theorists claim that logical positivism "is at this stage dead." Interview by Werner Callebaut with Philip Kitcher, Professor of Philosophy, University of California at San Diego, in San Diego, Cal. (June 1, 1990), *in* TAKING THE NATURALISTIC TURN, OR HOW REAL PHILOSOPHY OF SCIENCE IS DONE 128 (Werner Callebaut ed., 1993).

126 Lakatos, *supra* note 117, at 100–01, 116.

127 *Id.* at 100–01.

128 *Id.* at 100–01; POPPER, *supra* note 121, at 82–83.

129 POPPER, *supra* note 118, at 52; POPPER, *supra* note 121, at 42, 49–50.

test a second theory.[130] Popper further required that all such background knowledge be "well corroborated," i.e., subject to extensive testing itself, and that auxiliary hypotheses be investigated exhaustively.[131] A theory is therefore "scientific" insofar as it predicts or explains phenomena that may be tested against facts supported by another well corroborated theory.[132] Scientific inference according to Popper is relative to the existing information, theories, and auxiliary hypotheses; it is also ultimately a matter of convention supported by judgment, not of rigorous deductive logic alone. According to Popper, the fallibility of science cannot be overcome; it can only be counteracted by rigorous testing and critical analysis.

Thomas Kuhn built on Popper's work, but took it in a very different direction.[133] Where Popper claims discrete experiments can support scientific inference, Kuhn argues that experiments are unavoidably ambiguous because no theory is ever consistent with all of the available experimental evidence.[134] Kuhn also inverts Popper's priorities by arguing that the strength of science is its theoretical stability, not Popper's unrelentingly critical mode of inquiry. Kuhn believes scientific progress is made by scientists tenaciously investigating an accepted theory, both because it generates cumulative knowledge and because it propels theoretical innovation.[135]

Kuhn posits that science evolves through revolutionary shifts in which an accepted theory is replaced by one that is radically different.[136] He then develops this position by reconceptualizing important scientific developments through two central constructs: scientific paradigms and normal science.[137] Under this framework, the progress

130 POPPER, *supra* note 118, at 112, 238–40; POPPER, *supra* note 121, at 75–77; Lakatos, *supra* note 117, at 107–09.

131 POPPER, *supra* note 121, at 266–69; Lakatos, *supra* note 117, at 106–08.

132 POPPER, *supra* note 121, at 86–87, 104–05, 109–11.

133 *See* Thomas S. Kuhn, *Logic of Discovery or Psychology Research?*, *in* CRITICISM AND THE GROWTH OF KNOWLEDGE, *supra* note 117, at 1, 1–3 (acknowledging the close similarity of his and Popper's views on many issues).

134 *Id.* at 15–16. Drawing on historical examples, Kuhn shows that even for the most fully corroborated paradigms, significant, unexplained contrary evidence exists. *See* THOMAS S. KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS 79, 81–83 (3d ed. 1996); *see also* PAUL FEYERABEND, AGAINST METHOD 50–53 (3d ed. 1993) (stating that even the most well corroborated theories may sometimes be inconsistent with the evidence).

135 *See* KUHN, *supra* note 134, at 162–65. Kuhn also uses the natural selection metaphor. *See id.* at 146.

136 *See id.* at 92–94.

137 A paradigm has two basic elements: First, it represents the constellation of beliefs, values, and techniques shared by a scientific community; second, it is sustained by "concrete puzzle-solving" of problems discovered within a paradigm that arise as a

from one paradigm to the next does not proceed through logical extensions because paradigms are incommensurable with each other, meaning they are only imperfectly comparable or translatable.[138] Kuhn infers from this observation that theories cannot be rejected solely based on inferences from experimental results—in his words, the "decision to reject one paradigm is always simultaneously the decision to accept another."[139] The discontinuous nature of scientific developments leads Kuhn to conclude that adopting a new paradigm is an almost quasi-religious "conversion experience" that requires an act of "faith."[140] Kuhn does not, however, maintain that paradigm shifts are utterly irrational. He believes that such revolutionary thinking is justified by broad scientific values, such as simplicity, breadth, and consistency, and that these values are just as critical to scientific progress as rigorous methods.[141] Kuhn thus is not a scientific relativist; he merely rejects the belief that science progresses through, and is defined by, a shared "algorithmic decision procedure."[142]

For Kuhn, broad principles are integral to science, and it is the acceptance of their inherent indeterminacy that allows him to grasp what he believes is the great strength of science—normal science. The theoretical stability of normal science enables cumulative data collection and theory development to occur. Moreover, normal science either further corroborates a paradigm or (ultimately) precipitates a theoretical crisis, and paradigm shift, by exposing its anomalies.[143] Thus, although Kuhn rejects the notion that scientific progress can be prescribed algorithmically, through normal science he discovers the conditions necessary for science to progress.[144] Kuhn

---

theory is elaborated. *Id.* at 175, 182–84. Normal science is limited to concrete puzzle solving that is "firmly based upon one or more past scientific achievements," which are treated as fundamental and uncontested. *Id.* at 10.

138 *See id.* at 101–03, 122–23, 148–50; Thomas S. Kuhn, *Possible Worlds in History of Science, in* POSSIBLE WORLDS IN HUMANITIES, ARTS AND SCIENCES 9–11 (Sture Allén ed., 1989); Thomas S. Kuhn, *Reflections on my Critics, in* CRITICISM AND THE GROWTH OF KNOWLEDGE, *supra* note 117, at 231, 266–67. Kuhn has been convincingly criticized on this point, for there are many examples in which competing theories and ideas coexist (e.g., theories of matter). *See, e.g.*, John Watkins, *Against 'Normal Science,' in* CRITICISM AND THE GROWTH OF KNOWLEDGE, *supra* note 117, at 25, 34–36.

139 KUHN, *supra* note 134, at 77, 79.

140 *Id.* at 151, 156–58.

141 KUHN, *supra* note 11, at 330–34.

142 *See id.* at 326–27.

143 The virtue of normal science is paradoxical, for it is dogged commitment to a paradigm and to its exploration that exposes a paradigm's weaknesses and causes its failure. *Id.* at 96, 98, 165–70.

144 Despite their different frameworks, important parallels exist between Popper's and Kuhn's perspectives. First, both theories rely on some form of tentatively ac-

therefore distinguishes between scientific certainty, which he views as unattainable, and progress, which he defines through the combined workings of normal science and shifting scientific paradigms.

Significant disparities exist between how Popper and Kuhn portray science, on the one hand, and how science is typically understood within the legal community, on the other. The views of Breyer and Wagner illustrate these differences. They believe that chemical risk assessment and, in Breyer's case, regulatory science more generally, lie somewhere on the margins of rigorous scientific practices. Breyer is clearest on this point:

> Predicting risk is a scientifically related enterprise, but it does not involve scientists doing what they do best, namely developing theories about how *x* responds to *y*, others things being equal . . . . Scientists are happier looking for large differences in small populations over short periods of time than looking for small differences in large populations over long periods of time. To do the latter, they must make many simplifying assumptions that are often questionable.[145]

This view manifests a limited understanding of what scientists do, particularly with respect to the analytical and statistical methods they employ. First, as Jeffreys's work demonstrates, not all science involves running large numbers of carefully controlled experiments to test a theory.[146] Many areas of science, such as ecological fieldwork and atmospheric science, rely on discrete observations that cannot be carefully controlled or indefinitely repeated. Scientists have also constructed theoretical models and methods that require little empirical testing for them to be accepted.[147] Second, as Popper showed, scientists routinely make simplifying assumptions in the face of uncertainty with the objective of iteratively testing the validity of these assumptions and the confidence scientists should have in them. The so-

---

cepted theory: Kuhn relies on normal science; Popper utilizes background knowledge. Second, each theory posits a central mechanism that propels scientific progress: Kuhn's mechanism is the breakdown of normal science; Popper's is a strict critical mode of inquiry. Third, each approach has a model for scientific inference and theory choice. On this point, however, a fundamental conceptual difference emerges. Popper's model for scientific inference and theory choice is his convention for critical analysis. Kuhn, in contrast, treats the two separately: scientific theory choice is a matter of psychological conversion; scientific inference occurs within the bounds of normal science. *See* KUHN, *supra* note 134, at 144–46. As a result, Popper inherently places a higher premium on scientific judgment, while Kuhn views its role as being generally limited by the theoretical boundaries of normal science.

145 BREYER, *supra* note 20, at 42–43.

146 *See supra* Part I.

147 *See, e.g.*, LAUDAN, *supra* note 30, at 172–73.

called trans-scientific junctures that Wagner and Breyer identify are therefore central to all areas of science and not unique to chemical risk assessment.

Popper and Kuhn show that, at some level, all scientific theories are uncertain because they cannot be definitively proved or disproved. Science thus does not consist of mechanical true-false testing, but must turn on the degree of confidence a hypothesis warrants based on whether it has withstood (or failed) rigorous testing.[148] Further, scientific testing itself entails auxiliary hypotheses and background knowledge, both of which will vary in the degree to which they are corroborated. Risk assessors' assumptions about extrapolation models are, for instance, directly analogous to the auxiliary hypotheses climate scientists integrate into their atmospheric models.[149] In the former, risk assessors typically assume a linear relationship exists between chemical dose and risk, even at very low exposure levels; in the latter, climate scientists must make myriad assumptions to derive a relationship between greenhouse gas levels and mean global surface temperature. These judgments are unremarkable because experimental work is scientific insofar as models and hypotheses are rigorously tested and evaluated—science is "a process rather than the product of inquiry."[150]

The categorical division between science and trans-science, implicit in Breyer's work and explicit in Wagner's, conflicts with this view of science as process. Trans-science is typically defined as involving judgments that are unresolvable by science, and thus presumptively matters of policy.[151] This definition has intuitive appeal at the extremes—the question whether the sun is going to rise tomorrow is clearly not a matter of policy. The problem is that once one goes beyond such simple cases, it becomes increasingly difficult to determine what is science and what is policy. Is the decision to use a simpli-

148 FOSTER & HUBER, *supra* note 82, at 239–40.

149 Oreskes et al., *supra* note 83, at 642 ("The problem of deductive verification is that if the verification fails, there is often no way to know whether the principal hypothesis or some auxiliary hypothesis is at fault."). Climate scientists must contend with a global-scale system and integrate a vast range of interactions, including solar physics and energy balancing, atmospheric chemistry, and ocean-atmosphere interactions. L.D. DANNY HARVEY, GLOBAL WARMING: THE HARD SCIENCE 121–29 (2000). Calculating the human impact on climate requires scientists to model mathematically each of these features of Earth's climate system. *Id.*

150 GASKINS, *supra* note 83, at 152–53; Jan Beyea & Daniel Berger, *Scientific Misconceptions Among* Daubert *Gatekeepers: The Need for Reform of Expert Review Procedures*, 64 LAW & CONTEMP. PROBS., Spring/Summer 2001, at 327, 331–32.

151 *See supra* Part II.A; Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1619 nn.21–22.

fied model of experimental conditions a matter of scientific judgment or a matter of policy? Or, is the choice between two experimental methods, neither of which is ideal, a policy decision or a scientific determination? Scientists make innumerable decisions like these and, because of their own limited knowledge, also make many implicit judgments that they may not even be able to articulate.[152] It is difficult to see how trans-science can be defined in a coherent manner because, as Popper's work implies, scientific proof accrues in shades of gray; it does not fall neatly into resolvable and unresolvable categories.

Lawyers encounter an analogue of this line drawing problem in critiques of formalist theories of judicial review. As described in the Introduction, legal formalism limits judicial textual interpretation to the plain meaning of the Constitution (just as scientific positivism is limited to facts). Critics of legal formalism point out that it encounters a crucial problem when resolving how generally a constitutional provision should be construed.[153] For example, it is unclear whether the constitutional right to privacy should extend to married couples, to heterosexual couples, to all consensual adult relationships, or to all consensual relationships regardless of the age or maturity of the individuals.[154] Strict legal formalism, as a consequence, is viable for only trivial cases of judicial review, making antiformalist methods unavoidable. This uncertainty is just another instance of underdetermination: Judges are unable to rule out competing interpretations because specific constitutional language is consistent with several competing interpretations, which are supportable by various auxiliary principles a judge may invoke. The science/trans-science distinction is empty for the same reason: Developing science—meaning most science—falls under the category of trans-science, making the distinction heavily one-sided and of limited practical utility.

These problems suggest that the focus on trans-science is misdirected and, at best, a surrogate for what I believe is the real issue—the proper scope of scientific discretion in environmental decisionmaking. The legal analogy applies here as well: The concept of trans-science is used to establish a boundary for scientific expertise that circumscribes what is deemed unalloyed science in the same manner in which legal formalism is often used to limit judicial discretion.

152 Similarly, relying on scientific consensus to delimit science from trans-science, as Wagner does, is troublesome because it is unclear what degree of consensus (or conflict) is sufficient. Furthermore, some important judgments may not be subject to broad scientific debate, making the concept of consensus all but irrelevant.

153 *See* Brest, *supra* note 14, at 1084–85.

154 *Id.*

Once a question is declared trans-scientific, scientific judgment is presumptively owed less (perhaps no) deference and other factors, particularly societal values, are elevated for consideration.[155] There are, of course, very good reasons for questioning whether scientists should have the authority to make decisions that involve matters of policy that transcend their technical knowledge and expertise. Unfortunately, the debate over trans-science has not fostered much thoughtful reflection on scientific discretion; instead, its primary utility has been as an offensive weapon against purportedly overzealous reliance on scientific methods.

An examination of the need for and scope of scientific discretion requires a different approach. Assessing the proper scope of scientific discretion is dependent on one's theory of science, just as judicial discretion is tied to one's theory of judicial review. Scientific positivism, for example, limits scientific discretion to a narrow reading of experimental evidence, whereas Kuhn's theory, which is premised on a balance between standard analytical methods (normal science) and broad principles, affords scientists relatively broad discretion. This association makes it all the more important for lawyers and policymakers to have a concrete understanding of how science is conducted, as they are often in the position of determining how scientific judgments are utilized in environmental law. However, the current debate over environmental science ignores this relationship because it lapses either into positivist caricature, which ignores scientific judgment altogether, or relativist critique, which portrays scientific judgments as inherently political and a matter of policy. The debate over the proper role of scientific expertise in environmental policymaking ought to move beyond these misleading images of science.

A relatively straightforward case can be made for the importance of scientific discretion and judgment in good science. The preceding discussion suggests three basic arguments. First, scientists routinely make judgments in the face of uncertainty, as opposed to slavishly relying on empirical results from which they mechanically derive their conclusions. Consequently, the experience and knowledge upon which scientists draw are of significant value in setting environmental policy. Second, as both Popper and Kuhn demonstrate, science is a process that is reliant on individual judgments and discretion. Protecting the integrity of this process, which must include substantial latitude for individual judgments, is essential if environmental science

155 Breyer's position on this point is mystifying, as he described what scientists like to do in narrow terms and then granted them extremely broad discretion to set environmental policy. *See* BREYER, *supra* note 20, at 42–43, 64–68.

is going to progress.[156] Third, the authority of science does not derive from some form of analytical determinism—scientific methods are fallible and uncertain—but from the potential for science to achieve a status of accepted truth.[157] In short, the scientific processes used to establish facts as accepted truths are contingent on, and thus warrant, scientific judgment.

These points in no way diminish the importance of political concerns about relying on scientific expertise. Instead, they reveal that important tradeoffs exist between deferring to and overriding scientific judgments in environmental policymaking. The term trans-science proves misleading because it presumes the scientific process can be separated from policymaking. In particular, while it is true that well accepted science can be separated from policy, the process used to attain them cannot, as it entails numerous scientific judgments that invariably have important implications for policy. As environmentalists have often argued, the scientific process (like its judicial counterpart) encompasses matters that are hybrid in nature—both questions of science and questions of policy. The challenge inherent in environmental law is to maintain the integrity of both the political and scientific processes at the same time. In this light, the proposals that Breyer and Wagner advocate represent contrasting judgments about

156 Concerns about the corruption of science by politics is evident in objections to the Bush Administration's selection processes for members of federal agency scientific advisory boards and in the volatility of debates over climate change and genetically modified organisms. *See supra* note 106 and accompanying text; Roger A. Pielke, *Policy, Politics and Perspective: The Scientific Community Must Distinguish Analysis From Advocacy*, 416 NATURE 367, 367–68 (2003) (decrying the politicization of science and urging the scientific community to establish formal institutional mechanisms that buffer science from policymaking). Lysenkoism in Russia, from about 1935 to 1964, is the classic example of politics undermining the integrity of science. Trofim Lysenko was a leading proponent of the Lamarckian theory of evolution, which held that beneficial traits could be cultivated and developed during an organism's lifetime and passed on to its offspring. Lysenko was responsible for popularizing this theory and obtaining the support of the Soviet leadership. He was also instrumental in the censorship, disappearance, and imprisonment of numerous scientists who advocated Darwinian natural selection and Mendelian genetics. Russian biology, as a result, languished for three decades under the burden of the governing communist ideology. *See* Bert Black et al., *Science and the Law in the Wake of* Daubert: *A New Search for Scientific Knowledge*, 72 TEX. L. REV. 715, 769–71 (1994).

157 The political scientist Ian Shapiro makes a similar point: "Science holds out the hope that we can get beyond the welter of conflicting opinions and ideological claims to the truth of the matter, that we can come to hold a set of beliefs about an entity, event, or action that is most reasonable under the circumstances . . . . [A]lthough this is often difficult in practice, there is no reason to rule it out in principle." IAN SHAPIRO, POLITICAL CRITICISM 274 (1990).

protecting scientific and political processes, and the weaknesses in their respective approaches are a testament to just how difficult it is to formulate a framework for achieving such a balance. In Part IV, I enter the fray further by proposing an alternative to the approaches advocated by Breyer and Wagner.

The third rationale for scientific authority and discretion is perhaps the most important and least understood. Skepticism about science is often fueled by the general public's lack of understanding about how science is practiced and how it progresses. Skeptics with relativist leanings, in particular, challenge whether scientific truths exist at all and whether science should be accorded the level of authority it often receives. The next section describes scientific methods in detail and makes the case for the power of science to generate accepted truths. Again, my objective is not to argue that science should be set above everything else, but rather to substantiate the claim that important tradeoffs exist between scientific progress and political considerations, and to show how scientific truths emerge from the uncertainties Popper and Kuhn exposed.

C. Experimental Bootstrapping as an Answer to Underdetermination

Breyer and Wagner are correct in portraying chemical risk science as methodologically weak, but the problem is not its simplifying assumptions per se. The flaw in chemical risk models is that current testing methods are neither sufficiently stringent nor well corroborated. An important objective for environmental policy in the area of chemical risk assessment therefore ought to be the development of more powerful testing methods. This would require both a major shift in the science, given the inherent limitations of current chemical testing methods, and significant additional resources.[158] The views of legal scholars such as Breyer and Wagner, however, provide little incentive for such a forward looking approach. To the contrary, debates about environmental science among lawyers have come precariously close to rejecting the contributions science can make in deepening our understanding of chemical toxicology.

Legal commentators fail to acknowledge that basic experimental methods can circumvent the broad theoretical uncertainties Popper and Kuhn identified.[159] Specifically, independent and diverse testing

158 *See* BREYER, *supra* note 20, at 43–47; Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1620–27.

159 This approach rejects Popper's extreme form of underdetermination, which treats "the logically possible and the reasonable [as] coextensive." LAUDAN, *supra* note 30, at 29; *see also id.* at 37–40. Popper's underdetermination thesis amounts to

methods diminish the importance of theoretical underdetermination in obtaining evidence of the truth or falsity of a scientific hypothesis. The power of this approach is illustrated by the late nineteenth century dispute over biometrics and Mendelism.[160] The scientific debate over biometrics and Mendelism was highly politicized because biometrics provided the scientific basis for the eugenics movement, which was popular at the time.[161] The dispute stalled, however, because neither theory was well grounded and only limited experimental methods were available.[162] Scientists debated the form in which the data were quantified: Mendelian scientists used a binary metric that categorized traits into opposites, such as "hairy" or "hairless" and "yellow" or "green."[163] Biometricians used continuous variables, replacing, for example, the hairy-hairless categories with a continuous measure of the number of hairs per unit area.[164] Both groups adopted metrics based on their respective theories to validate them.

---

the "assumption that, unless we can show that a scientific hypothesis cannot possibly be reconciled with the evidence, then we have no epistemic grounds for faulting those who espouse that hypothesis." *Id.* at 29. In other words, if I can contrive a hypothesis, no matter how implausible, that is equally supported by the existing data, then no epistemologically valid reason exists to prefer the hypothesis under investigation over this alternative. The standard example illustrates this point using a collection of observations all finding that emeralds are green. According to the underdetermination thesis, the hypothesis that all emeralds are green is epistemically equivalent to the "Grue" hypothesis, which maintains that all emeralds are green until a new observation is made, after which all emeralds will be blue. While it is true that both theories are logically equivalent, insofar as they both entail the existing data equally well, it stretches credulity to assert that both are equally reasonable. *See id.* at 43–44. This section explains how experimental methods differentiate competing theories.

160 Biometrics and Mendelism were hotly contested at the time. "The biometricians . . . believed that evolution was a process of gradual change, taking place by the selection of continuous differences." DONALD A. MACKENZIE, STATISTICS IN BRITAIN 1865–1930, at 130 (1981). Mendelism, in contrast, held that evolution occurred through discontinuous changes in species as certain traits were activated or deactivated. *Id.* at 131–33.

161 *See id.* at 29–31, 36–40, 52–56.

162 *See id.* at 120–22.

163 *Id.* at 129. "The two sides could not always agree even on the facts that stood in need of explanation or description." *Id.* at 123.

164 *Id.* at 129. At the time, an important methodological debate existed over the statistical validity of two methods for summing up data. *Id.* at 153, 161–75. One group of statisticians argued for the statistical validity of "interval" (continuous) variables, which clearly supported biometrics; a second group urged "nominal" (categorical) variables, which provided a statistical model for Mendelism. The two methods of statistically summing up the data naturally reflected the ideological biases of the individuals who developed them, and the two statistical methods reflected the theory of evolution each supported. *Id.* at 168–75. Biometrics was premised on the continuous

These different approaches to scaling the data made the experimental results incompatible with each other.[165] Data collected by a Mendelian scientist presupposed Mendelian traits that fell into strict categories, whereas data collected by a biometrics scientist was premised on continuously variable traits that could not be fit into Mendelian categories. This incommensurability arose because each group of scientists used their respective theories to construct experimental statistics. As such, neither statistical test was severe in Popperian terms. First, the observation method used, here embodied in the sample statistics, was premised solely on the theory being tested.[166] Second, neither theory was well corroborated. The resulting experiments were weak and circular.[167] However, circular in this context does not mean tautological, as the data could have contradicted the Mendelian theory if, for example, no consistent patterns of heredity were observed.[168] Rather, this circularity implies the test was narrow, much as if one were to base a broad legal doctrine on a single case or incident.

The indeterminacy of the science and the surrounding politics split the scientific community into ideological factions for years.[169] The weaknesses of both sets of experiments led to a stalemate because no independently acceptable basis existed to assess the relative merits of the two theories. Indeed, this same circularity is evident in many

---

variation of biological traits because it maintained that evolution occurred incrementally, whereas Mendelism asserted that biological traits, and thus evolution, were discontinuous, shifting from one state to another.

165 *See id.* at 122–24 ("The incommensurability of the two positions [did lead] to difficulties of understanding and communication.").

166 IAN HACKING, REPRESENTING AND INTERVENING 183–85 (1983) (providing an example of the same dependence of observation and theory in solar physics).

167 *See* KARL R. POPPER, THE POVERTY OF HISTORICISM 108–11 (1957) (raising the same objection in a different context).

168 Kuhn describes this circularity in a slightly different manner:

> If . . . there can be no scientifically or empirically neutral system of language or concepts, then the proposed construction of alternate tests and theories must proceed from within one or another paradigm-based tradition. Thus restricted it would have no access to all possible experiences or to all possible theories. As a result, probabilistic theories disguise the verification situation as much as they illuminate it. . . . Verification is like natural selection: it picks out the most viable among the actual alternatives in a particular historical situation.

KUHN, *supra* note 134, at 146.

Statistical testing, in short, is subservient to substantive theory, and is no more able to resolve scientific questions beyond the theoretical framework from which it originates than scientists operating within a given paradigm.

169 *See id.* at 120–22 (discussing how scientists may come up with different conclusions based on the same observations, to show the indeterminacy of science).

risk assessments and cost-benefit analyses, which lack good independent observational methods and well-established theory. Two critical points emerge from this example: (1) The process of defining a statistical variable imposes significant analytical constraints and reflects the subservience of statistical quantification to scientific theory, and (2) statistical methods are far less powerful if they are dependent on the theory being tested.[170] The Mendelism-biometrics dispute also reveals how and why weak scientific methods reduce questions of quantification to ideological factionalism.

Independent observational techniques overcome circular testing methods and minimize problems that derive from the theory-laden nature of facts.[171] In doing so, this approach decouples theory from fact to a high degree.[172] A standard example of this strategy involves the first observation of "dense bodies" in red blood cells.[173] When dense bodies were initially observed, the scientist conducting the experiment believed they were an artifact of his observational method, here an electron microscope.[174] To test this hypothesis, he selected a different observational technique, a fluorescence microscope, which

170 These practical and epistemological limits derive from chance being in part a measure of human knowledge. The temperature of a gas, for example, operates as a meaningful measure of its total kinetic energy, and thus represents a discrete physical property that is accurately represented by a single metric. One can make such a claim only because the governing physical theory is well established. Other statistical metrics, such as IQ or risk, are much less well theoretically and empirically corroborated. (It is interesting to note that IQ tests were often characterized as a "clinical thermometer." JOANNE BROWN, THE DEFINITION OF A PROFESSION: THE AUTHORITY OF METAPHOR IN THE HISTORY OF INTELLIGENCE TESTING, 1890–1930, at 76–77, 81 (1992).) While it is natural for Bayesians to accept these epistemological constraints, frequentists' aspirations for objectivity run counter to accepting these deeper philosophical limitations..

Both statistical schools have nevertheless characterized statistical techniques in terms of "summarizing" data or evidence, which both qualifies the purported objectivity of quantitative statistical techniques and highlights their reductive functions. Frequency-type theorists, for example, have stated that statistical information is best understood as an "extremely brief summary of the [available] data bearing on the true value of some magnitude [of a system or thing]." HACKING, *supra* note 33, at 173; *see also* HACKING, *supra* note 12, at 214–15. Similarly, Bayesians have claimed that Bayes's method "sum[s] up the evidential meaning of new information." *Id.* at 181; *see also* HOWIE, *supra* note 37, at 114–15 (preferring a "Simplicity Postulate" to the "frequency theory").

171 *See* HACKING, *supra* note 166, at 184–85 (arguing both that observation-theory independence generates more compelling experiments and that the experiments themselves are not necessarily dependent on theory); LAUDAN, *supra* note 30, 48–49.

172 *See* MAYO, *supra* note 23, at 16–17.

173 See Hacking, *supra* note 166, at 200–02.

174 *Id.*

operated according to completely different physical principles.[175] As it happened, the scientist observed the dense bodies with the fluorescence microscope, refuting his original hypothesis that they were an artifact.[176] The logic behind this approach is straightforward:

> Two physical processes . . . are used to detect [dense] bodies. These processes have virtually nothing in common between them. They are essentially unrelated chunks of physics. It would be preposterous coincidence if, time and again, two completely different physical processes produced identical visual configurations which were, however, artifacts of the physical processes rather than real structures in the cell.[177]

Under this approach, the crucial assumption (i.e., auxiliary hypothesis) is the independence of the distinct fields from which the observational methods derive. Further, by taking advantage of the theoretical disunity of science in this manner, the significance of high-level theories (which invariably require more simplifying assumptions to test) is minimized because an overarching theory is unnecessary.[178]

The examples discussed so far bear out the value of this approach. Since the Mendelism-biometrics dispute arose in the nineteenth century, scientists have made huge advances in biochemistry and analytical chemistry, which in turn have enabled genes to be observed and tested through a variety of experimental methods. These developments avoided the circularity of the earlier debate and defused it altogether. In short, science reduced an ideology-laden debate over statistical quantification to one based on well corroborated, independently validated facts. Such shifts in theory and experimental methods continue to be critical today. For example, recent developments in toxicogenomics, which involves the application of rapid genetic screening methods to the field of toxicology,[179] are likely to revolutionize current chemical toxicity testing methods by offering

---

175 *Id.* at 201.

176 *Id.*

177 *Id.*

178 *Id.* at 208–09; *see also* MAYO, *supra* note 23, at 8–9, 16–17.

179 Toxicogenomics is based on obtaining a profile of

> the cell-wide changes in gene expression following exposure to toxins. This approach creates the potential to provide a molecular "fingerprint" of exposure or toxicological response to specific classes of toxic substances . . . . The toxicological significance of gene expression changes must be validated, including an evaluation of the robustness of [gene expression] results between or across different laboratories, species, individuals, tissues and time periods.

Gary E. Marchant, *Toxicogenomics and Toxic Torts*, 20 TRENDS IN BIOTECH. 329, 329–30 (2002).

new faster, less expensive alternatives.[180] These methods also have the potential to achieve for toxic risk assessment what biochemistry did for Mendelian genetics—the replacement of a weak testing regime with a set of diverse, independent methods.[181] The critical point here is that science can lead to an array of independent testing methods, which both generate reliable, objectively verifiable data and enhance our theoretical understanding of challenging problems in environmental science.

The value of this approach lies not just in the credibility of the science, but also in efficiency gains. Research on climate change, for example, is heavily dependent on statistical methods because the principal metric of climate change is a statistically averaged global surface temperature.[182] Predictably, this single metric approach has been criticized because it obscures a great deal of important regional information.[183] The response of scientists has been pragmatic and pluralistic: develop the best information possible using a single metric and, at the same time, work towards developing alternative predictors of climate change.[184] Under this approach, climate change research is not limited to a group of chemists and physicists developing massive computer models for simulating future increases in the average global surface temperature. Instead, studies in areas as diverse as lake ecology, glaciology, tropospheric chemistry, and volcanism are being conducted under the umbrella of climate change research.[185] These studies do not provide data that are fed directly into large-scale climate models; instead, they create a composite picture, based on inde-

180 *See id.* at 330–32; *see also* Jocelyn Kaiser, *Tying Genetics to the Risk of Environmental Diseases*, 300 SCIENCE 563, 563 (2003); Gary E. Marchant, *Genetics and Toxic Torts*, 31 SETON HALL L. REV. 949, 980 (2001); Gary E. Marchant, *Genomics and Toxic Substances: Part I—Toxicogenomics*, 33 ENVTL. L. REP. 10,071 *passim* (2003).

181 Marchant, *supra* note 179, at 329.

182 *See supra* note 77; HARVEY, *supra* note 149, at 75; Stephen H. Schneider, *Detecting Climatic Change Signals: Are There Any "Fingerprints"?*, 263 SCIENCE 341 (1994). The limits of existing climate models force a methodological compromise: scientists must balance their interests in exploiting more detailed information using a multivariate method against obtaining data relevant to existing climate models. *See id.* at 341.

183 Schneider, *supra* note 182, at 341, 345. To their credit, climate scientists have openly acknowledged the limits of a single metric approach. *See id.*; CLIMATE CHANGE: THE IPCC SCIENTIFIC ASSESSMENT xii–xiii, xxv, xxxv–xxxix, 247–48 (J.T. Houghton et al. eds., 1990) [hereinafter IPCC].

184 *See* Schneider, *supra* note 182, at 345–46; IPCC, *supra* note 183, at 247–48. Scientists must "[w]ork across many scales and disciplines to understand physical, chemical, biological, and relevant societal processes, [and] their interactions . . . ." Schneider, *supra* note 182, at 345.

185 HARVEY, *supra* note 149, ch. 2.

pendent measurements, that is far more powerful than any one of the studies on its own ever could be. This pluralistic approach is critical because climate models are decades away from being able to predict regional impacts with a reasonable level of accuracy. Complementary studies fill in some of these gaps and enhance the credibility of the science as a whole.

The advances in climate change research have been extraordinary given the relative youth and difficulty of the field. In an area arguably far more complex than chemical toxicology, scientists have produced predictive models, numerous experimental methods, and a huge amount of valuable information. This success derives in significant part from the multidisciplinary approach scientists have taken and their cultivation of independent measurements. A secondary benefit of this strategy is that it often allows scientists to make simpler arguments—explaining why two methods are distinct often is easier than explaining their mechanisms in detail. Similar to the complementary microscope measurements described in the dense body example, most people can understand intuitively that atmospheric physics and glacial ice core data derive from distinct areas of science and that complementary results from independent fields strongly reinforce each other. This accessibility and the decoupling of fact from theory lessens, though by no means eliminates, the ideological nature of the debate over climate change in the scientific community and, to a lesser degree, beyond it as well.

As lawyers, we inevitably focus on institutional and legal mechanisms for solving difficult problems in environmental policy. This tendency is encouraged by current polarized conceptions of science, which characterize science either as mechanistic or as infused with difficult value judgments and theoretical indeterminacies. As I have tried to show above, neither science nor its quantitative methods fall neatly into these extremes. Simple experimental strategies can generate reliable facts even where significant theoretical indeterminacies exist. Environmental law and policy would benefit from lawyers and policymakers having a greater appreciation of the basic experimental methods used to resolve important controversies in environmental science.

III. Misconceptions About Traditional Statistical Methods in Environmental Policy

Statistical methods, either Bayesian or frequentist, provide the basic analytical framework for scientists to draw inferences from the experiments they conduct. Laypeople typically associate statistical

inference with frequentist significance tests, which assess whether experimental data deviate "significantly" from the predictions of a test hypothesis. Statistical significance for most people therefore implies that a scientist has discovered an effect not predicted by that scientist's starting hypothesis. Few lawyers, and indeed relatively few scientists, have experience with Bayesian methods, which directly measure the probability that a hypothesis is true rather than providing a rigorous test of its validity. This Part will examine frequentist methods of statistical inference and will focus on presumptions about burdens of proof in drawing scientific inferences from discrete scientific studies—stage two of the three-stage framework described in the Introduction. Bayesian methods will be considered in the final Part of the Article.

Frequency-type statistical inference, as discussed in Part I, is inextricably tied to a world view in which phenomena are defined by their frequencies in abstract populations or classes. Environmentalists find this view problematic for two central reasons. First, statistical inference becomes dependent on this dubious "bingo game" model of the universe and science is practiced by experimentally isolating and randomly sampling such populations.[186] For environmentalists, this model appears to conflict with more holistic ecological models, as it is premised on a disconnected and atomized world ruled by chance. Second, and most importantly for this Part, frequentist methods almost invariably presume that environmental impacts are benign until proven guilty.[187] In order to connect the discussion to a familiar legal doctrine, I will examine criticism of frequentist methods that are based on the Precautionary Principle, which draws on and is intended to alter traditional methods of scientific inference.[188] The tensions between frequentist statistical methods and the Precautionary Principle will be evaluated (largely agnostically), and a novel approach to

186 *See* HOWIE, *supra* note 37, at 74; Collins, *supra* note 30, at 336. Recall the guiding metaphor of frequentist statistics is Mendelian genetics, under which long-run frequencies are determined by random selections of specific traits, as opposed to specific relationships or causes (i.e., biological, chemical, or physical). *See supra* Part I.

187 *See* Katherine Barrett & Carolyn Raffensperger, *Precautionary Science*, *in* PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 106, 111–12; Carl F. Cranor, *Asymmetric Information, The Precautionary Principle, and Burdens of Proof*, *in* PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 74, 79; D.H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1345 (1986).

188 Barrett & Raffensperger, *supra* note 187, at 108–09, 115; Andrew Jordan & Timothy O'Riordan, *The Precautionary Principle in Contemporary Environmental Policy and Politics*, *in* PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 15, 17 (noting that the Precautionary Principle "challenges the established scientific method").

frequentist statistical testing will be proposed that addresses concerns about systemic biases in traditional methods of scientific inference.

The Precautionary Principle embodies the old adage "better safe than sorry" by placing protection of public health and the environment above other interests even when evidence of harm is not proven definitively.[189] The Precautionary Principle is premised on the belief that "[i]f there is a potential for harm from an activity and if there is uncertainty about the magnitude of impacts or causality, then anticipatory action should be taken to avoid harm."[190] The Rio Declaration on Environment and Development describes the Precautionary Principle as follows: "In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation."[191]

In at least some of its myriad formulations, the Precautionary Principle proposes a balancing test of sorts, under which the potential level of harm, degree of scientific uncertainty, and likely alternatives for a product or action are assessed to determine the appropriate regulatory strategy.[192] If, for example, the potential level of harm from a product is great, the scientific uncertainty significant, and numerous low cost alternatives available, the Precautionary Principle would favor a ban on the product. Conversely, if the level of harm is low, the scientific uncertainty minimal, and the alternatives limited and very expensive, the Precautionary Principle would favor less stringent

189 Frank B. Cross, *Paradoxical Perils of the Precautionary Principle*, 53 WASH. & LEE L. REV. 851, 851 (1996).

190 *Introduction* to PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 1.

191 *The Rio Declaration on Environment and Development*, U.N. Conference on Env't & Dev., U.N. GAOR, 46th Sess., U.N. Doc. A/CONF.151/5/Rev.1 (1992), *reprinted in* 31 I.L.M. 874, 880 (1992). Different versions of the Precautionary Principle appear in a variety of other international agreements. *See* David Santillo et al., *The Precautionary Principle in Practice*, *in* PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 36, 41–45.

192 *See* Nicholas A. Ashford, *A Conceptual Framework for the Use of the Precautionary Principle in Law*, *in* PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 198, 199–200 (discussing the elements of the Precautionary Principle); Jordan & O'Riordan, *supra* note 188, at 25 ("[P]recaution is often linked to some consideration of risks, financial costs, and benefits."); Deborah Katz, *The Mismatch Between the Biosafety Protocol and the Precautionary Principle*, 13 GEO. INT'L ENVTL. L. REV. 949, 956–57 (2001) (illustrating the elements of the Precautionary Principle).

regulation. More complicated balancing is required when cases fall between these extremes.

The Precautionary Principle has both Bayesian and frequentist characteristics. The balancing test described above, for example, suggests a straightforward weighting of the evidence that would be compatible with a Bayesian approach. However, perhaps because frequentist methods are better known, the Precautionary Principle is more often discussed in frequentist terms. Thus, proponents of the Precautionary Principle justify it on the ground that the uncertainty of risk ought to be borne by the regulated industry, rather than the "potential victims."[193] This rationale is often expressed in terms borrowed from frequency-type probability theory:

> When a regulator makes a decision under conditions of uncertainty, there are two possible types of error. The regulator can overregulate a risk that turns out to be insignificant or the regulator can underregulate a risk that turns out to be significant. If the regulator erroneously underregulates, the burden of this mistake falls on those individuals who are injured or killed, and their families. If a regulator erroneously overregulates, the burden of this mistake falls on the regulated industry[,] which will pay for regulation that is not needed. This result, however, is fairer than setting the burden of uncertainty about a risk on potential victims.[194]

As this account suggests, the Precautionary Principle incorporates basic rules about minimizing error rates from frequentist inference methods.[195] In this context, erroneous overregulation and underregulation are variants of statistical significance (i.e., Type I error or false positives) and power (i.e., Type II error or false negatives).[196] Environmentalists have used the frequentist framework to argue that

---

193 Sidney A. Shapiro, *Keeping the Baby and Throwing Out the Bathwater: Justice Breyer's Critique of Regulation*, 8 ADMIN. L.J. AM. U. 721, 732 (1995).

194 *Id.* Under this view, differences in who bears the risk (victims versus stockholders, employers, and consumers), the number of people who bear the risk (few versus many), and the types of risks (financial versus emotional and psychological) justify affording higher protection to "potential victims." *Id.*

195 *See* Talbot Page, *A Generic View of Toxic Chemicals and Similar Risks*, 7 ECOLOGY L.Q. 207, 219–20, 230–39 (1978). This early articulation of the Precautionary Principle drew directly on frequency-type probability theory:

> The costs of wrong decisions are asymmetric for environmental risk in inverse proportion to the potential net costs and benefits associated with each of the two hypotheses. The cost of a false negative—deciding that the benign hypothesis is true when it is not—is much higher than the cost of a false positive—deciding that the catastrophic hypothesis is true when it is not.

Id. at 220 (footnote omitted).

196 *See infra* Part III.A.

Type II errors, meaning the risks from underregulation, should be accorded much greater weight than Type I errors in standard statistical tests used in regulatory environmental science.[197] Of course, one could, and many people do, disagree with this approach as a general rule, as instances will exist in which the net societal harm from overregulation is greater than underregulation.[198] For the purposes of this discussion, disagreements over this point are unimportant; one need only accept that the risks from underregulation sometimes will clearly outweigh those from overregulation.

The conventional levels for statistical significance are an obvious target because they are arbitrarily set.[199] If one accepts the Precautionary Principle, raising Type I errors and lowering Type II errors in the regulatory context is thus perfectly acceptable to account for asymmetries between potential victims and regulated industries.[200] However, while this rationale is valid, it often ignores the indirect nature of frequentist concepts and overemphasizes their role in scientific determinations. To begin with, statistical significance is a measure of the reliability of a statistical test; it is not a *direct* standard of persuasion like "beyond a reasonable doubt."[201] Thus, the direct result from raising the significance level of a statistical test is that the threshold for rejecting a test hypothesis is lowered.[202] This change is only indirectly

197 *See* Ashford, *supra* note 192, at 202–03; Barrett & Raffensperger, *supra* note 187, at 117; Lene Buhl-Mortensen, *Type-II Statistical Errors in Environmental Science and the Precautionary Principle*, 32 MARINE POLLUTION BULL. 528, 529–31 (1996); Cranor, *supra* note 187, at 79; Mark Geistfeld, *Reconciling Cost-Benefit Analysis With The Principle That Safety Matters More Than Money*, 76 N.Y.U. L. REV. 114, 118–19 (2001); Reed F. Noss, *Some Principles of Conservation Biology, As They Apply to Environmental Law*, 69 CHI.-KENT L. REV. 893, 896–97 (1994); Randall M. Peterman & Michael M'Gonigle, *Statistical Power Analysis and the Precautionary Principle*, 24 MARINE POLLUTION BULL. 231, 231–33 (1992); K.S. Shrader-Frechette & E.D. McCoy, *Statistics, Costs and Rationality in Ecological Inference*, 7 TRENDS IN ECOLOGY & EVOLUTION 96, 97 (1992); Michele Territo, *The Precautionary Principle in Marine Fisheries Conservation and the U.S. Sustainable Fisheries Act of 1996*, 24 VT. L. REV. 1351, 1351–52 (2000).

198 *See, e.g.*, Cross, *supra* note 189, at 859–61.

199 *See infra* Part III.A.; HACKING, *supra* note 12, at 225; Collins, *supra* note 30, at 339.

200 Page, *supra* note 195, at 230–39.

201 *See* David F. Parkhurst, *Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation*, 51 BIOSCIENCE 1051, 1057 (2001); *see also* Lawrence H. Lehmann, *The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?*, 88 J. AM. STAT. ASS'N 1242, 1243 (1993) (noting that mathematical uncertainty or probability is not an adequate expression of uncertain inferences of every kind).

202 *See infra* Part III.A (noting that statistical testing thus involves a one-sided competition between the null hypothesis and the conjecture that turns on the fidelity of the null-hypothesis model in matching the experimental data).

related to a legally required burden of persuasion, and its impact on Type II errors is not as simple as it might seem initially.[203] The following sections clarify these relations, suggest a direct method for addressing environmentalists' concerns about Type II errors and allocating the burden of proof, and examine the respective limits of statistical inference and the Precautionary Principle in scientific decisionmaking.

A. *The Indirect Nature of Frequentist Statistical Inference*

The frequentist definition of probability is central to understanding traditional methods of statistical significance testing. Frequentists define probability as the long-run frequency or propensity of a population, system, or thing.[204] The properties that may be studied are almost infinitely variable, limited only by imagination and what can be measured. The concept of long-run frequency has been aptly characterized as "combin[ing] individual irregularity with aggregate regularity," such that measurement of a system's long-run frequency converges to a fixed value as the number of observations increases.[205] The long-run frequency of a fair coin turning up heads, for example, converges to one-half as the number of trials approaches infinity.[206] Scientists thus collect repeated measurements (i.e., sample) of a population to obtain an accurate measure of such long-run frequencies. Statistical significance testing assesses the correspondence of such statistical samples with hypotheses regarding the true long-run population frequency being measured.

Statistical inference for frequentists revolves around determining the degree to which an experimental sample statistic is approximated by a normal distribution model.[207] For example, suppose you believe

---

203 *See, e.g.*, George Casella & Roger L. Berger, *Reconciling Bayesian and Frequentist Evidence in the One-Side Testing Problem*, 82 J. AM. STAT. ASS'N 106 (1987); Morris H. DeGroot, *Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio*, 68 J. AM. STAT. ASS'N 966 (1973).

204 HACKING, *supra* note 33, at 1–2.

205 *Id.* at 5 (quoting JOHN VENN, THE LOGIC OF CHANCE 4 (London, Chelsea 1866)); *see also* HACKING, *supra* note 12, at 145, 190–91, 196–97.

206 HACKING, *supra* note 12, at 191–92; HACKING, *supra* note 33, at 214. Propensity theorists ignore long-run frequencies, focusing instead on those attributes that cause fixed frequencies. *See* HACKING, *supra* note 12, at 145; PORTER, *supra* note 29, at 121–22. A die, for example, has symmetry properties that dictate its probabilistic tendencies. According to this account, probabilistic models, like abstract physical theories, embody mathematically specific properties of the systems or things they accurately represent.

207 The "binomial distribution" is a mathematically precise representation of a coin tossing system; the normal distribution is a fair approximation to the binomial

the coin in your pocket is fair and you want to test the validity of this starting hypothesis by flipping the coin a thousand times. For any fixed number of observations, the normal distribution offers a simplified model for the distribution between heads and tails, which in this case predicts that there is about a two-thirds probability of the number of heads lying between 495 and 505 and a 0.95 probability of it lying between 490 and 510.[208] The normal distribution in this case provides a mathematical approximation of experimental conditions in which variability is purely random and the coin has an equal probability of obtaining a heads or tails on each toss (fairness). If your experimental result were 491 heads, you would be reasonably confident in the fairness of the coin; conversely, if your experimental result were 400 heads, you would likely question your initial hypothesis about the coin's fairness.[209]

Testing a pesticide's toxicity provides a more informative and realistic example of significance testing.[210] The basic approach, however, is the same: Just as a normal distribution can be used to model the behavior of a fair coin, it may be used as a model of experimental conditions limited by random errors, which requires a carefully controlled testing regime.[211] Frequentists utilize the following convention for hypothesis testing: (1) a "null" hypothesis, which assumes no effect exists (i.e., the pesticide is harmless); and (2) a "conjecture," which assumes some effect exists (i.e., the pesticide has a discernable

---

distribution for tests involving at least thirty trials (i.e., flips of the coin). MAYO, *supra* note 23, at 171.

208 HACKING, *supra* note 12, at 203–04, 206.

209 The detailed analysis is actually a little more complicated. The assumption that the coin is fair implies that an event of probability much less than one percent would have occurred if you obtained 400 heads. One does not, however, reject one theory in a vacuum, but only rejects it if a better one exists. HACKING, *supra* note 33, at 79–81. If a statistical analysis reveals the chance of the experimental results occurring is very small, say 0.0001, there are two possible inferences one can make. *Id.* at 65, 83. One might attribute the low value to the observation theory, i.e., the statistical and experimental methods; for example, the result could imply that the flips of the coin in the preceding example were not independent. *Id.* at 83. Alternatively, one might attribute it to the explanatory theory (the fairness of the coin) if independence is well founded. In either case, a low statistical value merely shows that "if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say 0.5 . . . you will be very much more inclined to consider that the original hypothesis is not true." *Id.*

210 HACKING, *supra* note 12, at 213–15.

211 The experiment must be designed to ensure that the experimental observations are independent (i.e., each experimental test for toxicity is independent of the others) and that systematic errors are minimized (e.g., randomization, double-blind testing, etc.). *See* MAYO, *supra* note 23, at 4–7, 16–17.

toxic effect).[212] In this scheme, the null hypothesis incorporates the normal distribution as a model of the population (i.e., a human population insensitive to pesticide exposure) that the experiment is sampling; the experimental data are then compared against this population model.[213]

The null-hypothesis method leads to a counterintuitive result: The probability calculated is *not* the probability that the pesticide is harmful, but rather the probability of obtaining the experimental data assuming the null hypothesis is true.[214] In the pesticide example, the incidences of harm observed experimentally are compared against the likelihood of that frequency of harm occurring if the pesticide had no effect. As a result, a high probability of obtaining the experimental results under the null-hypothesis model of the experiment implies "we cannot tell which hypothesis is correct"—a different, untested hypothesis could have a higher probability.[215] Conversely, a low probability indicates "the null hypothesis seems likely to be false."[216] In either case, frequency-type statistical testing does not provide a straightforward assessment of the probability that the pesticide is harmful; it is instead based on two measures of the null-hypothesis model's error rates—significance and power.[217]

---

212 HACKING, *supra* note 12, at 214. The relevant data for this test are the incidences of harm (e.g., carcinogenisis) among individuals exposed to the pesticide and among those not exposed, generally referred to as the control group. Under the null hypothesis, the two populations (exposed and unexposed) are assumed to have equal incidences of harm.

213 *Id.*; *see also* Collins, *supra* note 30, at 335. The null hypothesis is a matter of faith, not of logic or science, and thus is not an ultimate criterion of truth because "[t]here is no way to test a statistical model statistically . . . . [Such an effort] leads only to logical regress." *Id.* at 336. Moreover, while it makes sense that disparate causal chains should be treated as completely independent, many gradations exist in between. As Keynes observed, "[a] remote connection or a reaction quantitatively small is a matter of degree and not by any means the same thing as absolute independence." KEYNES, *supra* note 58, at 283. Other theorists have acknowledged the importance of "non-normal" distributions, particularly in heterogeneous systems, but these efforts have been largely ignored. *See* PORTER, *supra* note 29, at 264–65, 307–10.

214 HACKING, *supra* note 12, at 214–15. In a 1986 article, Professor David H. Kaye provided a very clear exposition of the confusion that often arises in the context of legal actions over the meaning of statistical significance. *See* Kaye, *supra* note 187.

215 Parkhurst, *supra* note 201, at 1057.

216 *Id.* Stated otherwise, the statistical testing of the null-hypothesis model asks the question: "Do we lack evidence that the [pesticide] is not safe . . . ?" *Id.* at 1052. Accordingly, interpreting failure to reject the null hypothesis as proof of its validity is the "equivalent of failing to find a pair of pliers in a quick search of a messy garage and claiming that failure to be good evidence that the pliers were not there." *Id.* at 1053.

217 HACKING, *supra* note 12, at 209–15, 223–25.

The principle that underlies this approach is simple: "[T]here should be very little chance of mistakenly rejecting a true hypothesis . . . [and] a good chance of rejecting false hypotheses."[218] The significance of a test is thus defined as the probability of rejecting the null hypothesis when it is true (i.e., a Type I error).[219] Similarly, the power of a test is defined as the probability of accepting the null hypothesis when it is false (i.e., a Type II error).[220] Following this principle, the general rule is that experiments should have low significance and high power.[221] This rule is difficult to implement for two reasons. First, it is often difficult to formulate an appropriate measure for power, which leads investigators to ignore power altogether.[222] Second, an inherent tradeoff exists between minimizing significance and maximizing power—the basic mathematics makes it impossible, as a general rule, to minimize them simultaneously.[223] The term "significance test" is not arbitrary in this respect; it implies that traditional frequentist testing focuses on statistical significance, not power.

Statisticians have responded to these constraints by adopting a convention: They minimize only Type I errors and, where possible,

---

218 HACKING, *supra* note 33, at 92. This approach is referred to as the "Neyman-Pearson" theory, as it was first developed by Jerzey Neyman and Egon S. Pearson. Karl Popper's falsificationalism adopts an analogous approach to scientific testing using frequency-type probability. *See* POPPER, *supra* note 121, at 198–205; Lakatos, *supra* note 117, at 109 & n.6.

219 HACKING, *supra* note 12, at 212–13, 223–25; HACKING, *supra* note 33, at 92.

220 HACKING, *supra* note 12, at 224–25.

221 HACKING, *supra* note 12, at 225; HACKING, *supra* note 33, at 92–93.

222 *See* HACKING, *supra* note 12, at 224–25; Lehmann, *supra* note 201, at 1244–45; McBride, *supra* note 28, at 19. If we return to the pesticide example, delimiting the potential alternative hypothesis is far from straightforward. The alternative to "harmless" is not "harmful," it is actually a host of alternative hypotheses (and degrees of potency) that entail some kind of harmful interaction. These problems arise for the same reason that scientific inference generally is difficult: It is impossible to rule out all possible alternative hypotheses. *See supra* Part II.B.; *see also* R. Lewin, *Santa Rosalia Was a Goat*, 221 SCIENCE 636, 639 (1983) (providing an example of poor information and theory for development of alternatives to the null-hypothesis model in ecological science). Nevertheless, in certain well defined experiments these indeterminacies can be minimized, and the power of an experiment may be reduced to a relatively simple function of the sample size. MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 182–88 (2d ed. 2001); Dennis, *supra* note 23, at 1101.

223 *See* HACKING, *supra* note 12, at 224–25; HACKING, *supra* note 33, at 92–93. The reason for this tradeoff becomes apparent if one considers the extreme cases of obtaining zero Type I or II error. If Type I errors were set at zero, the test would effectively reject the null hypothesis all the time, causing Type II errors to increase substantially, and an analogous increase in Type I errors would occur if Type II errors were set at zero. In between these extremes, a tradeoff exists between the two types of errors and no general method exists for simultaneously minimizing them. *Id.*

formulate a null hypothesis for which Type I errors are the more serious ones.[224] In practice, however, the starting hypothesis in a significance test is by default a no-effect null hypothesis, meaning that the Type I error being minimized in most statistical tests is identifying a risk where none exists—not failing to discover a risk that is real. Following this convention, statistical tests are characterized by their "significance level" (i.e., Type I error rate), such that a test is "significant at the one-percent level" when the null-hypothesis model of the experiment predicts that there is a 1% chance of observing the experimental result.[225] More concretely, if a pesticide were, in fact, not harmful, there would be only a 1% chance of observing the relative increase in incidences of harm observed experimentally. Significance levels are typically either 0.05 (95%) or 0.01 (99%) but, as suggested above, these standard levels are neither driven by principle nor logical necessity.[226] To the contrary, they represent an arbitrary rule established by convention that early on was likely dictated by mathematical simplicity.[227]

Environmentalists focus on Type I and II error rates because the no-effect null hypotheses used almost universally in significance testing are contrary to the Precautionary Principle. In the pesticide example, for instance, the starting hypothesis was that the pesticide was harmless. This formulation fails to minimize the errors of greater

224 *See* MAYO, *supra* note 23, at 372–74; J. NEYMAN, FIRST COURSE IN PROBABILITY AND STATISTICS 261–64 (1950).

225 HACKING, *supra* note 12, at 212–13. As such, "a hypothesis or significance test determines whether an observed result is so unlikely to have occurred by chance alone that it is reasonable to attribute the result to something else." Kaye, *supra* note 187, at 1333. More precisely, a 1% significance level means the following: "If a designated null hypothesis is true, then, using a certain statistic that summarizes information from an experiment like ours, the probability of obtaining the data that we obtained, or less probable data, is 0.01." HACKING, *supra* note 12, at 215. In the absence of a conjectured theoretical model, significance testing takes on a mindless quality because it amounts merely to finding that "[*e*]*ither* the null hypothesis is true, in which case something unusual has happened by chance (probability 1%), *or* the null hypothesis is false." *Id.* at 243.

226 HACKING, *supra* note 12, at 216–18. Confidence limits, which are related to significance but used for point estimates, often also employ 95% or 99% limits by convention. A "confidence limit" of 95% represents the following: the point estimate with which it is associated was made using a procedure that gives a correct estimate 95% of the time. *Id.* at 234–36, 240–41.

227 *See* Collins, *supra* note 30, at 337, 339; Kaye, *supra* note 187, at 1343–45; Lehmann, *supra* note 201, at 1244. The choice of 0.05 and 0.01 is at least partly a "mathematical accident" based on the normal distribution, for which "it is unusually easy to compute the 99% and 95% accuracy probabilities for some phenomena." HACKING, *supra* note 12, at 217.

concern to environmentalists (i.e., failing to regulate when the pesticide is in fact harmful) because they are treated as Type II errors. Environmentalists argue that asymmetries in the severity of Type I and II errors can be corrected by relaxing a statistical test's significance level, which they believe shifts the presumption away from the null hypothesis and, in effect, lowers the burden of persuasion for finding harm.[228] This reasoning illustrates two important misconceptions about frequentist methods. First, it conflates the frequentist and Bayesian theories by interpreting the indirect statistical error rates of frequentist significance testing as Bayesian degree-of-belief probability.[229] Second, it presumes that a simple relationship exists between Type I and II errors.

Frequentist methods, as described above, employ null-hypothesis error rates, not standards of proof. Thus, the proper interpretation of

---

228 *See supra* note 197 and accompanying text. For lawyers, the logic of this position appears self-evident, especially in light of longstanding Supreme Court jurisprudence on burdens of persuasion. A good example of this is Justice Harlan's opinion in *In re Winship*:

> The standard of proof influences the relative frequency of these two types of erroneous outcomes. If, for example, the standard of proof for a criminal trial were a preponderance of the evidence rather than proof beyond a reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but far greater risk of factual errors that result in convicting the innocent. Because the standard of proof affects the comparative frequency of these two types of erroneous outcomes, the choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect an assessment of the comparative social disutility of each.

397 U.S. 358, 371 (1970) (Harlan, J., concurring); *see also* David H. Kaye, *Statistical Significance and the Burden of Persuasion*, 46 LAW & CONTEMP. PROBS., Autumn 1983, at 13, 14–17.

229 D.H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 CORNELL L. REV. 54, 57 (1987) (noting that the "unholy union" of frequency and belief-type theories of probability leads to incoherence and "yields arbitrary and unjustifiable results").

> The burden of persuasion [i.e., degree of reasonable belief] is . . . not the likelihood that the effect found was due to random error. Using statistical significance as the equivalent of the burden of persuasion is, as David Kaye has trenchantly stated, like 'trying to find the shortest path from Oxford to Cambridge by scrutinizing a map of London.'

Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation*, 86 NW. U. L. REV. 643, 686 (1992); *see also* Kaye, *supra* note 228, at 21–23. Numerous examples of this confusion exist. *See, e.g.*, K.S. SHRADER-FRECHETTE, RISK AND RATIONALITY: PHILOSOPHICAL FOUNDATIONS FOR POPULIST REFORMS 132–34 (1991); *To Foresee and to Forestall*, *in* PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 1, 3; Barrett & Raffensperger, *supra* note 187, at 111–12; Cranor, *supra* note 187, at 79.

a significance test with a 95% significance level is not that failing the test (i.e., being statistically significant) means the null hypothesis has a 95% chance of being false. Instead, meeting this error rate means that the null-hypothesis model has less than a 5% chance of generating the observed data. In the pesticide example, the fact that the null hypothesis has a low probability of predicting the experimental results does not preclude it from being the most likely hypothesis—the experimental results could simply represent a rare event.[230] Interpreting frequentist significance levels as quantifying the degree of support for a hypothesis is equivalent to concluding that where *A* implies *B* it necessarily follows that *B* implies *A*. Significance tests quantify how likely a test hypothesis is to predict the observed data; they do not quantify how well the observed data support a test hypothesis. Only under certain limited circumstances may frequentist null-hypothesis error rates be quantifiably related to a burden of persuasion and, even where they can, the relationship is not a simple one in which unique rates of Type I and II errors correspond to a specific burden of persuasion.[231] Typically, frequentist error rates will change from experiment to experiment for a given burden of persuasion.[232]

The effect of varying Type I and II errors must be carefully considered for several additional reasons. First, arguments regarding statistical error rates generally devolve into a rejection of conventional significance levels with little or no consideration for how Type II errors are affected. While it is true that increasing the significance level of a test lowers Type II errors, a simple one-to-one relationship does

230 The fact that a hypothesis explains observed data well does not necessarily imply that it is the most probable account. An exceedingly rare genetic disorder might be consistent with certain observed symptoms, but if the symptoms also were reasonably consistent with a very common virus, a doctor will choose the latter in her diagnosis of the patient because it is so much more likely to occur. Similarly, the fact that a hypothesis is only marginally consistent with experimental results does not necessarily imply that it is not the most probable explanation. This is no different than if you were to role double sixes five times consecutively in a game of backgammon. The likelihood of this occurring with fair dice is exceedingly low, but if you have no other reasons to believe that the dice are fixed, you could reasonably conclude that a remarkably rare event just occurred rather than that the dice are unfair. These examples involve what are often referred to as "base-rate" problems, one of which (the taxi cab example) is discussed in detail in Part IV.A.

231 *See* Kaye, *supra* note 187, at 1355–56, 1362–63. Moreover, where multiple hypotheses are at issue, other analytical problems may arise. *See* David Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation*, 1982 AM. B. FOUND. RES. J. 487, 508–13.

232 *See* M. DEGROOT, PROBABILITY AND STATISTICS 373–82 (1975); Kaye, *supra* note 229, at 71–73; Kaye, *supra* note 228, at 17, 21–23.

not exist between them.[233] The relationship between the two types of errors is complicated by the fact that Type II error is determined by several other independent factors, such as the size of the data set and the background incidence of the phenomena being studied.[234] Second, indiscriminately raising the significance level of an experiment can lead to perverse results: increased total experimental error (i.e., combined Type I and II errors) with only a marginal decrease in Type II errors.[235] In such cases, statistical reliability is sacrificed without environmental concerns benefiting from a more rigorous vetting of the data.[236] Precautionary Principle proponents must thus be careful in how they relate significance levels to legal burdens of persuasion and how they seek to balance perceived asymmetries between Type I and II errors in a regulatory context. It is essential to understand that frequentist methods test hypotheses stringently; they do not quantify their probability of being true directly. For frequentists, confidence in a hypothesis instead accrues qualitatively through a hypothesis satisfying multiple tests.

B. Equivalence Testing: A Direct Method for Minimizing Type II Errors of Environmental Significance

The Precautionary Principle has undeniably helped to expose the systemic bias in traditional significance testing methods, which employ, generally by default, a no-effect null hypothesis. One can disagree in specific cases whether an asymmetry exists between underregulation and overregulation, but few people would deny that

233 *See* DEGROOT, *supra* note 232, at 275–78 (noting that a large increase in significance level may not have a marked effect on an experiment's power and, within a certain range, may have little effect at all); Green, *supra* note 229, at 684–85.

234 *See* Green, *supra* note 229, at 685. As a general rule, experiments containing larger statistical samples and studying phenomena with low background rates, or significant impacts, will have lower Type II error rates. A scientist, for example, studying breast cancer deaths associated with an industrial chemical drawing on a patient population of ten thousand individuals will be in a much better position to discern an effect than a scientist studying mild cognitive impairments from lead exposure with a patient population of one hundred individuals.

235 *See id.* at 687–89; Kaye, *supra* note 229, at 66–73.

236 The challenges of controlling statistical power are demonstrated by scientists' recent efforts to refocus attention on statistical power by undertaking post hoc power analyses, under which statistical power is calculated using the experimental data as an alternative to directly improving the statistical power of their experiments. While well intentioned, this approach is analytically flawed and logically inconsistent for reasons related to the interpretive problems discussed here. *See* John M. Hoenig & Dennis M. Heisey, *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis*, 55 AM. STATISTICIAN 19, 19–21 (2001).

in some situations underregulation poses the more serious risk of harm. Fortunately, the apparent bias of frequentist methods is neither necessary nor, as a historical matter, consistent with how significance testing was originally conceived. The statistician Jerzy Neyman, one of the co-developers of modern significance testing, addressed the importance and meaning of Type I and II errors in his 1950 introductory text on statistics:

> It is essential to notice *there are two different kinds of error possible.* The adoption of [the null] hypothesis H when it is false is an error qualitatively different from the error consisting of rejecting H when it is true. This distinction is very important because, with rare exceptions, the importance of the two errors is different, and this difference must be taken into consideration when selecting the appropriate test. . . .
>
> As already mentioned, the situation where the consequences of the two kinds of errors are of unequal importance is of a very general occurrence. It is true that in many cases the relative importance of the errors is a subjective matter . . . . However, this subjective element lies outside of the theory of statistics. The essential point to notice is that, in most cases, the person applying a test of a statistical hypothesis considers one of the possible errors more important to avoid than the other. . . .
>
> . . . .
>
> Postulating this to be the ordinary case we will use the expression *error of the first kind* [i.e., Type I error] to describe that particular error in testing hypotheses which is considered more important to avoid. The less important error will be called the *error of the second kind* [i.e., Type II error] . . . .
>
> . . . .
>
> This convention of labeling the two kinds of error is supplemented by a parallel convention concerning the use of the term *hypothesis tested.* Let H be a statistical hypothesis and H its negation. *The term hypothesis tested is attached to H or to H in such a way that the rejection of the hypothesis tested when it is true is an error of the first kind.*[237]

Neyman carefully distinguished Type I and II errors because, as discussed above, they cannot be jointly minimized. Accordingly, a judgment must be made regarding the appropriate tradeoff between the two types of error.[238] In significance testing, as Neyman indicates, Type I errors are minimized first, that is they are given priority. Neyman also made it clear, however, that the hypothesis tested and

237 NEYMAN, *supra* note 224, at 261–64.

238 *See supra* Part III.A.

Type I error are connected—minimizing the more important error requires that the appropriate hypothesis be tested.

Addressing Type II errors in environmental science therefore also entails formulating an appropriate null hypothesis to test.[239] In the standard significance tests, the null hypothesis is either that no effect exists or that an effect does not exceed a specific level, such as a regulatory limit.[240] In these cases, the null-hypothesis model is constructed by positioning a normal distribution at the value in question (i.e., zero or some other number). The Type I error then is the error of obtaining a positive result that is false (e.g., regulating a chemical that is nontoxic), which will result in the less important type of error being minimized, if one either accepts the Precautionary Principle generally or believes in the specific instance that underregulation poses greater risks.[241]

The bias of conventional frequentist significance testing is compounded by the common interpretive mistakes discussed earlier. Recall that significance testing supports one of two conclusions: either (1) the null hypothesis is false or (2) the null hypothesis is not inconsistent with observed experimental data—from which one generally *cannot* conclude that the null hypothesis is true.[242] Nevertheless, many people assume that failure to falsify the null hypothesis (i.e., lack of statistical significance) implies that no effect exists.[243] This interpretive error, in effect, places the burden of proof on anyone wishing to refute the null hypothesis.

Equivalence testing uses a null hypothesis that resolves both of these problems.[244] The typical null-hypothesis model of an experi-

239 Philip M. Dixon, *Assessing Effect and No Effect with Equivalence Tests*, *in* RISK ASSESSMENT: LOGIC AND MEASUREMENT 275, 275–76 (Michael C. Newman & Carl L. Strojan eds., 1998).

240 It is important to recognize that shifting the starting hypothesis to a nonzero value, such that some degree of harm is assumed at the outset, does not get you very far. In such cases, the test minimizes the error associated with finding, for example, the chemical does not have the specific nonzero value when in fact it does. The error minimized remains regulating when the nonzero harm does not actually exist, not failing to regulate when the chemical is harmful. If there is significant uncertainty about what the actual level is, minimizing the error associated with a discrete nonzero value is not terribly effective. To be effective, the null hypothesis needs to encompass a range of values all at once.

241 NEYMAN, *supra* note 224, at 275–76; Page, *supra* note 195, at 231–33.

242 *See supra* Part III.A.; Dixon, *supra* note 239, at 275–76.

243 Parkhurst, *supra* note 201, at 1053, 1055.

244 *See* Berger & Hsu, *supra* note 28, at 283–84; McBride, *supra* note 28, at 20–21; Parkhurst, *supra* note 201, at 1053–54. The test described here is also sometimes referred to as a "reverse equivalence test." *See id.*, at 1054–56.

ment, as discussed above, is based on a point estimate. Equivalence tests replace the point estimate with an interval. A zero-valued point estimate, for example, would be replaced by an interval of, say, magnitude 0.01, which would range from 0.00 to 0.01.[245] Just like a point estimate, an equivalence interval also can be used for nonzero values, either bracketing them, ± 0.05, or extending to one side, $x + 0.01$. The null hypothesis for an equivalence test is not "the chemical is toxic." It is "the chemical's toxicity is equal to or greater than x," where the interval is 0 to x and the value x is presumably set by a regulatory entity.[246] The conjectured hypothesis is "the chemical's toxicity is less than x."[247] Because the null hypothesis assumes that the chemical is harmful, equivalence tests minimize the "more important" error, which here is the error of declaring the chemical harmless when its toxicity is beyond the regulatory interval (i.e., erroneously determining that the chemical should not be regulated).[248] Similarly, the interpretive mistakes discussed above err in favor of protecting the environment and human health, which in this case is presumptively the more important direction to err.

An additional virtue of equivalence testing is that it is a well established statistical method under governing Food and Drug Administration (FDA) regulations.[249] Consistent with Neyman's reasoning, FDA requires equivalence testing to ensure that the risk of allowing a harmful drug to be sold is minimized, i.e., the more serious error is controlled. Accordingly, given that the FDA is one of the most highly regarded and scientifically sophisticated federal agencies, equivalence testing should not raise problems from either a scientific or regulatory standpoint. Moreover, while it is somewhat surprising that equivalence testing has not been used beyond the FDA, it does not derive from inherent limitations of the methodology, which could be applied in a broad range of environmental sciences.[250] Instead, it is likely that

245 *See* Dixon, *supra* note 239, at 276–77; McBride, *supra* note 28, 20–21.

246 *See* Berger & Hsu, *supra* note 28, at 283–84; Parkhurst, *supra* note 201, at 1054. The example is admittedly oversimplified insofar as it suggests that toxicity can be measured on a single metric. These complexities are not relevant here, as the central point of the example is independent of considerations about processes for quantifying the data.

247 McBride, *supra* note 28, at 20–21.

248 *Id.*

249 *See id.*; Dixon, *supra* note 239, at 279. The FDA requires generic drug manufacturers to use equivalence testing to determine whether a generic drug is bioequivalent to an existing brand-name drug. *See, e.g.*, FDA Bioavailability and Bioequivalence Requirements, 21 C.F.R. § 320 (2003).

250 Dixon, *supra* note 239, at 279; McBride, *supra* note 28, at 19–20, 23; Parkhurst, *supra* note 201, at 1054–56.

the arcane nature of statistical methods and general ignorance about them simply obscured the relevance of equivalence testing to other legal and regulatory areas.[251]

Despite these important virtues, some environmentalists may nevertheless object to the use of equivalence intervals.[252] Specifically, the interval from *0* to *x* described in the example above is, in effect, an interval in which the chemical's (nonzero) toxicity is determined to be de minimis.[253] If the toxicity of the chemical falls entirely within the equivalence interval, the null hypothesis for the equivalence test (i.e., that the chemical's toxicity is equal to or greater than *x*) is likely false and the chemical will be considered safe; otherwise, the test is inconclusive and the presumption remains that the chemical is harmful. The problem raised by the equivalence interval is that—like the convention of using a 5% significance level—no objective basis exists for determining its magnitude.[254] The size of the interval would presumably be set by the relevant agency, which is the current practice at the FDA.[255] For some environmentalists, the specter of allowing federal agencies to establish a priori de minimis levels for industrial chemicals will be grounds for rejecting the method, as de minimis levels are contrary to the chemical risk models propounded by environmentalists.[256]

Such opposition would not be warranted. First, the significance levels of traditional frequentist tests raise precisely the same problem, just less transparently. In fact, many people consider statistical significance levels to be defined objectively when they are set by convention. An equivalence interval, in contrast, would be established up front as a matter of agency policy, not under the guise of arcane statistical rules as significance levels are.[257] Second, and more importantly, traditional significance testing methods lack the benefit derived from shifting the de facto burden of proof to the regulated entity and minimizing the more environmentally significant type of error. Equivalence testing both rectifies the systemic bias in traditional significance

251 *See, e.g.*, Hoenig & Heisey, *supra* note 236, at 23; Parkhurst, *supra* note 201, at 1056–57.

252 *See* Dixon, *supra* note 239, at 279 ("All equivalence tests force the user to specify some region of equivalence before the data are analyzed.").

253 *See* McBride, *supra* note 28, at 21–26; Parkhurst, *supra* note 201, at 1054.

254 Dixon, *supra* note 239, at 279; Parkhurst, *supra* note 201, at 1054.

255 *See* Berger & Hsu, *supra* note 28, at 284.

256 *See* BREYER, *supra* note 20, at 44–45; Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1618–26.

257 Dixon, *supra* note 239, at 298.

testing while at the same time making the judgments and conventions in significance testing more transparent.[258]

Decisions regarding the use of equivalence tests over traditional methods will nevertheless remain contentious if the longstanding battle over the Precautionary Principle is at all representative. One can only hope that the added flexibility equivalence testing offers will allow this debate to evolve, as it will afford direct comparisons between traditional methods and a statistically valid alternative that is consistent with the Precautionary Principle.

C. *The Limits of Frequentist Methods and the Precautionary Principle*

Equivalence testing is ultimately only a partial response to the dictates of the Precautionary Principle. Frequentist methods support inferences from discrete scientific studies (the second stage of our framework) and are thus of limited value to integrated scientific determinations.[259] Consider an example in which the results of two experiments on a chemical's toxicity both satisfy a 95% significance level, but their estimates of its toxicity differ markedly. Assume further that one of the experiments involved dosing rats under controlled conditions, while the other was a human epidemiological study for which exposure levels could not be controlled as stringently. These experimental differences prove critical because the data are not directly comparable, i.e., they are not commensurable. Statistical significance will be irrelevant to how a scientist weighs the credibility of the two studies and integrates their results to estimate the chemical's toxicity. To make an integrated (stage three) determination, a scientist undertakes a qualitative assessment of how well each experiment was designed and implemented.[260] Accordingly, while statistical significance serves an important purpose, its role in rigorously testing hypotheses (i.e., gatekeeper) is removed from the final, third-stage scientific judgment.[261]

This simplified example is directly applicable to the EPA's process for setting chemical toxicity levels under its Integrated Risk Infor-

---

258 McBride, *supra* note 28, at 26.

259 *See* HACKING, *supra* note 33, at 111–13 (observing that frequentist statistical tests do not resolve the question of how stringent a test must be in a given context and thus cannot, on their own, be used to determine whether to reject or accept a given scientific theory); *see also* Green, *supra* note 229, at 693–94 (discussing problems with judges and juries limiting scientific analysis to "simple [statistical] screening devices").

260 *See* FOSTER & HUBER, *supra* note 82, at 33; MAYO, *supra* note 23, at 122–26; Collins, *supra* note 30, at 336–37.

261 MAYO, *supra* note 23, at 375–77.

mation System (IRIS) program.[262] IRIS toxicological reviews are designed to generate a consensus opinion on the potency of the toxic chemicals the EPA regulates. The IRIS process assesses all of the available toxicological studies performed on a chemical.[263] When integrating the available data to arrive at a consensus opinion, scientists consider a variety of experimental factors, such as whether the data are derived from animal or human studies, the degree to which the conditions for the experiments were controlled, assumptions made to determine exposure levels, and any confounding exposures that could bias the results.[264] Statistical significance is independent of these considerations—even poorly crafted or weak experiments can generate statistically significant results.[265] Thus, while a lower level of statistical significance may permit scientists to consider more data, it provides no guidance on the more important judgment of how the data are assessed relative to each other or as a whole.[266] This point is critical because scientific judgments on the value of specific experimental results "count most, not some meeting of, or failure to meet, an arbitrary level of statistical 'significance.'"[267]

The Precautionary Principle clearly is not limited to stage-two inferences from discrete experiments, or interpreted solely in terms of relative error rates and frequentist significance testing. Although it is often described in frequentist terms, the Precautionary Principle is targeted at scientific judgment generally.[268] Indeed, advocates of the Precautionary Principle consider its singular virtue to be that it is "imperfectly translatable into codes of conduct," and thus is resistant to

262 See Envtl. Prot. Agency, *What is Iris?*, *at* http://www.epa.gov/iris/intro.htm (last updated July 8, 2003), for the EPA's description of the IRIS program.

263 A chemical's "reference dose" is the highest dose for which its toxic effects are not observed, corrected for uncertainties in its derivation. The EPA uses potencies/reference doses and modeling methods to calculate regulatory standards for each of the chemicals it regulates. As such, the IRIS toxicological reviews provide the final toxicological information used by the EPA to calculate regulatory standards for toxic substances.

264 *See, e.g.*, BREYER, *supra* note 20, at 43–44; Green, *supra* note 229, at 649–53; Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1621–27.

265 *See supra* Part II.C (noting that the early Mendelian and biometrics experiments are just two examples).

266 *See* MAYO, *supra* note 23, at 313 n.8 (noting that the exclusion of nonsignificant results actually creates a bias in the scientific literature because negative results are often not reported and thus not considered in meta-analyses of multiple experimental studies); Collins, *supra* note 30, at 337.

267 *See* Collins, *supra* note 30, at 337.

268 *See* Barrett & Raffensperger, *supra* note 187, at 115–20; Jordan & O'Riordan, *supra* note 188, at 16–19.

expert co-option.[269] Formulated in this manner, however, the Precautionary Principle risks compromising legal and scientific procedures by treating obscurantism as a virtue necessary to counteract expert authority. The underlying premise is a familiar one, namely, that all "research priorities, data, and conclusions are shaped by social contexts and values."[270] In short, because environmental science is qualified by uncertainties and thus subject to value judgments, the Precautionary Principle should direct all scientific determinations.[271]

Probably the most common criticism of the Precautionary Principle is that it risks advancing a model for scientific inference that lacks both objective measures and quantitative clarity.[272] This vagueness is not, however, unique to the Precautionary Principle, but instead is a general feature of efforts to formulate interpretive principles based on broad fundamental principles or rights.[273] For conservative or procedurally oriented legal scholars, reliance on fundamental rights (e.g.,

269 Jordan & O'Riordan, *supra* note 188, at 15–16 (noting that the Precautionary Principle does not have "much coherence other than it is captured by the spirit that is challenging the authority of science, the hegemony of cost-benefit analysis, the powerlessness of victims of environmental abuse, and the unimplemented ethics of intrinsic natural rights and intergenerational equity").

270 Barrett & Raffensperger, *supra* note 187, at 116; *see also* R. Michael M'Gonigle, *The Political Economy of Precaution*, *in* PROTECTING PUBLIC HEALTH & THE ENVIRONMENT, *supra* note 85, at 123, 129–30.

271 The presumption that science is inseparable from social factors is a highly debatable one. *See, e.g.*, GASKINS, *supra* note 83, at 161–62; LAUDAN, *supra* note 30, at 104, 201–02. The environmentalists' argument is a non-sequitur. Environmentalists demonstrate the uncertainties in science and then employ these arguments to show that values must fill the gaps. The problem is that they never demonstrate that the choice is necessarily limited to either science or social values.

272 *See, e.g.*, Daniel Bodansky, *Scientific Uncertainty and the Precautionary Principle*, ENVIRONMENT, Sept. 1991, at 4–5 (asserting that "the precautionary principle . . . is too vague to serve as a regulatory standard"); Kenneth R. Foster et al., *Science and the Precautionary Principle*, 288 SCIENCE 979, 979 (2000) ("[The Precautionary Principle's] greatest problem, as a policy tool, is its extreme variability in interpretation."); Mark Geistfeld, *Reconciling Cost-Benefit Analysis With The Principle That Safety Matters More Than Money*, 76 N.Y.U. L. REV. 114, 174–76 (2001) ("The vagueness of the precautionary principle provides ample room for disagreement, making it hard to justify regulations based on the principle."); John Lemons et al., *The Precautionary Principle: Scientific Uncertainty and Type I and Type II Errors*, 2 FOUND. SCIENCE 207, 210 (1997) (claiming that the Precautionary Principle is not "concrete enough" to allow for consistent implementation); Sheila Jasanoff, *A Living Legacy: The Precautionary Ideal in American Law*, *in* PRECAUTION, ENVIRONMENTAL SCIENCE AND PREVENTATIVE PUBLIC POLICY 227, 229 (Joel A. Tickner ed., 2003) ("Critics charge not only that [the Precautionary Principle] is too vague to be useful, but also that it rejects science and threatens innovation.").

273 ELY, *supra* note 14, at 50 ("'[A]ll theories of natural law have a singular vagueness which is both an advantage and disadvantage in the application of the theories.'

privacy or equality) in judicial review exemplifies this kind of approach.[274] Objections to the Precautionary Principle do not differ in substance from those raised in the judicial context: the Precautionary Principle is used to guide scientific judgment just as fundamental rights are used to resolve interpretive ambiguities in constitutions and to guide judicial review generally.[275]

The deficiencies of a rights-based, or natural law, approach to judicial review have been enumerated many times. John Hart Ely provides one of the most deft and clear critiques:

> [T]he only propositions with a prayer of passing themselves off as "natural law" are those so uselessly vague that no one will notice—something along the "No one should needlessly inflict suffering" line. "[A]ll the many attempts to build a moral and political doctrine upon the conception of a universal human nature have failed . . . . [They] are too few and abstract to give content to the idea of the good, or they are too numerous and concrete to be truly universal. One has to choose between triviality and implausibility."[276]

The same uncertainties arise with the Precautionary Principle: "While it is applauded as a 'good thing,' no one is quite sure about what it really means, or how it might be implemented."[277] The challenges of applying the Precautionary Principle are in fact potentially more acute, as environmental policymaking is already rendered difficult by the technical nature of the underlying scientific determinations. Moreover, insofar as proponents of the Precautionary Principle accept as dogma that science is unavoidably infused with value judgments, the potential for science to resolve uncertainties will be undervalued or ignored.[278]

The advantage, one gathers, is that you can invoke natural law to support anything you want. The disadvantage is that everybody understands that.") (citation omitted).

274 *Id.* at 48–49.

275 *See* Jordan & O'Riordan, *supra* note 188, at 16 (characterizing the Precautionary Principle as implementing the "ethics of intrinsic natural rights and intergenerational equity"); Santillo et al., *supra* note 191, at 46 (noting that the Precautionary Principle is "an overarching principle to guide decision making in the absence of analytical or predictive certainty").

276 ELY, *supra* note 14, at 51–52 (citations omitted).

277 Jordan & O'Riordan, *supra* note 188, at 22 (noting that critics "claim its popularity derives from its vagueness").

278 *See* Barrett & Raffensperger, *supra* note 187, at 115 ("[R]esearch methods, theories, and empirical bases in ecology, as well as in more reductionist sciences, are underdetermined. As a result, isolated scientific disciplines cannot provide a strong basis for environmental policy."). See Part II.C for an opposing argument.

The problem with this critique is that it also applies to science. As Kuhn and others have shown, science consists of a mix of rigorous techniques and broad principles. Kuhn refers to the balance between them as "the essential tension" in good science.[279] These broad scientific principles (e.g., simplicity, consistency, and breadth) are not demonstrably more or less vague than the Precautionary Principle. Scientists, for example, seek to elaborate theories that are both internally consistent and consistent with existing data, but this ideal is fraught with uncertainties and ad hoc qualifications because no scientific theory is ever without contrary data.[280] Consistency thus becomes a matter of degree, but developing a coherent measure is complicated by the fact that competing theories will be consistent with different data. The different empirical support for competing theories makes it far more difficult to ascertain which of them is the "more consistent" because one ends up having to make judgments that amount to comparing apples and oranges. Objecting to the Precautionary Principle because of its vagueness is therefore self-defeating, for it implicitly condemns established scientific principles as well.

The basic sentiment behind the Precautionary Principle—consideration of the nature, uncertainties, and potential magnitude of the risks implicated in a scientific analysis—is not inherently antiscientific. Established scientific methods like statistical significance (and equivalence) testing, for example, contemplate a precautionary approach that considers the risks at issue in an experimental study.[281] Many advocates of the Precautionary Principle, however, have much more grandiose objectives, such as curing science of its reductionist bias and democratizing how science is practiced.[282] Indeed, some "strong con-

279 *See supra* Part II.B.

280 *See supra* Part II.B.

281 *See supra* Part III.B.

282 *See* Barrett & Raffensperger, *supra* note 187, at 115–17 ("In the precautionary model, scientists act as co-problem solvers in a broad community of peers. This community extends not only beyond the boundaries of individual disciplines but also beyond the traditional boundaries of the scientific community."); Joel A. Tickner, *The Role of Environmental Science in Precautionary Decision Making*, *in* PRECAUTION, ENVIRONMENTAL SCIENCE AND PREVENTATIVE PUBLIC POLICY, *supra* note 272, at 3, 16 ("To support precautionary decision making, the current fragmentation and narrow focus of science and policy will need to be dissolved, allowing a much broader framing and examination of questions."). It is worth remembering that Thomas Kuhn's theory of science was not an endorsement of scientific relativism. Kuhn's belief in science was grounded in the workings of normal science, which is an expert-community model, not a fully democratic one. *See supra* Part II.B. Kuhn understood that the singular virtue of science is that it sometimes does generate methods for objectively substantiating facts—despite universal theories remaining elusive. *Id.*

ceptions" of the Precautionary Principle restrict scientists to "a very limited role in decision making."[283] These stronger versions of the Precautionary Principle raise more difficult questions insofar as they propose radical departures from existing scientific methods and processes. The heavy ideological baggage that often attends the Precautionary Principle provides further grounds for being circumspect.[284] As in the balancing of scientific and political processes described in Part II.B, important tradeoffs exist between maintaining the integrity of scientific processes and addressing the objectives of these stronger versions of the Precautionary Principle.

The Precautionary Principle is a prominent example of how the public, lawyers, and scientists are struggling to define the appropriate scope of their respective roles in environmental policymaking. Part II of this Article focused on scientific methods and regulatory models relevant to stage-one quantitative judgments. This Part has discussed some of the limits of scientific methods by exploring important systemic biases and interpretive constraints found in frequentist statistical methods. I have proposed equivalence testing as a technical response to the bias of traditional frequentist methods, but it cannot address the subsequent stage-three judgments that ultimately must be made. Because significance testing does not quantify directly the probability that a hypothesis is valid, qualitative judgments—not quantitative assessments—of the support for a hypothesis must be made following a finding of statistical significance. The need for, and difficulty of making, these qualitative judgments is central to Bayesian criticisms of frequentist methods. Bayesian methods offer an alternative methodological approach to defining, and arguably broadening, the role of scientific expertise in environmental policy.

IV. Frequentist and Bayesian Methods: Competing Models for Integrating Expert Judgment Into Environmental Science

This final Part brings together the scientific and legal debates. It begins with a discussion of the opposing Bayesian and frequentist positions advocated by scientists, drawing initially on the models proposed by Breyer and Wagner, respectively. Part IV.A then provides a brief introduction to Bayesian methods, which is followed by a critique and more detailed comparison of frequentist and Bayesian methods in Part IV.B. The legal debate reenters the discussion in Part IV.C, where I propose a model for structuring scientific judgment that

283 Jordan & O'Riordan, *supra* note 188, at 25, 30–31.
284 *See id.*

exploits the parallels between the competing scientific and legal frameworks. However, the scientific and legal debates differ in one critical respect: the dispute among scientists is not divisible into the traditional antiregulatory and pro-environment political camps. Well known environmental scientists, for example, are ardent advocates of Bayesian methods, which complement the expert model of environmental decisionmaking advocated by Breyer and uniformly criticized by environmentalists. This undercutting of the existing political divisions is part of the appeal of linking the legal and scientific debates.

The dispute among scientists over Bayesian and frequentist methods can be understood through the lens of the proposals advocated by Justice Breyer and Professor Wagner discussed in Part II. Like the scientists debating statistical methods, both Breyer and Wagner are concerned about the proper role of scientific judgment in drawing inferences from limited scientific information. Breyer adopts a Bayesian approach by delegating broad authority to scientific experts for setting administrative standards and regulations. As described in Part I, Bayesian methods require that expert judgment be quantified and integrated directly into the analysis—the starting point for all Bayesian analyses is a judgment about the relative probability of the hypotheses being evaluated. Bayesians do not fixate on objective experimental methods, but opt instead to ground their methods on expertise and logical rules.[285] Like Breyer, Bayesians do not attempt to distinguish science from policy; Bayesian methods seamlessly mix judgment and experimental data together to derive the conditional probability (i.e., conditional on data and judgment) that a hypothesis is true.

Wagner's approach is frequentist in its procedural detail, commitment to transparency, and skepticism about expert judgment. Recall that the centerpiece of Wagner's proposal was requiring agency experts to distinguish the science from the trans-science, or science from policy judgments.[286] While frequentism does not distinguish between science and trans-science, it adopts an analogous approach by separating the formal statistical analysis of experimental data from scientific judgment, that is stages two and three of the framework discussed in the Introduction to this Article.[287] However, neither Wagner nor fre-

285 *See supra* Part I; *infra* Part IV.A.

286 *See supra* Part II.A.

287 Frequentism is not simply a collection of mathematical techniques for analyzing data; it incorporates a philosophy of scientific method similar to Popper's critical mode of inquiry. *See* MAYO, *supra* note 23, at 13–17; Dennis, *supra* note 23, at 1100. However, instead of developing severe tests for particular hypotheses, frequentists have adopted a set of experimental rules (e.g., randomization, control groups, 95%

quentism denies the importance of scientific judgment. Both instead adopt formal procedures for separating objective inferences from subjective judgments. Frequentism accomplishes this by quantifying how well a test hypothesis predicts experimentally observed results, which scientists use, along with a variety of qualitative factors, to judge the validity of their test hypothesis. Wagner, however, does not propose an analytic framework, but, somewhat circularly, relies on the judgment of scientific experts and judges to distinguish science from policy. Nevertheless, Wagner's proposal and frequentist methods share the same philosophy: Objective facts should be separated from subjective judgments to maximize the transparency and verifiability of the determinations.

The debate among environmental scientists over frequentist and Bayesian methods is hotly contested and intensifying. Frequentists object to Bayesian techniques because they do not adhere to frequentist conceptions of proper scientific methods. Frequentists warn that:

> [Environmental scientists] should be aware that Bayesian methods constitute a radically different way of doing science . . . . Bayesians categorically reject various tenets of statistics and the scientific method that are currently widely accepted in ecology and other sciences. The Bayesian approach has split the statistics world into warring factions . . . and it is fair to say that the Bayesian approach is growing rapidly in influence.[288]

Frequentists, as it turns out, have good reason to fear that Bayesian methods will become more important in environmental sci-

---

significance levels) to ensure that tests are severe and transparent. *See* MAYO, *supra* note 23, at 4–7, 12–13, 16–17. Frequentist experiments also are designed to minimize Popper's problem with auxiliary hypotheses by "isolat[ing] the effect of interest so that only a manageable number of causal factors (or types of factors) may produce the particular experimental outcome." *Id.* at 15. These rules provide standard methodological tools that ensure experiments are adequately controlled. Scientific objectivity is maximized by limiting scientific bias and making error rates a critical measure of scientific validity. Frequentists use statistical error as a basis for accepting experimental results "'because they are confident that *error ramifies.* If the hypotheses that they are accepting in order to attack new problems are mistaken, the results of related, though partially independent, research are likely to signal that something is wrong.'" *Id.* at 41–42 (quoting DAVID HULL, SCIENCE AS A PROCESS: AN EVOLUTIONARY ACCOUNT OF THE SOCIAL AND CONCEPTUAL DEVELOPMENT OF SCIENCE (1988)). Frequentists seek to minimize bias by "remov[ing] the scientist's beliefs from the conclusions as much as possible and let[ting] the data do the talking." Dennis, *supra* note 23, at 1100.

288 Dennis, *supra* note 23, at 1095.

ence.[289] It was not until new computational methods became available in the late 1980s that Bayesian methods began to receive significant attention from environmental scientists,[290] and packaged programs made available through new computational methods are making Bayesian methods increasingly accessible.[291]

Bayesian advocates are no less committed to their approach and its significance to environmental science.[292] They argue that it is absurd to assume that policymakers will have the expertise necessary to interpret frequentist statistics, which require a detailed understanding of the experimental methods and science to determine the weight experimental findings should be accorded.[293] A frequentist working on global change research, for example, would be limited to supplying policymakers with statistically significant results on one or more scenarios for global warming, each of which would be supported by a multiplicity of experimental and observational data. Moreover, pursuant to the frequentist philosophy of letting the data tell the story, the scientist would not indicate which of the scenarios he or she believes is the most probable; this judgment would be left to the policymaker. In contrast, a Bayesian analysis would generate relative probabilities for each of the scenarios based on expert judgment and the available data. Policymakers could inquire about and challenge the methods used and bases for the scientific judgments, but would not be required to interpret the data de novo.

Bayesians challenge the viability of frequentist methods in environmental science and laud the clarity of Bayesian results:

> Serious doubts have been raised about the utility of abstract, general theories that have been shown repeatedly to have little predictive value in field or laboratory situations. . . . Bayesian statistical inference can be used to estimate ecologically meaningful parameters and provides an explicit expression of the amount of uncertainty in these parameter estimates.[294]

The essence of this debate turns on two practical concerns: the credibility of science, which frequentist fear is compromised by Baye-

289 *See* Malakoff, *supra* note 23, at 1460–61 (noting that using the Bayesian method enables researchers to achieve results not easily obtainable using the frequentist method).

290 Brian Dennis, *Statistics and the Scientific Method in Ecology*, *in* THE NATURE OF SCIENTIFIC EVIDENCE (M.L. Taper & S.R. Lele eds., forthcoming 2004) (manuscript at 20, on file with author).

291 Malakoff, *supra* note 23, at 1460–61.

292 *See* Schneider, *supra* note 23, at 18.

293 *See id.* at 18–19.

294 Ellison, *supra* note 23, at 1036–37.

sian methods;[295] and the ability of environmental science to provide meaningful information for policymakers, which Bayesians fear frequentist methods preclude.[296] It also involves two competing models of science: the Bayesian model structured around preliminary scientific judgments and logical rules, and the frequentist approach premised on rigorous testing and an objectivist approach to statistical analysis.

Bayesians have encountered significant opposition from environmental scientists who believe that Bayesian methods threaten the objectivity of environmental science by incorporating subjective judgments. The issue at the center of this dispute is the normative authority of environmental science, particularly as it pertains to environmental policymaking. Frequentists believe the authority of environmental science depends on its objective credentials; Bayesians assume that it is derivative of the good judgment of scientists themselves. To urge the acceptance of Bayesian methods in environmental science, Bayesians rely on two central arguments: (1) Bayesian methods generate objective results because scientists' estimates converge as more evidence is collected, and (2) frequentist methods are ill suited to the needs and constraints of environmental science and policy. These arguments will be discussed following a brief explanation of Bayesian methods. This Part concludes by proposing a new model for effectively integrating scientific judgment into environmental policymaking.

A. Bayesian Statistical Inference in Practice

The foundational principle of Bayesian, or belief-type, probability is that information should be incorporated into decisionmaking pursuant to the logic of Bayes's theorem.[297] Reverend Thomas Bayes's approach to statistical inference is particularly novel because it can be applied even when one begins with no information about the probable outcomes of a decision. If I have several competing hypotheses about the toxicity of a chemical (e.g., the chemical is benign, harmful only at high doses, or harmful at any level of exposure) but no starting data, Bayes postulated, in effect, a Cartesian starting point from which

295 *See* Dennis, *supra* note 23, at 1099–100 ("I object to calling [Bayesian estimates] science. Science is not about decisions; science is about making convincing conclusions. . . . [T]he Bayesian philosophy of science is scientific relativism.").

296 Ellison, *supra* note 23, at 1036–38.

297 HACKING, *supra* note 12, at 172–77. The principal question for Bayesian analysis "is whether we are reasonable in modifying these opinions in the light of new experience, new evidence." *Id.* at 256–57.

to incorporate inductively new information and to infer logically which of the competing hypotheses is most probable.[298] Bayesian theorists view probability in two distinct ways, logically and subjectively.[299] The two theories will be treated together because they draw on the same principles and methods.

A crucial aspect of Bayesian methods is the initial scientific judgment from which the analysis starts. This judgment reflects a scientist's understanding of the system being studied and must be reduced to a "prior distribution," which represents the distribution of probabilities assigned to the set of hypotheses or events being considered.[300] The prior distribution, for example, of a coin assumed to be

---

298 It should also be noted that frequency-type information is routinely used as a basis for belief-type probability. *Id.* at 137.

299 *Id.* at 146–47. Logical belief-type probability is defined objectively in terms of logical relations between evidence and propositions, such that a proposition's probability is evaluated relative to the available evidence. *Id.* at 142–43; Ian Hacking, *The Theory of Probable Inference: Neyman, Peirce and Braithewaite, in* SCIENCE, BELIEF AND BEHAVIOUR 141, 142–44 (D.H. Mellor ed., 1980). Restated more formally, "any body of evidence *e* uniquely determines a probability for any hypothesis *h*," such that *c(h,e)* = degree to which *e* confirms *h*. HACKING, *supra* note 32, at 148. This approach has a distinctly legal character, having been variously described as the degree of reasonable belief in or credibility of a proposition conditioned on the available evidence. *See* HACKING, *supra* note 33, at 190–91, 194, 202; HACKING, *supra* note 32, at 13, 43, 85–87, 134–36.

By contrast, subjective belief-type theorists define probability as a subjective estimate of the likelihood that an event will occur or theory is true. *See* HACKING, *supra* note 33, at 148, 208–09, 213–14. This radically subjective approach presumes that individuals arrive at personal estimates on whatever bases they choose. *See id.* at 190, 194. Subjectivism denies (in a variant of idealism) that probability exists as a property independent of individual subjective consciousness, and in so doing purports to avoid Hume's induction problem. HACKING, *supra* note 12, at 256–60; MUSGRAVE, *supra* note 116, at 146–47. Accordingly, where the logical theory turns on evidentiary support, the subjective theory considers "betting rates" (i.e., individual judgments constitute the relevant data) to resolve personal estimates regarding an event or hypothesis. *See* HACKING, *supra* note 33, at 191–93, 202. If a person is "rational," his betting rates will satisfy the logical axioms of belief-type probability theory. *See id.* at 208–09. The central rule is the following: "The *probability* of any event is the ratio between the value at which an expectation depending upon the happening of the event ought to be computed, and the value of the thing expected upon its happening." *Id.* at 192. The probability, *P(*)*, of an event is defined in the following terms: *P(*) = [contingent expectation value of a thing ("fair stake")] / [value of the thing expected when it occurs ("prize")]*. *Id.* at 192–94. The *"fair betting rate" = fair stake / prize = P(E)*, where *E* is a contingent event upon which the expectation is based. *Id.* Subjectivists assume that a fair betting rate is independent of the prize's size—i.e., that the relations are mathematically linear. *See id.* at 96–97.

300 HACKING, *supra* note 12, at 173–74. A prior distribution may involve two hypotheses or events (e.g., a chemical is either harmful or not) or a number of them

fair would distribute the total probability evenly between heads and tails, i.e., 50% for a toss coming up heads and 50% for tails. A prior distribution also may be a continuous function and can take on a wide variety of mathematical forms—although, more complex distributions may complicate Bayesian calculations. The prior distribution, for instance, of a chemical's toxicity could be a simple normal distribution centered around the toxicity level a scientist believes is the most probable. Where a prior distribution is well circumscribed by existing scientific knowledge, judgments about it will likely be relatively consistent and uncontested. As a practical matter, however, scientists often differ substantially in their judgments (influenced by personal biases or specific experiences) regarding the appropriate prior distribution for a specific system.[301] The discussion that follows begins with a brief explanation of Bayes's theorem and then turns to Bayes's postulate and the strengths and limitations of the Bayesian approach.

1. Derivation and Logic of Bayes's Theorem

Bayes's theorem is the unifying logical formula that governs Bayesian statistical inference.[302] The derivation of Bayes's theorem follows directly from the three fundamental axioms of probability theory and the definition of "conditional probability."[303] For Bayes's theorem to

---

(e.g., a chemical is benign, it has a threshold below which it is not harmful, it is harmful and harm increases linearly with exposure, or it is harmful and harm increases exponentially with exposure) depending on the circumstances.

301 *See* MAYO, *supra* note 23, at 75–77, 119–20. For example, a toxicologist working for a pesticide manufacturer is likely to have a very different understanding of an industrial chemical's biological activity than a toxicologist working for an environmental organization, and this understanding will be reflected in their judgment about a prior distribution.

302 *See* HACKING, *supra* note 33, at 190.

303 *See supra* note 33. Conditional probability is the probability of an event conditioned on another event occurring, such as the likelihood of rain under certain atmospheric conditions. HACKING, *supra* note 33, at 14, 59. The probability of A occurring conditioned on the occurrence of B, that is $P(A / B)$, is defined to be $P(A\&B) / P(B)$. *See id.* at 127. Conditional probability satisfies Kolmogoroff's first axiom, $0 \leq P(A / B) \leq 1$, because the probability of B and A occurring cannot exceed the probability of B occurring on its own. The conditional probability of A relative to B, $P(A / B)$, is thus defined as the probability of A and B occurring together, "normalized" (i.e., transformed onto a scale from 0 to 1) relative to the probability of B occurring on its own.

The definition of conditional probability also implies that $P(A) = P(B)P(A / B) + P(\sim B)P(A / \sim B)$, where $\sim B$ is defined as B does not occur. For hypotheses A and B that are mutually exclusive and exhaustive, the derivation is as follows: First, it follows from A and B being mutually exclusive and exhaustive that $P(A) = P(A\&B) + P(A + \sim B)$—all cases of A occurring are covered by circumstance under which either B oc-

apply, however, the hypotheses being evaluated must satisfy two conditions: first, they must be mutually exclusive, meaning only one hypothesis can be true or only one event may occur; second, they must be jointly exhaustive, meaning one of the hypotheses must be true or one event must occur.[304] Bayes's theorem states that the probability of a hypothesis *H*, such as it will rain today, conditioned on an event *E*, such as news of diminishing barometric pressure, is simply the probability of *H* prior to the new information, *P(H)*, multiplied by the "likelihood" of observing *E* if *H* is true.[305] Bayes's theorem has the following general form:

$P(H_j / E) = P(H_j)P(E / H_j) / [\Sigma P(H_i)P(E / H_i)]$, for any set of mutually exclusive and exhaustive hypotheses $H_1, H_2, H_3, \ldots, H_k$, where $P(H_i) > 0$ for each i and $P(E) > 0$.[306]

The logical appeal and elegance of Bayes's theorem are apparent in the following standard example.[307] Assume you have been selected to be a juror in a case in which the plaintiff's car was sideswiped by another car on a misty winter night. The sole witness to the collision claims she observed a blue cab collide with the plaintiff's car and then drive off. You also learn that there are only two taxi companies in the town, Green Cabs, Ltd., which owns only green cabs and operates 85% of the cabs in town, and Blue Taxi, Inc., which owns only blue cabs and operates 15% of the cabs in town. Finally, you are informed that the witness selected the correct color of car, whether green or blue, 80% of the time when tested under conditions similar to those during

---

curs or *B* does not occur. Using the definition of conditional probability, substitute $P(A / B)P(B)$ and $P(A / \sim B)P(\sim B)$ for $P(A\&B)$ and $P(A + \sim B)$, respectively, to obtain $P(A) = P(B)P(A / B) + P(\sim B)P(A / \sim B)$, also known as the rule of total probability. For any mutually exclusive and exhaustive set of hypothesis $H_1, H_2, H_3, \ldots, H_k$ for which $P(H_i) > 0$ for each i, this rule generalizes as follows: $P(H_i) = \Sigma P(H_i)P(E / H_i)$.

304 HACKING, *supra* note 12, at 70.

305 *See* HACKING, *supra* note 33, at 190–91.

306 EARMAN, *supra* note 31, at 33–35. The derivation is relatively straightforward: Starting with $P(H + E) = P(E + H)$, multiply the left side by $P(E) / P(E)$ and the right side by $P(H) / P(H)$ to obtain $P(H + E)P(E) / P(E) = P(E + H) P(H) / P(H)$. Next, using the definition of conditional probability, substitute $P(H / E)$ for $P(H + E) / P(E)$ on the left side and $P(E / H)$ for $P(E + H) / P(H)$ on the right side to obtain $P(H / E)P(E) = P(E / H)P(E)$, which becomes $P(H / E) = P(E / H)P(E) / P(E)$. Finally, using the rule of total probability substitute $P(H)P(E / H) + P(\sim H)P(E / \sim H)$ for $P(E)$ to obtain Bayes's theorem for the simplest system of a binary set of exhaustive hypothesis (e.g., a coin is fair or unfair): $P(H / E) = P(E / H)P(E) / [P(H)P(E / H) + P(\sim H)P(E / \sim H)]$. Bayes's theorem generalizes as follows: $P(H_j / E) = P(H_j)P(E / H_j) / [\Sigma P(H_i)P(E / H_i)]$, for any set of mutually exclusive and exhaustive hypotheses $H_1, H_2, H_3, \ldots, H_k$ for which $P(H_i) > 0$ for each i and where $P(E) > 0$. HACKING, *supra* note 12, at 58–62, 70.

307 HACKING, *supra* note 12, at 72–73.

the night of the accident. As a juror, you want to determine the probability that the witness's account of the collision is correct.

Bayes's theorem may be used to integrate this information. Assuming that the car involved was a cab, the mutually exclusive and exhaustive set of scenarios (hypotheses) are that the cab was blue, *B*, or that the cab was green, *G*. In this example, you are given the prior distribution, namely, the relative likelihoods of randomly encountering a green cab, *P(G)*, and a blue cab, *P(B)*, which are 0.85 and 0.15, respectively. In addition, you know the probability that the witness's observation is correct, $P(W_b / B)$, is 0.80 (implying that $P(W_b / G) = 0.20$), where W_b identifies the witness as observing a blue cab. Bayes's theorem may be used to determine the probability of the cab being blue based on the witness's statement as follows:

$$P(B / W_b) = P(B)P(W_b / B) / [P(B)P(W_b / B) + P(G)P(W_b / G)]$$
$$(0.15 \times 0.8) / [(0.15 \times 0.8) + (0.85 \times 0.2)] = 0.41$$
$$P(B / W_b) = 0.41$$
$$P(G / W_b) = 1 - P(B / W_b) = 1 - 0.41 = 0.59$$

The determination that it is more likely that the cab is green (59% probability) than blue (41% probability) is counterintuitive for many people. The result, however, follows directly from the fact that 85% of the cabs on the road are green. The significance of the relative numbers of blue and green cabs becomes evident if you consider it in the context of testing the witness's observational accuracy. Assume that this testing required the witness to make one hundred observations. Based on the relative numbers of green and blue cabs, approximately eighty-five of the observed cabs are green and fifteen blue. The witness's observations, however, are correct only 80% of the time, implying that of the eighty-five green cabs, she observes about sixty-eight as green and about seventeen as blue, and of the fifteen blue cabs, she observes about twelve as blue and about three as green. Adding these results together, the witness therefore observes twenty-nine cabs as blue, but only twelve (or about 41%) of the cabs are in fact blue. In this example, the relative numbers of the blue and green cabs offset the relatively high accuracy of the witness.

The cab example involves a very simple system for application of Bayes's theorem. It satisfies precisely the basic requirements for Bayes's theorem to hold: the potential options are mutually exclusive, cabs are blue or green (not both), and exhaustive, each cab is either blue or green (not some other color).[308] Perhaps most importantly,

308 This analysis implicitly rejects (i.e., sets their probabilities at zero) other hypotheses, such as that the defendant's car was not a cab, thereby irreversibly excluding them from the analysis.

the prior distribution is provided by the data, obviating any need for judgment by the juror. Few, if any, real world situations neatly meet these conditions.[309] In particular, as Popper demonstrated, scientists cannot be certain that their set of hypotheses is complete.[310] These epistemic limits necessitate that a number of conventional rules be employed to resolve uncertainties and to make Bayes's theorem workable.[311] The next subsection discusses Bayes's postulate and the limits of Bayesian analysis when confronted with such practical constraints.

2. Bayes's Postulate and the Limits of Bayesian Analysis

According to Bayes's postulate, Bayes's theorem may be used to incorporate experimental evidence inductively and to infer the relative probabilities of any set of mutually exclusive and exhaustive hypotheses or events.[312] When starting with near or complete ignorance, Bayes determined that a uniform prior distribution, meaning all hypotheses or events are given the same starting probability, should be used.[313] This strategy is analogous to assigning a one-sixth probability to each face of a die when no basis exists to believe that the die favors one side over another. For a variety of reasons, Bayesians are criticized for their reliance on such uniform distributions as a justifiable starting point for Bayesian analysis.[314] More importantly, even where information exists, scientists often have divergent views on what the prior distribution should be. Take the example of a Bayesian analysis of contaminant levels in a river. An environmentalist would likely adopt a precautionary approach by assuming, for instance, that

309 EARMAN, *supra* note 31, at 140–41, 148–49.

310 *See supra* Part II.B.; HACKING, *supra* note 33, at 223–25. Moreover, for subjectivists, it is not at all clear what it means to "bet" on a hypothesis, which in all but the most trivial cases will not be conclusively verifiable. *See id.* at 215.

311 Among the most important conventions is the assumption of linear scaling, which ignores certain indeterminacies in choosing different scales and runs contrary to economic and sociological data. *See* EARMAN, *supra* note 31, at 17; HACKING, *supra* note 33, at 171–72, 199–201, 210; MAYO, *supra* note 23, at 90.

312 It is worth noting that, despite the dubious status of positivism, Bayes's postulate is based on a positivist approach to scientific progress. *See* EARMAN, *supra* note 31, at 63–64.

313 *See* HACKING, *supra* note 12, at 70, 141–44; HACKING, *supra* note 33, at 200–06.

314 *See* HACKING, *supra* note 33, at 202–04, 208–10; MAYO, *supra* note 23, at 75–76, 83–85. A common criticism, for example, is that such uniform distributions are arbitrary in their choice of how the data are scaled, such that a uniform distribution in one coordinate system or scale may be highly nonuniform in a different one. *See* HACKING, *supra* note 33, at 202–04, 208. Certain theorists, most notably Ronald A. Fisher, deplored Bayesians' overreliance on starting uniform distributions. *See* MACKENZIE, *supra* note 160, at 208–10.

the contaminants do not biodegrade rapidly, that they are not sequestered, and that river flow dynamics cause them to concentrate in certain areas. Industry scientists, in contrast, would likely begin with a very different, less conservative set of assumptions to construct their prior distributions. These kinds of subjective judgments are intrinsic to Bayesian methods, and they cause scientists with different perspectives to arrive at different starting prior distributions.

Bayesians have a simple response to such concerns: The starting distribution is irrelevant because Bayesian analyses converge to the most probable hypothesis irrespective of the starting distribution.[315] In its fully idealized form, Bayes's theorem, according to its proponents, operates as follows:

> It is of fundamental importance to any deep appreciation of the Bayesian viewpoint to realize that the particular form of the prior distribution expressing beliefs held before the experiment is conducted is not a crucial matter. . . . For the Bayesian, concerned as he is to deal with the real world of ordinary and scientific experience, the existence of a systematic method for reaching agreement is important. . . . The well-designed experiment is one that will swamp divergent prior distributions with the clarity and sharpness of its results, and thereby render insignificant the diversity of prior opinion.[316]

Bayesians in essence claim that given sufficient data, Bayes's theorem will produce objective results that are independent of initial estimates of the probabilities for a set of starting hypotheses. It achieves this objectivity by causing expert opinion to converge as more data are collected, meaning at some point Bayesian assessments of any group of experts would derive the same most-probable hypothesis. Thus, an environmentalist and industry scientist may start with divergent estimates of contaminant levels, but once sufficient data are collected, their estimates will converge to the same value.

A relatively simple example of how Bayesian analysis is applied in environmental science suggests that such convergence is far from guaranteed and that prior distributions generally cannot be ignored. Assume that a river is found to have been contaminated with copper from an industrial plant and that ten samples with a mean value of 50.6 microgram/liter (μg/l) and variance of 25.0 μg/l have been col-

315 EARMAN, *supra* note 31, at 57–58, 141–42.

316 Patrick Suppes, *A Bayesian Approach to the Paradoxes of the Ravens*, *in* ASPECTS OF INDUCTIVE LOGIC 198, 204 (Jaakko Hintikka ed., 1966); *see also* FOSTER & HUBER, *supra* note 82, at 121–24 ("In Bayesian terms: Prior probabilities (that is, initial guesses) play smaller and smaller roles as new evidence accumulates.").

lected to test the concentration of copper in the water.[317] Bayesian methods can be used to determine the probability that the mean concentration of copper in the river exceeds a federal regulatory standard of, say, 46 μg/l.

To begin the analysis, the scientist, whom we will assume is an employee of the plant, would specify a prior distribution delineating the probabilities of a continuous range of mean copper concentrations in the river.[318] For simplicity, we will assume that the scientist judges that the prior distribution, *P(H)* above, is a normal distribution with a variance of 4 μg/l centered around a mean copper concentration of 20 μg/l (i.e., a low estimate with a narrow spread of values).[319] The scientist would then use Bayes's theorem to combine the sample data with his prior distribution, which here would generate a most probable mean concentration of ~36.1 μg/l.[320] Analogous to the cab example, the prior distribution in this case offsets the much higher observed values and leads to a prediction that the federal standard is not violated.[321] Further, if another scientist were to start with a prior distribution much closer to the sample value, convergence of the two estimates would not be attained even with a significant increase in the quantity of data—convergence is purely speculative.[322]

The water sampling example demonstrates that where a disparity exists between a scientist's prior distribution and observed values, which may occur where data are of unknown quality, the prior distribution will dominate the final result if data are limited. Accordingly, the claim that well designed experiments will ensure that divergent assessments of competing hypotheses will converge appears empty. It

317 Dennis, *supra* note 290 (manuscript at 6–7).

318 *Id.* (manuscript at 15–16). The prior distribution need not be a normal distribution; its use here is solely to simplify the example.

319 *Id.* (manuscript at 15). Unlike the cab example, the prior distribution in this case is continuous across a range of potential contaminant levels, such that each contaminant level encompassed by the normal distribution constitutes a hypothesis for the likely concentration of copper in the river and is given a discrete probability. While mathematically more complex, the basic logic for applying Bayes's theorem is the same.

320 *Id.* (manuscript at 18). In this example, the expected value of the posterior distribution was used for the predicted value.

321 For comparison, the equivalent frequentist estimate (i.e., using a null (normal) distribution centered at 46 μg/l) would generate a 95% confidence interval that ranged from 47 to 54 μg/l. *Id.* From this result, a frequentist would conclude that the null hypothesis is false and that one can infer that the regulatory standard is being violated.

322 In the paper from which this example was taken, the author estimates that a Bayesian would have to collect more than six times the number of measurements before concluding that the regulatory standard was violated. *Id.* (manuscript at 23).

assumes, among other things, that such experiments exist and will be accepted without controversy, as well as that all concerned will agree on the hypotheses that should be considered. These constraints markedly qualify the power of Bayesian analysis to provide objective probability estimates for statistical inference.[323] Part IV.B examines these limitations in greater detail.

B. Bayesian and Frequentist Methods in Environment Policy

Bayes's theorem, according to its proponents, provides a rigorously logical technique for attaining agreement among scientists on the most probable of several competing hypotheses (e.g., the toxicity of a chemical for a range of exposure levels).[324] An obvious question is *how rapidly* Bayesian analysis causes the opinions of scientists to converge; in particular, if a large amount of data is required, agreement may be foreclosed as a practical matter. Several requirements must be met in order for Bayesian analysis to converge, but just two, probability assignment and hypothesis selection, will be discussed here.[325] For purposes of this discussion, description of the other requirements is unnecessary because the two selected factors amply illustrate the practical limitations of Bayesian methods.

The first requirement is that scientists assign a probability to each of the hypotheses being evaluated.[326] While this judgment may be relatively straightforward for simple experimental systems, such as a series of coin-flipping experiments, it becomes much more variable and complex for nonuniform experimental information:

> It is not just that different Bayesian agents will give different estimates of rates of convergence but that there may be no useful way to form the estimates. To form an estimate for a given [case] we need to know what kind of evidence is received and also what bits are received in what order . . . . [I]n the general case, the relevant evidence can come in myriad forms, and within a form the order [in which it is obtained] can matter critically.[327]

To understand why this is the case, it is helpful to consider why simple experimental systems are distinctive. In a simple stochastic ex-

---

323 Examples include uncertainties in the additivity of probabilities and the limits of strict conditionalism. *See* EARMAN, *supra* note 31, at 41; FOSTER & HUBER, *supra* note 82, at 129–30. Note that for subjective Bayesians this is not a problem, as they do not believe that objective probabilities can be obtained. *See supra* note 299.

324 Suppes, *supra* note 316, at 204.

325 EARMAN, *supra* note 31, at 139–42, 148–49.

326 *Id.* at 143, 148; B. Efron, *Why Isn't Everyone a Bayesian?*, 40 AM. STATISTICIAN 1, 2 (1986).

327 EARMAN, *supra* note 31, at 148–49.

periment, the only information that is relevant is the final tally of the experimental observations, e.g., the number of heads versus tails for a coin or the relative number of green and blue taxis in the earlier cab example.[328]

Most experimental data are not uniformly equivalent and thus cannot be simply added together like a sequence of coin-flips, which are simply reported as either heads or tails. In the case of a chemical risk assessment, all available toxicological studies must be evaluated, but few (if any) of them will have been conducted under identical conditions. Experimental parameters will vary, including study type (e.g., animal versus human, in vivo versus in vitro), exposure levels and controls, and confounding variables that could bias the results.[329] For example, some data may derive from epidemiological studies of exposed workers for which exposure conditions were poorly controlled and inaccurately known; other testing may involve carefully controlled animal studies, but the animal model may not be well justified.[330] These differences preclude data from distinct experiments being quantitatively inter-translatable. As a result, interpretation of data from multiple experiments requires scientists to judge how data from each experiment are to be weighted and integrated. Data will therefore influence scientists' opinions to different degrees and in ways that depend on other information.[331] Consequently, because many different paths will exist for information to be obtained, this interpretive interdependence makes it impossible to predict when Bayesian convergence of opinion will occur.

The second requirement for convergence is that scientists be "equally dogmatic" at the outset, that is assign zero probability to (i.e., reject) the same candidate hypotheses.[332] Such agreement might be attained by establishing "a rule of mutual respect that enjoins members of a scientific community to accord a nonzero [probability] to any hypothesis seriously proposed by a member of the community."[333] Indeed, this approach might be viable with a small group of scientists

328 *See supra* Part IV.A.1.

329 *See, e.g.*, Green, *supra* note 229, at 649–53; Wagner, *Toxic Risk Regulation*, *supra* note 5, at 1621–27.

330 Experimental variability also is significant in climate change science, where scientists must rely on climate measurements taken under a wide range of conditions and using a broad variety of techniques. HARVEY, *supra* note 149, at 76–77; IPCC, *supra* note 183, at 249.

331 EARMAN, *supra* note 31, at 56, 149, 151–53.

332 *Id.* at 142. Requiring equal dogmatism is also clearly contrary to Popper's critical mode. *Id.* at 139, 148–49, 160; HACKING, *supra* note 12, at 181, 257; *see supra* Part II.B.

333 EARMAN, *supra* note 31, at 142.

working on an esoteric problem. It is difficult to imagine, however, that one could convince the community of toxicologists, including those working in industry, government, and nonprofits, to agree in principle on such an approach. Industry scientists, for example, might reject (i.e., assign zero probability to) a hypothesis that chemicals are harmful even at extremely low levels, whereas environmentalists would rule out the hormesis hypothesis, which maintains that low-level exposure to industrial chemicals can be beneficial. Bayesian analysis requires that such differences be resolved up front. Moreover, even if it were possible to achieve such an agreement, you would still encounter widely divergent interpretations of the available data, which would make the likelihood of convergence of opinion at best speculative. The rigor of Bayes's theorem is therefore achieved at a significant price, for it requires that all of the relevant theories be fully elaborated and agreed to at the start.[334]

The limited power of Bayes's theorem to harmonize divergent opinions magnifies the significance of a scientist's starting prior distribution in Bayesian analyses.[335] The water-sampling example presented in Part IV.A.2 illustrated this problem. The industry scientist's prior distribution in that example lowered the estimated mean contaminant level by almost 29% relative to the value for the collected data.[336] Scientists' subjectively derived prior distributions are therefore bound to influence greatly, if not determine, the outcome of Bayesian assessments in fields like environmental science where data are often limited.[337] Drawing on Thomas Kuhn's work, a charitable assessment of Bayesian methods might be that they have significant value for normal science; however, even normal science generates conflicting hypotheses and divergent opinions among scientists.[338] It is difficult not to conclude that the failed objectivity of Bayesian methods seriously undermines their utility.

Undeterred, Bayesians are quick to point out that frequentist methods have important drawbacks, too. In particular, frequentist methods can be ill suited to environmental science:

> From an ecological perspective, there are many difficulties with [frequentism]. Within experiments, true randomization is difficult,

334 *Id.* at 123–25; Efron, *supra* note 326, at 2.

335 EARMAN, *supra* note 31, at 138.

336 *See supra* Part IV.A.2.

337 EARMAN, *supra* note 31, at 161. Some theorists have argued that this substantive bias ought to be acknowledged up front and that debate over and development of potential hypotheses should be integrated into the Bayesian program. *See id.* at 182–85, 198–203.

338 *Id.* at 172–73, 198.

> replication is often small, misidentified, or by virtue of circumstance, nonexistent. Ecological experiments rarely are repeated independently. No two organisms are exactly alike, and consequently they are unlikely to respond to our treatment in exactly the same way. Evolution virtually guarantees that even if they were alike today, their offspring will be measurably different. Thus, the idea that there is a true, fixed value [i.e., long-run frequency] for any ecologically meaningful statistical parameter is a Platonic phantom.[339]

These problems undoubtedly pose critical dilemmas for how frequentist methods are applied and interpreted. It is less clear, though, whether they justify abandoning frequentist methods for Bayesian analysis. Bayesians typically make arguments about the logical rigor of Bayesian analysis, as well as its ability to handle many types or even small amounts of data.[340] While it is difficult to argue against logic, this virtue is not realized without its attendant costs. Bayesian logic requires that candidate hypotheses be mutually exclusive and exhaustive and that scientists agree on the hypotheses that will be considered. These requirements are problematic for environmental science, which is often speculative and hotly contested. Furthermore, the value of rigorous inductive logic under these circumstances is less compelling, as it risks meaning little more than being consistently wrong. Simple deductive or inductive logic on its own, as Popper and Kuhn showed, does not guarantee good science.[341]

Bayesian environmental scientists argue most passionately about the power of Bayesian methods to generate clear quantitative data, which they believe are essential if environmental science is to be used effectively in policymaking.[342] Bayesian advocates, such as Dr. Stephen Schneider (mentioned in the Introduction) argue "that policy analysts need probability estimates to assess the seriousness of the implied [environmental] impacts; otherwise they would be left to work out the implicit probability assignments for themselves."[343] Bayesians claim, and apparently believe, that policymakers should concur, that it is better "to put more trust in the probability estimates of [environmental scientists]—however subjective—than those of the myriad special interests that have been encouraged to make their own [predictions]."[344] For many environmental scientists, such claims ignore the risk that Bayesian methods will further politicize environ-

339 Ellison, *supra* note 23, at 1037 (citations omitted).
340 *Id.* at 1037–38, 1043.
341 *See supra* Part II.B.
342 *See* Schneider, *supra* note 23, at 17–18.
343 *Id.*
344 *Id.* at 19.

mental science and damage its already often-tenuous credibility. Bayesian probability estimates appear to them as hopelessly muddled because Bayesian probability amalgamates objective frequencies (experimental data) with subjective judgments (prior distributions), thereby contaminating both with individual biases.[345] Much like the Mendelian-biometrics dispute discussed in Part II.C, frequentists and Bayesians talk past each other because they cannot agree on a basic framework or principles for scientific inference.

The choice between Bayesian and frequentist methods ultimately turns on one's view of the merits of a holistic logical approach versus procedural objectivist methods. Their relative values cannot be judged in the abstract. Jeffreys used Bayesian methods very successfully in his work in meteorology and astrophysics, while Fisher's frequentist methods greatly improved plant genetics research and the agricultural sciences.[346] At bottom, however, environmental science is not ideally suited to either frequentist or Bayesian methods. The interpretive circuitousness and experimental constraints of frequentist methods inhibit their effective use by environmental scientists and policymakers. Similarly, the contentious politics, theoretical uncertainties, and limited data found in environmental policymaking all suggest that the Bayesian mixing of judgment and data could further fuel existing disputes by obscuring the solid data and theories that do exist. Frequentists' separation of data analysis (i.e., stage-two statistical inference) from stage-three scientific judgments is often of vital importance to the credibility of environmental science. Ultimately, a flexible approach that is responsive to specific contexts is what is needed.

C. Effectively Integrating Science into Environmental Policymaking

Institutional developments in climate change science policy provide a potential model that integrates the proposals found in the legal and scientific debates over the proper role of scientific judgment in environmental science.[347] Drawing on the parallels already noted between Justice Breyer's expert model and Bayesian methods and between Professor Wagner's procedural scheme and frequentism, this approach rests on the belief that the legal and scientific debates should not be treated separately. Instead, statistical methods should be used in tandem with legal and institutional mechanisms to control

345 HOWIE, *supra* note 37, at 161–62; Efron, *supra* note 326, at 3–4.

346 *See supra* Part I.

347 *See* Mark Schrope, *Consensus Science, or Consensus Politics?*, 412 NATURE 112, 113–14 (2001).

how scientific judgment is used in environmental policymaking. The climate change research community is already, in effect, considering a strategy that would combine a well developed institutional structure with a Bayesian approach to calculating climate change predictions.[348] Under the Intergovernmental Panel on Climate Change (IPCC), formal procedures exist for drafting scientific reports and attaining consensus among its scientific membership.[349] The use of Bayesian methods is already being debated within the IPCC, and if embraced would be particularly noteworthy given the scientific complexity and political importance of climate change science.[350]

The IPCC system represents a promising approach to integrating scientific judgment into environmental policy. This model recognizes that the function of statistics in science is analogous to the role of procedures and institutional structures in law—they both operate as frameworks for ensuring that judgments are transparent and consistent with applicable rules and principles. At the same time, it is clear that environmental policy is reliant on scientific judgments shaped by statistical methods, which may enhance or detract from existing legal measures. Bayesian probabilities, for instance, are less transparent to public scrutiny because they incorporate subjective judgments and objective data into a single quantitative estimate. By considering statistical and legal mechanisms together, lawyers, scientists, and

348 *See, e.g.*, L. Mark Berliner et al., *Bayesian Climate Change Assessment*, 13 J. CLIMATE 3805 (2000); Chris E. Forest et al., *Quantifying Uncertainties in Climate System Properties with the Use of Recent Climate Observations*, 295 SCIENCE 113 (2002); Giles, *supra* note 23, at 476–78; Roger N. Jones, *Managing Uncertainty in Climate Change Projections—Issues for Impact Assessment*, 45 CLIMATIC CHANGE 403 (2000); Schneider, *supra* note 23, at 17–18; T.M.L. Wigley & S.C.B. Raper, *Interpretation of High Projections for Global-Mean Warming*, 293 SCIENCE 451 (2001). These references encompass Bayesian methods that are applied within specific climate models and used as a meta-analytic method to combine information from diverse models and experiments and to generate a final quantitative prediction of future climate change. It appears that the latter meta-analytical method is receiving more attention and opposition than the former. *See* Giles, *supra* note 23, at 476–77; Schneider, *supra* note 23, at 17–18.

349 *See* IPCC, *supra* note 183, at iii–vi; Intergovernmental Panel on Climate Change, *About IPCC*, *at* http://www.ipcc.ch/about/about.htm (last visited Nov. 21, 2003). The IPCC draws scientists from universities, the private sector, and nongovernmental organizations. *Id.* Climate change research has required a careful balancing of consensus processes and critical debate. *See* IPCC, *supra* note 183, at v–vi; Thomas J. Crowley, *Causes of Climate Change Over the Past 1000 Years*, 289 SCIENCE 270, 271–72 (2000) (discussing competing hypotheses for climate change that scientists have had to rule out to conclude that climate change is from anthropogenic sources).

350 *See* Giles, *supra* note 23, at 476–77; Arnulf Grübler & Nebojsz Nakicenovic, Letter to the Editor, *Identifying Dangers in an Uncertain Climate*, 412 NATURE 15 (2001); Schneider, *supra* note 23, at 17–18.

policymakers will gain added flexibility and be in a position to integrate the different mechanisms more effectively.

A complementary strategy, as the IPCC example suggests, could employ legal procedures and institutional structures to counterbalance expertise based Bayesian methods.[351] Just as in the legal context, the added procedures would mitigate potential problems with individual bias. Procedural mechanisms could include justifications of critical scientific judgments (e.g., how experimental results are weighted), consideration of a range of hypotheses or prior distributions, or calculation of Bayesian probabilities under a range of starting assumptions. Institutional measures might consist of balanced membership requirements or consensus processes, much like those associated with the IPCC. Integrating legal procedures and statistical methods in this manner provides a general analytic framework, Bayesian or frequentist, for matching legal procedures with critical scientific judgments. Importantly, this approach differs from Wagner's judicial review proposal, which is reliant on scientists and judges, because it uses broadly applicable statistical frameworks to identify important scientific judgments, rather than a case-by-case approach based on separating science from policy. Similarly, it has advantages over Breyer's proposal insofar as it incorporates procedural mechanisms that mitigate the potential influence of individual bias and political pressures, which Breyer's approach gives short shrift.

Such an integrated approach also could match statistical and legal mechanisms based on whether the relevant science is better served by Bayesian or frequentist methods. Ecological studies, for example, that are less amenable to frequentist analyses could be conducted using Bayesian methods and subjected to additional procedural requirements, such as requiring scientists to explain the rationale for their choice of prior distribution. As a general rule, frequentist methods will be better suited to areas of science where mathematical sophistication is lower, divisive disciplinary controversies are common, and large-scale controlled experimental testing dominates. Bayesian methods will be preferred where mathematical sophistication is higher, disciplinary consensus is more attainable, and observational studies dominate or, as in the case of many ecology studies, statistical long-run frequencies are elusive. These are clearly generalizations, though, that must be re-examined in particular contexts.

351 Alternatively, one could combine Breyer's expert model with frequentist methods (i.e., preclude Bayesian analysis); however, without some kind of further integration this would be little different than the status quo, given the dominance of frequentist methods in environmental science.

Taking advantage of the interplay between statistical methods, legal procedures, and institutional mechanisms in shaping expert judgment has the potential to expand the range of options available to guide scientific judgment in environmental policy. Moreover, by linking law and science in this manner, this approach will promote a deeper appreciation among lawyers and policymakers of the limits and strengths of scientific methods and among scientists of the important role that legal procedures and institutions play in environmental law. Both stand to improve how science is used in environmental policymaking.

CONCLUSION

This Article has analyzed the relationship between scientific judgment and statistics in environmental policy, and has described the competing Bayesian and frequentist approaches to statistical inference. As we have seen, expert judgments are important at each stage of the three-part framework described in the Introduction—experimental quantification, inferences from discrete scientific studies, and integrated scientific assessments. Stage one shows how substantive scientific uncertainties introduce ambiguities that weaken statistical tests, and reveals the importance of using a variety of experimental methods that are independent of the hypothesis being investigated. It also exposes the difficult balancing that is required to protect the integrity of science while ensuring transparency and political accountability. Stage two demonstrates the role of frequentist methods in shaping scientific judgments and reveals how common misapprehensions about them have hampered efforts to alleviate the systemic benign-until-proven-guilty bias of traditional significance tests. It also reveals that statistical tests are more flexible than their skeptics appreciate and proposes a remarkably underutilized method—equivalence testing—to address this bias. Equivalence testing ought to be a standard technique in environmental science, as it already is in medicine under FDA regulations.

Stage three exposes the limitations of both Bayesian and frequentist methods. Just as Breyer's and Wagner's regulatory proposals have proven incomplete in environmental law, so too are Bayesian and frequentist methods imperfectly suited to environmental science. While these limitations cannot be completely overcome, they can be mitigated by using statistical methods in tandem with legal procedures and institutional mechanisms. Indeed, many similarities exist between the legal debate over the appropriate role of expert judgment in environmental policymaking and the growing dispute among envi-

ronmental scientists over statistics. Lawyers and scientists have both adopted procedural and expert based strategies, but where lawyers rely on institutional structures and administrative procedures, scientists employ Bayesian and frequentist methods, respectively. These parallels should aid lawyers in grasping how statistics influences environmental policy. They also suggest that decisions regarding statistical methods and legal measures should not be isolated from each other and that the integrated model proposed in Part IV.C promises to strike the right balance between science and politics in environmental policymaking.