

CONTENTS

HIGH TECHNOLOGY LAW JOURNAL FALL 1992 VOLUME 7 NUMBER 2

ARTICLES

Science and Toxic Torts:

Is There a Rational Solution to the Problem of Causation?

Susan R. Poulter 189

Antitrust and International Competitiveness:

Is Encouraging Production Joint Ventures Worth the Cost?

Donald K. Stockdale, Jr. 269

Software Litigation in the Year 2000:

The Effect of Object-Oriented Design Methodologies on Traditional Software Jurisprudence

David M. Barkan 315

COMMENT

The Experimental Use Exception to Infringement Applied to Federally Funded Inventions

Suzanne T. Michel 369

ARTICLE

SCIENCE AND TOXIC TORTS: IS THERE A RATIONAL SOLUTION TO THE PROBLEM OF CAUSATION?

SUSAN R. POULTER[†]

Table of Contents

I.	INTRODUCTION	190
II.	HARD CASES MAKE BAD LAW	197
III.	ACTIVE REVIEW OF SCIENTIFIC EVIDENCE.....	205
	A. Active Review and the Rules of Evidence	205
	B. Active Review and Scientific Reasoning	207
IV.	ACTIVE REVIEW OF CAUSATION EVIDENCE IN TOXIC TORTS	213
	A. Validity, Reliability, and the Determination of Probative Value	213
	B. Validity and Reliability of Causation Evidence in Toxic Torts	216
V.	DIVERGENCE OF OPINION	241
	A. Deferential Review and the Accumulation of Errors	241
	B. Active Review Exemplified	250
VI.	ACTIVE REVIEW: THE ANTIDOTE FOR JUNK SCIENCE.....	252
	A. Courts' Ability to Review Scientific Evidence	252
	B. Overcompensating for the Deficiencies and Inequities of the Tort System	254
	C. The Costs of Overcompensation	264

© 1993 Susan R. Poulter.

[†] Associate Professor, University of Utah College of Law; J.D. 1983, University of Utah College of Law; Ph.D., 1969, University of California, Berkeley; B.S. 1965, University of California, Berkeley. The author wishes to thank the following colleagues and friends for their thoughtful review and many helpful comments on drafts of this article: Dean Lee Teitelbaum, Professors Leslie Francis and Wayne McCormack and Associate Professor Paul Cassell of the University of Utah College of Law, and Professor Gary Yost of the Department of Pharmacology and Toxicology of the University of Utah. Any errors are, of course, the author's own.

I. INTRODUCTION

Recent controversies over the safety of breast implants,¹ electrical power transmission lines,² and even cellular phones³ portend yet another period of protracted litigation in which the courts will confront issues of what constitutes admissible and sufficient evidence⁴ of causation in toxic torts.⁵ Questions have surfaced regarding the safety of each product, but there is no clearly established causal link between chronic exposure to any of them and disease or injury. News reports indicate that while anecdotal reports abound regarding breast implants, little if any systematic testing has been done to confirm suspicions of harmful effects.⁶ Concerns about cellular phones were prompted by an even sparser array of anecdotal reports and studies.⁷ Electromagnetic radiation from electrical power lines has been studied more extensively, but many scientists remain unconvinced of the purported link between such

1. See, e.g., Philip J. Hilts, *Experts Suggest U.S. Sharply Limit Breast Implants*, N.Y. TIMES, Feb. 21, 1992, at A1.

2. See Bill Richards, *Elusive Threat: Electric Utilities Brace for Cancer Lawsuits Though Risk Is Unclear*, WALL ST. J., Feb. 5, 1993, at A1.

3. See Natalie Angier, *Cellular Phone Scare Discounted*, N.Y. TIMES, Feb. 2, 1993, at C1.

4. The United States Supreme Court recently granted the plaintiffs' petition for certiorari in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992). In *Daubert*, the Ninth Circuit held that animal testing and chemical studies provided insufficient foundation for expert testimony that Bendectin causes limb reduction defects. 951 F.2d at 1131. The court also held that unpublished reanalyses of epidemiologic studies which had not been peer reviewed and which were generated solely for use in litigation were inadmissible on the issue of causation. *Id.*

5. This article uses the term "toxic tort" for cases, including products liability and environmental exposure cases, in which disease or injury is alleged to have resulted from exposure to harmful substances (i.e., chemicals). See 1 MICHAEL DORE, THE LAW OF TOXIC TORTS § 2.02 (1992). The toxic tort rubric also applies to cases involving radiation exposure. See, e.g., *Allen v. United States*, 588 F. Supp. 247 (D. Utah 1984), *rev'd*, 816 F.2d 1417 (10th Cir. 1987), cert. denied, 484 U.S. 1004 (1988). For discussion of the characteristics of toxic torts cases, see *infra* notes 43-49 and accompanying text.

6. This statement is intended to apply to the issue of whether breast implants or their constituents pose systemic risks. There are, of course, cases in which the implants have ruptured or produced localized effects, where the injuries and the causal role of breast implants is not subject to the same level of doubt.

As the breast implant controversy came to a head, *Chemical & Engineering News* reported:

After 30 years of silicone gel breast implant use, the biological, physiological, physical, and chemical reactions of silicones in the human body are likely, finally, to be systematically studied. A major goal of these studies will be determining how often the devices rupture, and what happens when they do.

Lois Ember, *Breast Implants: Silicone Effects in Body to Be Probed*, CHEMICAL & ENGINEERING NEWS, Mar. 2, 1992, at 4. Almost a year later, the *Wall Street Journal* reported that some researchers have identified diseases that they believe are unique to or more common in breast implant recipients. Joan Rigdon, *Breast Implants Raise More Safety Issues: Saline Implants Appear to Carry Hazard as Well*, WALL ST. J., Feb. 4, 1993, at B1.

7. See *infra* notes 10-13 and accompanying text.

exposure and disease.⁸ Nonetheless, all three exposures are the subject of recently filed, and in some cases adjudicated, lawsuits.⁹

The current scare over cellular phones is instructive. The primary "evidence" of a causal link between the phones and brain cancer is the fact that a number of cellular phone users have been diagnosed with brain cancer, several with the cancer located near the location of the phone's antenna in use. Using newspaper estimates of over three million users of hand held portable cellular phones in the United States¹⁰ and 11,000 expected deaths from brain cancer this year,¹¹ it is hardly surprising that several cases of brain cancer in cellular phone users have been reported. One reported laboratory study which reported that radio-frequency radiation increased the growth rate of tumor cells is consistent with the possibility that such radiation could increase the growth rate of preexisting cancers,¹² but it does not prove that there is any effect in humans from cellular phone use.¹³

8. See Richards, *supra* note 2.

9. *Plaintiffs in Georgia, Texas Sue Makers, Contending Devices Caused Various Ailments*, Current Report, Toxics L. Rep. (BNA) 937 (Jan. 8, 1992) (breast implants). On February 4, 1993, the *Wall Street Journal* reported a plaintiffs' lawyer's estimate that 2000 breast implant cases have been or soon will be filed in consolidated court proceedings in Birmingham, Alabama. Rigdon, *supra* note 6. At least one California case produced a verdict for the plaintiff. *Federal Court Upholds \$7.3 Million Award, Says Verdict Supported, Punitives Proper*, Toxics L. Rep. (BNA) 1480 (May 6, 1992). Regarding radiation from electrical power transmission lines, see *Suit Seeks to Hold Two Utilities Liable for Injuries to Family Living Near Substation*, Toxics L. Rep. (BNA) 927 (Jan. 8, 1992). See also Richards, *supra* note 2, at A1 (describing a "nationwide group of law firms eager to turn [electromagnetic field radiation] into a legal battleground").

Cellular phones are at issue in at least one lawsuit. See Angier, *supra* note 3.

10. See Stephen Nolhgren et al., *A Lethal Connection?*, ST. PETERSBURG TIMES, Jan. 10, 1993, at 1A (reporting estimates of 10 million owners of cellular phones, approximately one third of which are hand-held portables).

11. See Mary Lu Carnevale, *Scientists Doubt Phones Cause Brain Tumors*, WALL ST. J., Feb. 3, 1993, at B1. Richard Adamson, a researcher at the National Cancer Institute, was quoted as predicting 11,800 deaths from brain cancer in the U.S. this year. *Id.* Estimating the population of the U.S. at 250 million, the brain cancer death rate would then be approximately 47 per million, leading to an expected mortality of approximately 140 cases per year among the 3 million hand-held cellular phone users. Even if the age-adjusted cancer rates are lower for the age groups who use cellular phones, it is not unexpected that there would be a number of cases of brain cancer among cellular phone users each year. Further, incidence of brain cancer in the United States is undoubtedly somewhat higher than mortality from the disease.

12. See *supra* note 11.

13. Even the study's author, Stephen Cleary, a physiology and biophysics professor at the Medical College of Virginia, was quoted by the *Wall Street Journal* as stating that he does not believe that portable cellular phones cause cancer. Carnevale, *supra* note 11, at B1. The *Journal* cited scientists from the National Cancer Institute, the Food and Drug Administration, the Environmental Protection Agency, and the Federal Communications Commission as stating that they do not believe that phone use causes brain cancer, but they might pose a small risk of increasing the growth rate of existing cancers. *Id.*

Despite the obvious lack of evidence to prove that cellular phone use causes brain cancer given the current state of knowledge, the evidence available today on cellular phones does not differ substantially in quantity or quality from the evidence that courts have found admissible and sufficient in other recent toxic tort cases. Those problematic cases are likely to be supported only by a combination of anecdotal evidence that amounts to no more than coincidence, speculation in the guise of scientific explanation, and testing based on unvalidated methodology or studies that have limited predictive value for human disease. Sometimes, as in the Bendectin litigation, such evidence is urged upon and accepted by courts in the face of overwhelming scientific consensus, supported by evidence, that a substance is unlikely to be a cause of injury. In other cases, very tenuous evidence is deemed sufficient where more probative positive or negative evidence is unavailable. Such unprobative and insufficient evidence and testimony, termed "junk science" by some observers,¹⁴ has been the subject of increasing commentary and criticism.¹⁵

Erroneous plaintiffs' verdicts and the corresponding overcompensation and overdeterrence are not just academic concerns. The prospect of useful products being driven from the market or of economic resources being diverted from productive uses is real, as the cases of vaccines¹⁶ and

14. The term "junk science" has been popularized by Huber. See PETER HUBER, GALILEO'S REVENGE: JUNK SCIENCE IN THE COURTROOM (1991). At least one court has used the term in a toxic tort case as of this writing. Landrigan v. Celotex Corp., 605 A.2d 1079, 1086 (N.J. 1992).

15. See generally Bert Black, *A Unified Theory of Scientific Evidence*, 56 FORDHAM L. REV. 595 (1988); Jude P. Dougherty, *Accountability Without Causality: Tort Litigation Reaches Fairy Tale Levels*, 41 CATH. U. L. REV. 1 (1991); Peter Huber, *Junk Science in the Courtroom*, 26 VAL. U. L. REV. 723 (1992). For commentary on the courts' tendency to ignore probative evidence in favor of unproven mechanistic explanations and medical testimony, see Troyen A. Brennan, *Causal Chains and Statistical Links: The Role of Scientific Uncertainty in Hazardous Substance Litigation*, 73 CORNELL L. REV. 469 (1988).

16. The cost of litigation and the threat of liability have discouraged research and development of new vaccines, as well as production of existing vaccines, activities that are already of marginal interest to pharmaceutical companies because of high production costs and low return on investment. Louis Lasagna, *The Chilling Effect of Product Liability on New Drug Development*, in THE LIABILITY MAZE 335, 341-45 (Peter W. Huber & Robert E. Litan eds., 1991). In 1991, there was only one U.S. manufacturer of vaccines for measles, mumps, rubella, and polio, down from three to six for each. *Id.* at 344. The high price of vaccines for childhood diseases has recently become the focus of public health concerns about low immunization rates among children in the United States. See Richard L. Berke, *President Assails "Shocking" Prices of Drug Industry*, N.Y. TIMES, Feb. 13, 1993, § 1, at 1. Those prices are attributable in part to liability concerns. See Lasagna, *supra* note 16 at 344; James V. Aquavella, *Profits Don't Explain High Drug Costs*, N.Y. TIMES, Feb. 23, 1993, at A20 (letter to the editor) (attributing high costs to product liability insurance and limited life of patent protection).

Bendectin¹⁷ illustrate. Submission of a case to the jury may result in a plaintiff's verdict where even the most cursory examination of the evidence reveals its deficiencies.¹⁸ Verdicts may be very large,¹⁹ and an occasional plaintiff's verdict may even encourage other suits and increase the settlement value of other cases.²⁰ The social and economic significance of breast implants, electrical transmission lines and cellular phones varies considerably, but clearly the costs to society of an erroneous conclusion that any of them causes harm are significant, potentially even catastrophic.

To deal with the problems of junk science in court, several commentators have suggested that courts regularize the standard for admissibility of scientific evidence. One frequent suggestion is that courts reinstate or continue to apply the standard announced in *Frye v. United States*,²¹ which requires that novel scientific evidence have general acceptance within the relevant scientific discipline,²² an issue that the United States Supreme Court is expected to address this year in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*²³ As will be demonstrated in this article, however, many of the issues that arise are more properly viewed as questions about the sufficiency of relevant evidence to meet the more probable than not standard of proof. Thus, solutions that depend on tightening the criteria for admissibility will either require distortion of the

17. Bendectin was eventually withdrawn from the market despite defense verdicts in the overwhelming majority of cases. *Lasagna, supra* note 16, at 340; see also Joseph Sanders, *The Bendectin Litigation: A Case Study in the Life Cycle of Mass Torts*, 43 HASTINGS L.J. 301, 357 (1992).

18. See, e.g., *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984). Obviously, no plaintiff's verdict can result where a case is not submitted for a decision on the merits. It is understood among plaintiffs' lawyers that the objective is to get to trial. Thus, plaintiffs often propose to fully try a few "bellwether" cases, while defendants move for exclusion of evidence and summary judgment on causation issues. See, e.g., *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1547 (D. Colo. 1990), aff'd, 972 F.2d 304 (10th Cir. 1992).

19. In *Ealy v. Richardson-Merrell, Inc.*, Civ. A. No. 83-3504, 1987 WL 18743 (D.D.C. Oct. 1, 1987), rev'd, 897 F.2d 1159 (D.C. Cir.), cert. denied, 498 U.S. 950 (1990), the jury awarded compensatory damages of \$20 million and punitive damages of \$75 million to a boy born with limb reduction defects attributed to Bendectin. The district court allowed the compensatory verdict to stand, but granted remittitur as to the punitive verdict. The compensatory verdict was reversed on appeal.

20. See Sanders, *supra* note 17, at 357.

21. 293 F. 1013 (D.C. Cir. 1923).

22. See, e.g., *Black, supra* note 15, at 637-38; *Huber, supra* note 15, at 742-47. The *Frye* rule is still followed in many jurisdictions. See, e.g., *Christopherson v. Allied-Signal Corp.*, 939 F.2d 1106, 1110 (5th Cir. 1991) (en banc), cert. denied, 112 S. Ct. 1280 (1992). See generally *Black, supra* note 15, at 601 & n.23. *Frye* was also the basis of rejection of certain of plaintiffs' evidence in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992). See discussion of *Frye infra* notes 55-61 and accompanying text.

23. 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992).

admissibility inquiry to encompass sufficiency issues, or will address only part of the problem. Similar concerns are raised by proposals to change the rules of evidence to limit the use of expert testimony.²⁴

The problem of determining the sufficiency of evidence of causation is more directly addressed by proposals that courts use science boards, science panels or court-appointed experts to assist in resolving scientific issues.²⁵ Such proposals, however, except for the use of court-appointed experts, depart substantially from existing notions of civil jurisprudence because they involve delegation to experts of the traditional fact-finding functions of the lay trier of fact.

The thesis of this article is that measures such as the return to the *Frye* rule, or the use of science panels or science courts are unnecessary, because common law courts already possess the authority under the existing rules to "actively review"²⁶ scientific evidence by eliciting and scrutinizing the reasoning underlying scientific evidence and expert testimony and determining its validity and probative worth. As this article will demonstrate, much of the junk science that appears in toxic tort cases is readily apparent or easily uncovered by inquiry of which courts are quite capable.

If active review under the existing rules can uncover bad science, why do a significant number of courts take a lenient posture toward scientific evidence? There appear to be two major reasons for the deferential approach. First, some courts are philosophically indisposed to examine scientific reasoning or methodology, fearing that they are ill-

24. A working group of the Judicial Conference proposed the following amendment of Rule 702 of the Federal Rules of Evidence:

Testimony providing scientific, technical, or other specialized information, in the form of an opinion or otherwise, may be permitted only if (1) the information is reasonably reliable and will substantially assist the trier of fact to understand the evidence or to determine a fact in issue, and (2) the witness is qualified as an expert by knowledge, skill, experience, training, or education to provide such testimony. Except with leave of court for good cause shown, the witness shall not testify on direct examination in any civil action to any opinion or inference, or reason or basis therefor, that has not been seasonably disclosed as required by Rules 26(a)(2) and 26(e)(1) of the Federal Rules of Civil Procedure.

137 F.R.D. 83 (1991). The proposed changes seem more a shift in emphasis than a radical revision of the existing rule. *See also* Black, *supra* note 15, at 611-13 (proposing a modification of Rule 702 to require the court to determine the validity of reasoning as well as its reliability as a precondition to admitting scientific evidence).

25. See 2 AMERICAN LAW INST. REPORTERS' STUDY, ENTERPRISE RESPONSIBILITY FOR PERSONAL INJURY, APPROACHES TO LEGAL AND INSTITUTIONAL CHANGE 332-51 (1991). The ALI Reporters' Study recommendations were based on Brennan, *supra* note 15, and Troyen Brennan, *Helping Courts with Toxic Torts: Some Proposals Regarding Alternative Methods for Presenting and Assessing Scientific Evidence in Common Law Courts*, 51 U. PITTS. L. REV. 1 (1989).

26. For a discussion of "active review," see Black, *supra* note 15, at 674-77.

equipped to delve into scientific disciplines. As will be described below, however, scientific reasoning and legal factfinding employ the same rules of logic. Thus, lay judges need not fear that examination of scientific evidence to determine whether it is soundly reasoned and reliable is beyond their capabilities.

Moreover, the reasons for judicial control of evidence are more compelling where technical evidence is concerned than for non-technical evidence. Judges exhibit no hesitation in barring non-expert testimony based on hearsay and otherwise lacking in foundation even though juries could readily identify the flaws in such testimony with skilled cross-examination and argument by opposing counsel. Juries are less likely to identify the weaknesses in testimony cloaked in technical jargon from an expert with a lengthy list of credentials than in testimony on ordinary factual issues.²⁷ Thus, it is more important for the judge, who understands the legal requirements of proof, to discriminate between reliable and unreliable scientific evidence than between well founded and unfounded evidence on matters within the understanding of ordinary people.²⁸

A second reason for the lenient treatment of scientific evidence in some courts is the apparent desire to compensate for perceived inequities and deficiencies of the tort system. Much of the movement toward the adoption of lenient standards of admissibility and proof of causation in toxic torts has been prompted by the recognition of the difficulties faced by plaintiffs in meeting the traditional requirement that they prove, by a preponderance of the evidence, that their injuries were caused by chronic, low-level chemical or radiation exposures that were remote in time from the manifestation of injury. The paucity of scientific evidence on the causation of diseases such as cancer and birth defects, and the difficulty of distinguishing other identified or background risk factors for the disease, decrease the likelihood that deserving plaintiffs will be compensated. The level of concern about those difficulties was heightened by increasing scientific knowledge of the role of chemicals and radiation in diseases such as cancer and birth defects, as well as scientific speculation about

27. Courts that scrutinize scientific evidence more closely recognize that jurors are likely to be persuaded by the aura of infallibility that surrounds scientific evidence, or by the credentials and certainty expressed by the expert. *See Barefoot v. Estelle*, 463 U.S. 880, 926-28 (1983) (Blackmun, J., dissenting).

28. Courts' abandonment of the *Frye* standard increases the need for judicial scrutiny of scientific evidence because the *Frye* general acceptance standard assures that some evaluation of methods or theories other than that of the expert witness has occurred. Once courts unhinge the admissibility of scientific evidence from scientists' standards, it is incumbent on them to see that other safeguards are in place. *See Steven M. Egesdal, Note, The Frye Doctrine and Relevancy Approach Controversy*, 74 GEO. L.J. 1769, 1787 (1986) (suggesting the need to increase jurors' understanding of novel scientific techniques under the relevancy approach).

potential effects of the greatly accelerated dissemination of untested new chemicals in consumer products and the environment.²⁹ Taking their cue from the scientists,³⁰ legal scholars began to address the difficulties faced by plaintiffs in proving that exposure to toxic substances or chemicals caused their diseases or injuries,³¹ difficulties that can result in uncompensated injuries and the failure to adequately deter harmful activity.³² Lenient standards of admissibility and proof certainly facilitate plaintiffs' recoveries; further, they are consistent with courts' suspicions that mainstream scientists are too demanding in their requirements of proof, and that the unconventional scientists who testify that an exposure caused a plaintiff's disease may be correct.

More than a decade of scientific research into cancer incidence and causation, however, has failed to bear out the fears that prompted deferential review of causation evidence. Many of the assumptions that underlay the shift to more lenient standards for causation evidence in toxic torts are still unproven or are even contrary to current scientific thinking. The contribution of toxic synthetic chemicals and other hazards of the industrial age to cancer and other diseases and injuries is still an open question, but it appears unlikely that such substances cause anything approaching a majority of human cancer and birth defects.

As for the possibility that the unconventional expert may be right, even a superficial examination of much of the disputed evidence reveals that it amounts to speculation about possibilities that have not been tested or that fall far short of meeting the more probable than not standard of proof. Speculation about possibilities forms the beginning, not the endpoint, of factual inquiry, in either the scientific or legal realm. A causal explanation of disease or injury can be said to be probable only when it is supported by observations or data that distinguish between it and other possible explanations. When courts authorize or approve plaintiffs' verdicts without a factual basis for causal inference, they undermine traditional tort requirements for rational factfinding and the "more probable than not" standard of proof. The case for the abrogation of those standards has not been made, nor have courts given full consideration to the implications of such a radical change in the law.

29. See Bruce N. Ames, *Identifying Environmental Chemicals Causing Mutations and Cancer*, 204 SCIENCE 587, 588-89 (1979).

30. See, e.g., *id.* at 592 (recommending short-term mutagenicity testing to expedite identification of environmental mutagens and carcinogens).

31. See, e.g., Jeffrey Trauberman, *Statutory Reform of "Toxic Torts": Relieving Legal, Scientific and Economic Burdens on the Chemical Victim*, 7 HARV. ENVTL. L. REV. 177, 188 n.48 (1983) (citing law review articles and other writings).

32. See generally David Rosenberg, *The Causal Connection in Mass Exposure Cases: A "Public Law" Vision of the Tort System*, 97 HARV. L. REV. 849 (1984).

The purpose of this Article is to demonstrate that courts can and should actively review scientific evidence of causation in toxic tort cases. The next Part describes how courts have loosened the standards for expert testimony in an effort to compensate for the perceived problems faced by toxic tort plaintiffs. Part III then discusses active review and its relation to the rules of evidence and civil procedure and attempts to allay courts' fears that they are ill equipped to evaluate the basis of scientific opinion testimony. Part IV then describes the criteria against which the reliability of scientific evidence can be evaluated and then applies those criteria to the kinds of evidence offered on causation in toxic tort suits. Part V examines a sampling of recent cases that illustrate inadequate judicial scrutiny of scientific evidence, as well as cases that skillfully distinguish probative from nonprobative or insufficient evidence. Lastly, Part VI discusses in depth the factors that underlie courts' failure to examine adequately scientific evidence and shows that many of those concerns are unjustified or that, even where justified, the remedy of authorizing plaintiffs' verdicts that are unsupported by a factual foundation goes too far.

II. HARD CASES MAKE BAD LAW

In the 1960s and 1970s, mounting evidence on the harmful effects of chemicals such as asbestos, vinyl chloride, dioxin and many others, together with the dramatic increase in the use of new chemicals in products ranging from foods, to drugs and medical devices, to many other consumer products, raised concerns that chronic, low level exposures to those substances would lead, or might already have led, to widespread illness and injury.³³ As evidence mounted that exposure to substances such as asbestos and vinyl chloride could cause cancer and other debilitating or fatal conditions, the courts began to see an increasing number of toxic tort suits—tort actions seeking to recover for injuries attributed to toxic substances.

As numerous commentators have explained, proof of causation³⁴ has been the biggest stumbling block to recovery in toxic torts cases.³⁵

33. See R. Jeffrey Smith, *Government Says Cancer Rate Is Increasing*, 209 SCIENCE 998 (1980); Mostafa K. Tolba, *Chemicals in the Environment*, 1979 NAT'L PARKS & CONSERVATION MAG. 16. The controversy continues, as indicated by more recent publications. See Eliot Marshall, *Experts Clash over Cancer Data*, 250 SCIENCE 900 (1990); see also *infra* notes 332-38 and accompanying text.

34. This Article is addressed to issues of causation in fact, by which is meant the issue of whether there is an empirical linkage between the causative event and the claimed injury.

35. See Brennan, *supra* note 25, at 2; Daniel A. Farber, *Toxic Causation*, 71 MINN. L. REV. 1219, 1219-20 (1987); Jean M. Eggen, *Toxic Reproductive and Genetic Hazards in the Workplace: Challenging the Myths of the Tort and Worker's Compensation System*, 60 FORDHAM L. REV. 843, 861-64 (1992) (discussing causation problems in the worker's compensation system);

Both negligence and strict liability require the plaintiff to prove that the substance in question³⁶ caused the plaintiff's disease or injury.³⁷ That inquiry often involves a number of subissues,³⁸ including whether: (1) the toxic substance is capable of causing the harm complained of³⁹; (2) the plaintiff was exposed to the toxic substance in quantity sufficient to cause disease,⁴⁰ and (3) the toxic substance exposure caused the particular plaintiff's injury or disease.⁴¹ Proof of any of these propositions is likely to require expert testimony on scientific evidence.⁴²

Palma J. Strand, Note, *The Inapplicability of Traditional Tort Analysis to Environmental Risks: The Example of Toxic Waste Pollution Victim Compensation*, 35 STAN. L. REV. 575, 583-84 (1983); Note, *Tort Actions for Cancer: Deterrence, Compensation, and Environmental Carcinogenesis*, 90 YALE L.J. 840 (1981) (hereinafter Note, *Tort Actions for Cancer*).

36. Disputes over who produced the offending substance have also been cast as causation questions. These "indeterminate defendant" cases have arisen frequently in asbestos and DES litigation where the plaintiff may have difficulty identifying the producer of the substance to which the plaintiff was exposed, even where the causal connection between the substance and the injury is established. See Richard Delgado, *Beyond Sindell: Relaxation of Cause-in-Fact Rules for Indeterminate Plaintiffs*, 70 CAL. L. REV. 881 (1982); Eggen, *supra* note 35, at 890-91 & n.258.

37. Most courts require proof of causation to meet a "more likely than not" standard. See, e.g., *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1553 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992). See generally Bert Black & David E. Lilienfeld, *Epidemiologic Proof in Toxic Tort Litigation*, 52 FORDHAM L. REV. 732, 749-50 (1984). But see Black, *supra* note 15, at 659-69 (discussing the meaning of "reasonable medical certainty"). Additionally, most jurisdictions require the plaintiff to prove that her injuries would not have occurred "but for" the exposure to the toxic substance. Brennan, *supra* note 15, at 493-94. Where there are two or more contributing causes to a single harm, some courts will require proof only that the exposure was a "substantial factor" in causing the plaintiff's injury or that it "contributed to" the plaintiff's injury. See *Renaud v. Martin Marietta Corp.*, 749 F. Supp. at 1551 (plaintiff must prove that "the exposure caused, or contributed to, plaintiff's injuries"). Proof under the substantial-or-contributing-factor test nonetheless requires establishment of a "but for" causal relationship between the substance and the plaintiff's disease. See Bert Black et al., *Unravelling Causation: Back to the Basics*, 7 Toxics L. Rep. (BNA) 1061, 1063 (1993). A somewhat different formulation, perhaps more suited to the realities of toxic torts, is Calabresi's "causal linkage," that is, the belief that the causative event makes the occurrence of the injury result more likely. See Guido Calabresi, *Concerning Cause and the Law of Torts: An Essay for Harry Kalven, Jr.*, 43 U. CHI. L. REV. 69, 71 (1975).

38. See Black & Lilienfeld, *supra* note 37, at 737-38.

39. See Black, *supra* note 15, at 689. Although this framing of the question seems implicit, plaintiffs sometimes argue that evidence of causation of one type of harm is evidence of causation of other types of harm. *Id.*; see also *Christopherson v. Allied-Signal Corp.*, 939 F.2d 1106, 1115 (5th Cir. 1991) (en banc) (association of nickel and cadmium with small-cell carcinoma of the lung asserted as probative of causation of small-cell colon cancer), *cert. denied*, 112 S. Ct. 1280 (1992).

40. See Black & Lilienfeld, *supra* note 37, at 737-38. Courts sometimes frame the question more simply as whether the plaintiff was exposed to the toxic substance, and there is some divergence in the case law as to the specificity with which exposure must be proved. See *infra* notes 219-20 and accompanying text.

41. This statement, which appears all-inclusive, is intended to cover those aspects of causation-in-fact that remain after exposure and capability of the substance to cause harm ("general causation") are established, including primarily the issue of whether plaintiff's

Several characteristics of the typical toxic tort case diminish the prospects of recovery by deserving plaintiffs.⁴³ The long latency period between exposure and disease manifestation⁴⁴ decreases the likelihood that the plaintiff will even suspect the causal connection, as well as decreasing the likelihood that the plaintiff will be able to marshal the facts on issues such as exposure necessary to prove her case.⁴⁵ Typically there is no clinical evidence capable of linking the substance to the disease.⁴⁶ The situation is further complicated by the fact that exposure to the toxic

injury was the result of the toxic substance exposure or other causes. This issue is sometimes referred to as one of "individual causation" or "medical causation." *Renaud v. Martin Marietta Corp.*, 972 F.2d 304, 306 (10th Cir. 1992) (discussing medical causation); *see also Rosenberg, supra* note 32, at 855-56 (discussing "specific causation").

42. *See, e.g., In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990) (plaintiff's case depended upon expert testimony relating to exposure and causation), *cert. denied*, 111 S. Ct. 1584 (1991); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (expert testimony on exposure and individual causation), *aff'd*, 972 F.2d 304 (10th Cir. 1992). As a definitional matter, this Article will use the terms science and scientific evidence to encompass both science, in the sense of discovery of new factual information, and technology, which can be defined as application of established scientific principles to a particular problem. *See Howard T. Markey, Needed: A Judicial Welcome for Technology—Star Wars or Stare Decisis?*, 79 F.R.D. 209, 210-12 (1978). An additional assumption will be made that scientific evidence will be presented by expert witnesses, because that is most often the case.

43. *See generally Brennan, supra* note 25, at 20-26 (discussing cancer causation); *Strand, supra* note 35, at 578-86.

44. *See Strand, supra* note 35 at 580-81. More precisely, the lapse of time between exposure and the appearance of clinical symptoms may comprise both an induction period, the period of time between the exposure and disease initiation, and a latency period, the interval between disease occurrence and detection. *See KENNETH J. ROTHMAN, MODERN EPIDEMIOLOGY* 14-15 (1986). The period between first exposure and clinically detectable disease for many cancers is 20 to 30 years. *Ames, supra* note 29, at 587. Birth defects that are manifest at birth or soon thereafter would not exemplify this problem to nearly as great a degree.

45. Delay may, however, may increase the chance that epidemiologic evidence will be available. Nonetheless, latency also gives rise to problems under some formulations of statutes of limitation, although many jurisdictions employ the discovery rule to determine when the statute of limitations begins to run. *See Black & Lilienfeld, supra* note 37, at 780; *Strand, supra* note 35, at 580-81.

46. In cases involving "signature diseases," diseases that are almost exclusively associated with a toxic substance, the presence of the condition is highly probative of the causative agent. Examples of signature diseases are mesothelioma, associated almost entirely with asbestos exposure, and clear cell adenocarcinoma of the vagina, associated almost exclusively with diethylstilbestrol (DES) exposure in utero. *See Brennan, supra* note 25, at 21 & n.96. In most cases, however, either the toxic substance is no longer present when the disease manifests itself, as is the case with benzene and leukemia, or its presence, if persistent, is not the only or even most probable explanation of disease. *See Brennan, supra* note 15, at 502. An example of the latter is the almost ubiquitous presence of PCBs in human adipose tissue, apparently without effect in most cases. *See In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 843 (3d Cir. 1990) (discussing ATSDR studies), *cert. denied*, 111 S. Ct. 1584 (1991). *But see Landrigan v. Celotex Corp.*, 605 A.2d 1079, 1087 (N.J. 1992) (discussing the presence of asbestos near the tumor as probative of colon cancer causation).

substance, even at relatively high levels, may not result in disease in most persons.⁴⁷ Moreover, many of the diseases caused by toxic chemicals, particularly cancers and birth defects, occur in the general population.⁴⁸ The absence of any unequivocal linkage between the disease and the toxin, together with the absence of clinical tests that could establish a linkage, means that proof of causation, if it can be made out at all, must be made indirectly, from comparisons between exposed and unexposed groups, or from studies where surrogates such as animals or single-celled organisms are used. Further, there may be other known risk factors for the claimed injury, whose role in the disease process must be considered.⁴⁹

The obvious difficulties of proof in toxic tort cases provoked a flood of commentary and proposals for reform.⁵⁰ A number of commentators have focused specifically on limitations placed by courts on the kinds of

47. Occupational asbestos exposure in nonsmokers increases the risk of lung cancer by about a factor of five, from about 11 per 100,000, for nonsmoking industrial workers not exposed to asbestos, to about 58 per 100,000 for nonsmoking asbestos workers. See U.S. SURGEON GEN., U.S. DEP'T OF HEALTH & HUMAN SERVS., PUB. NO. 85-50207, HEALTH CONSEQUENCES OF SMOKING: CANCER AND CHRONIC LUNG DISEASE IN THE WORKPLACE 216 (1985); see also Rodolfo Saracci, *The Interactions of Tobacco Smoking and Other Agents in Cancer Etiology*, EPIDEMIOLOGIC REV. 175, 181-83 (1987).

48. See, e.g., Rubanick v. Witco Chem. Corp., 593 A.2d 733, 745 (N.J. 1991) (involving contention that PCB exposure caused colon cancer). Colorectal cancer is the second most common cancer in the United States. *Colonoscopy Recommended*, AM. MED. NEWS, Sept. 16, 1991, at 39, cited in Landrigan, 605 A.2d at 1082).

49. For example, although asbestos is recognized as a cause of lung cancer, *see supra* note 47, other causative factors such as smoking are well known. That fact often leads to contentions that the plaintiff's disease was caused by factors other than the toxic chemical exposure. For discussion of attributable risk and the problems of distinguishing among causes, see *infra* notes 206-18 and accompanying text.

50. Some commentators have proposed modification of the tort system's rules of liability, suggesting, for example, that courts recognize causes of action for tortiously created risk. See, e.g., Glen O. Robinson, *Probabilistic Causation and Compensation for Tortious Risk*, 14 J. LEGAL STUD. 779 (1985) [hereinafter Robinson, *Probabilistic Causation*]. Others have suggested that all victims of a disease attributable in part to toxic chemical exposure recover the fraction of their damages that corresponds to the proportion of disease incidence attributable to the toxic exposure. See, e.g., Delgado, *supra* note 36, at 892; Glen O. Robinson, *Multiple Causation in Tort Law: Reflections on the DES Cases*, 68 VA. L. REV. 713, 759 (1982); Rosenberg, *supra* note 32; cf. Farber, *supra* note 35, at 1221 (proposing compensation for the "most likely victim"). A number of courts have redefined damages or injury to include exposure, presumed subclinical injury, medical monitoring costs, or fear of cancer where clinically manifest disease or injury is absent. See 2 DORE, *supra* note 5, §2.02.

Other commentators have recommended the shifting burden of proving causation to defendants, once a threshold showing is made of the possibility of harm. See Note, *Tort Actions for Cancer*, *supra* note 35, at 855-62. Still others have suggested administrative compensation systems with reduced requirements for proof of causation. See Black & Lilienfeld, *supra* note 37, at 734 & nn.3-5 (discussing the Superfund Study Group's proposal for an administrative compensation scheme); see also E. Donald Elliott, *Why Courts? Comment on Robinson*, 14 J. LEGAL STUD. 799, 801 (1985).

evidence deemed admissible or sufficient to prove causation. One set of problems has been courts' reluctance to accept statistical evidence, such as epidemiologic studies, because statistical evidence does not provide mechanistic explanations of cause and because statistics do not provide a basis for distinguishing between persons in an exposed group whose disease was caused by the exposure from those whose disease was caused by background or other risk factors.⁵¹ Recognizing that epidemiologic evidence is often the best if not the only evidence linking a toxic substance exposure to disease, however, recent cases have been more accepting of epidemiologic evidence,⁵² in some cases evidencing quite a sophisticated understanding of epidemiologic evidence.⁵³

Other commentators have urged courts to liberalize the standards for admissibility of scientific evidence in general.⁵⁴ They have suggested that the traditional requirement under *United States v. Frye* that limits the admission of scientific evidence to that generally accepted in the relevant scientific discipline⁵⁵ may preclude recovery by deserving plaintiffs who

51. See, e.g., Black & Lilienfeld, *supra* note 37, at 767; Brennan, *supra* note 15, at 491-501.

52. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307, 313 (5th Cir.) (holding absence of "conclusive" epidemiologic evidence fatal to plaintiffs' case), modified, 884 F.2d 166, 167 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990). The Fifth Circuit subsequently modified *Brock*, stating that the plaintiffs' case was fatally flawed because of their failure to present "statistically significant" epidemiologic evidence. *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 884 F.2d 166, 167 (5th Cir. 1989) (denying plaintiffs' motion for rehearing en banc and modifying prior opinion), cert. denied, 494 U.S. 1046 (1990). Courts willing to accept statistical evidence as probative of the capability of a substance to cause harm have sometimes balked at accepting such evidence on the question of whether the substance caused the plaintiff's injury, on the basis the epidemiologic evidence cannot prove individual causation. See, e.g., *Landrigan v. Celotex Corp.*, 605 A.2d 1079, 1087 (N.J. 1992) (discussing the trial court's refusal to allow an epidemiologist to testify on individual causation). In *Landrigan*, the New Jersey Supreme Court, however, set forth a detailed summary of how epidemiologic reasoning could be applied to the question of individual causation and concluded that an epidemiologist could offer an opinion on that issue, provided the expert's qualifications and methodology withstood the trial court's scrutiny. *Id.* at 1087-89.

53. See, e.g., *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 946-49 (3d Cir. 1990) (discussing statistical significance); *Landrigan*, 605 A.2d at 1087 (discussing the concept of attributable risk derived from epidemiologic studies). Several courts have announced that epidemiologic evidence is the only sufficient evidence on the question of whether Bendectin causes human birth defects. See, e.g., *Brock*, 874 F.2d at 313-15.

54. Anne S. Toker, *Admitting Scientific Evidence in Toxic Tort Litigation*, 15 HARV. ENVTL. L. REV. 165 *passim* (1991); see also citations in *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 740 (N.J. 1991).

55. The traditional standard for determining the admissibility of novel scientific evidence was set forth in *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). The *Frye* court stated, in regard to evidence based on a forerunner of modern polygraph testing, that "the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field to which it belongs." *Id.* at 1014. The *Frye* test is most appropriately applied to the expert's methodology or reasoning, including but not limited to devices or techniques such as the breathalyzer or polygraph. See Black, *supra* note 15, at 627-29. It is sometimes applied to the expert's opinion or conclusions, however,

must rely on novel, yet valid and reliable evidence.⁵⁶ That line of reasoning was accepted in *Ferebee v. Chevron Chemical Company*,⁵⁷ in which the Court of Appeals for the District of Columbia Circuit upheld a jury verdict of liability based on expert opinion testimony on causation that did not enjoy general acceptance in the scientific community.⁵⁸

Ferebee coincided with a general move away from the *Frye* standard under the Federal Rules of Evidence toward the relevancy or reliability test articulated in *United States v. Downing*.⁵⁹ In *Downing*, the Third Circuit stated that the admissibility of scientific evidence should focus on the soundness and reliability of the expert's methodology, the strength of the connection between the evidence and the issues in the case, and the possibility of confusing or misleading the jury.⁶⁰ Acceptance of the

in such cases being stated to require that the expert's opinion or theory be generally accepted by the relevant scientific community. *See, e.g.*, Rubanick v. Witco Chem. Corp., 542 A.2d 975, 982 (N.J. Super. Ct. Law Div. 1988) (applying *Frye* analysis to scientific principle on which expert's opinion was based), *rev'd*, 576 A.2d 4 (N.J. Super. Ct. 1990), *modified*, 593 A.2d 733 (N.J. 1991); *see also* Black, *supra* note 15, at 629-38. Contrarily, some commentators have taken the position *Frye*'s general acceptance test should not be applied to an expert's reasoning or methodology, but only to particular techniques or devices. *See, e.g.*, Christopherson v. Allied-Signal Corp., 939 F.2d 1106, 1131-33 (5th Cir. 1991) (Rawley, J., dissenting), *cert. denied*, 112 S. Ct. 1280 (1992). Many jurisdictions still follow *Frye*. *See, e.g.*, Christopherson, 939 F.2d 1106.

56. *Frye* has proven to be a significant barrier to novel scientific theories and methodologies. Edward J. Imwinkelreid, *The Standard for Admitting Scientific Evidence: A Critique from the Perspective of Juror Psychology*, 28 VILL. L. REV. 554, 555-56 (1982-83). As Huber has pointed out, however, when the *Frye* inquiry is directed to the methodology and reasoning underlying scientific opinion, a novel opinion on causation will easily pass muster if it is based on well-established and properly conducted methods, such as epidemiologic studies. Huber, *supra* note 15, at 744.

57. *See Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984).

58. *Id.* at 1535-36. The *Ferebee* court did not reject *Frye* out of hand, however, but construed it as applicable only to novel techniques or methodologies, not scientific opinion testimony. *Id.* at 1535.

59. 753 F.2d 1224 (3d Cir. 1985); *see also* *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 856-60 (3d Cir. 1990), *cert. denied*, 111 S. Ct. 1584 (1991); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1242 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988); Rubanick v. Witco Chem. Corp., 593 A.2d 733, 746 (N.J. 1991).

60. In *Downing*, the Third Circuit articulated the proper test as follows:

In our view, Rule 702 [of the Federal Rules of Evidence] requires that a district court ruling upon the admission of (novel) scientific evidence . . . conduct a preliminary inquiry focusing on (1) the soundness and reliability of the process or technique used in generating the evidence, (2) the possibility that admitting the evidence would overwhelmingly confuse or mislead the jury, and (3) the proffered connection between the scientific research or test result to be presented and the particular disputed factual issues in the case.

Downing, 753 F.2d at 1238. The *Downing* reliability standard is inherently more flexible than *Frye* because it is not tied to "general acceptance." Nonetheless, courts recognize that acceptance in the expert community is an important indicium of reliability. *See, e.g., id.*

expert's techniques or methodology in the relevant scientific community is evidence of soundness, but need not be the sole basis for that determination.

Although *Frye* has been justifiably criticized as too simplistic and inflexible,⁶¹ the *Downing* standard is equally problematic when it is used to justify such minimal scrutiny of the reliability of scientific evidence, particularly of expert opinion testimony, that it amounts to no standard at all. The troublesome, deferential application of the reliability standard adopts the approach that if "qualified" experts are willing to testify that a causal relationship exists, the court is willing to uphold a plaintiff's verdict without examining whether a reasoned basis exists for the expert's opinion.⁶² This approach is undoubtedly the result of some courts' reluctance to delve into the reasoning underlying scientific evidence, a reluctance that results in deference to the expert with seemingly impressive credentials. The crucial determination then becomes whether the expert is qualified, a particularly weak screening device given the lenient standards for determining expert qualifications.⁶³

Deferential review is the gateway for the admission of junk science into the courts. When courts do not examine the reasoning of expert testimony, they are likely to accept medical opinion based on the facts in the case at hand, or supported by perhaps a few other case reports, facts

Thus, the *Frye* standard is related to reliability, though more limiting. See generally Imwinkelried, *supra* note 56.

61. Part of the difficulty with the *Frye* rule is the lack of consensus regarding the subject matter to which it applies. For example, is it the expert's opinion, the reasoning or methodology that underlies the opinion, or both that must be generally accepted? See *supra* note 55. The better rule would seem to be that the *Frye* general acceptance test applies to the expert's reasoning and methodology, but not to the opinion or conclusion derived from that methodology. Otherwise, the *Frye* rule effectively delegates part of the admissibility determination to the scientific discipline, obviating the need for the court to evaluate the expert's reasoning or methodology. On the other hand, as Black has pointed out, the general acceptance test of *Frye* is not an appropriate standard to apply to the uncertainty or accuracy (i.e., the reliability) of scientific methodology. See Black, *supra* note 15, at 629-57. The Ninth Circuit's opinion in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 951 F.2d 1128 (9th Cir. 1991), *cert. granted*, 113 S. Ct. 320 (1992), appears to commit this error, when it frames the admissibility standard regarding an unpublished, un-peer-reviewed reanalysis of epidemiologic data as follows: "Expert opinion based on a scientific technique 'is admissible if it is generally accepted as a reliable technique among the scientific community.'" *Id.* at 1129 (quoting *United States v. Solomon*, 753 F.2d 1522, 1526 (9th Cir. 1985)).

62. *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 429 U.S. 1062 (1984), is the leading case following this approach and is often cited by other courts taking similar approaches. In *Ferebee*, the Court of Appeals for the District of Columbia Circuit upheld a jury verdict for the plaintiff where testimony of causation was based on "tissue samples, standard tests, and patient examination." *Id.* at 1536. There is nothing in the opinion to suggest that the cited tests and examinations were capable of indicating the cause of the lung disease complained of, however.

63. FED. R. EVID. 702 provides that a witness may be qualified as an expert "by knowledge, skill, experience or training."

that cannot establish causation because the coincidence of exposure and disease may be the result of chance.⁶⁴ In some cases, courts accept as sufficient medical or similar opinions supplemented by reference to animal studies, chemical structure-activity analyses, mutagenicity testing, or other similar lines of reasoning that are subject to a large degree of uncertainty.⁶⁵ Affirmative epidemiologic evidence of a statistically significant association between the alleged causative agent and human disease is absent.⁶⁶ As a practical matter, only those cases based on studies in human populations of the association of suspected toxic substances and disease—e.g., epidemiologic studies or highly unusual disease clusters—have proven to be sound as new scientific information developed.⁶⁷

A reliability analysis should not result in uncritical acceptance of junk science.⁶⁸ Tort jurisprudence requires that there be a rational basis

64. See, e.g., *Ferebee*, 736 F.2d at 1535.

65. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 735-36 (N.J. 1991) (expert based opinion that PCBs caused plaintiff's colon cancer on animal test reports and other cancer cases at Witco). The Bendectin litigation has been characterized by plaintiffs' cases based on animal testing, structure-activity relationships, *in vitro* testing, and reanalysis of data from epidemiologic studies that failed to show statistically significant increased risks. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990). See generally *Sanders, supra* note 17. Brennan in particular has urged courts to admit animal studies and other methods used in cancer and other medical research. *Brennan, supra* note 25, at 41-57; see also Michael D. Green, *Expert Witnesses and Sufficiency of the Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and the Bendectin Litigation*, 86 NW. U. L. REV. 643, 680-81 (1992).

66. Occasionally, plaintiffs may offer a "reanalysis" of existing epidemiologic data. See *infra* notes 350-58.

67. It may be tempting to characterize the argument made herein as establishing a threshold requirement of epidemiologic evidence to support a toxic tort case. A number of commentators have characterized *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990), and cases that have followed it as creating such a threshold in Bendectin cases. See, e.g., Green, *supra* note 65, at 679-82. The intent of this Article, however, is to show why, given the present state of toxicological science, anecdotal evidence, animal test results, and other evidence offered when positive human evidence is missing are generally unreliable and insufficiently probative in the typical toxic torts case. The kind of analysis proposed herein can be applied to new information as it develops, without the rigidity of a *per se* rule about specific kinds of evidence.

68. In *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988), Judge Weinstein excluded the causation opinion testimony of several of plaintiffs' witnesses because he concluded that their testimony, which relied on animal tests and studies of industrial exposures, and which failed to consider and eliminate other causal explanations, was "insufficiently grounded in any reliable evidence." *Id.* at 1248-51. Although Judge Weinstein cited Rule 703 as the basis of his ruling, *see id.* at 1243-55, it is clear that he recognized the uncertainty associated with causal inferences derived from animal studies or human studies where exposures differed widely from plaintiffs', particularly where the experts ignored more relevant studies and alternative causal explanations. *See id.* at 1250. Under the analysis proposed in this Article, the factors cited by Judge Weinstein would be part of a reliability

for judicial findings of fact.⁶⁹ The relevancy or reliability standard's "soundness and reliability" inquiries bear directly on whether there is a rational basis for findings of fact and whether the evidence is sufficient to meet the more probable than not standard of proof.⁷⁰ Active review facilitates the inquiries necessary to decide those issues, while deferential review avoids them. Courts cannot and should not avoid those responsibilities by deferring to "qualified" experts.

III. ACTIVE REVIEW OF SCIENTIFIC EVIDENCE

A. Active Review and the Rules of Evidence

The active review contemplated by this article and being conducted by some courts is a process in which the court conducts two inquiries. First, the court examines the evidentiary basis and reasoning of scientific opinion testimony and determines whether there is a rational basis for the opinion. The evidentiary basis of the opinion, as well as the expert's reasoning, can be probed by the proponent of the testimony, the opponent, or the court,⁷¹ and will often be assisted by the defendants' experts.

The second inquiry focuses on the sufficiency of the admissible evidence to meet the plaintiff's burden of proof. This inquiry goes to the reliability or accuracy of the evidence and requires that the plaintiff present admissible evidence from which a reasonable juror could find that it is more probable than not that the defendant caused the plaintiff's

analysis. *See infra* notes 310-14 and accompanying text; *see also* Black, *supra* note 15, at 674-76.

69. *See, e.g.*, *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. at 1250.

70. FED. R. EVID. 104 requires the court to determine questions of admissibility of evidence. *See, e.g.*, *Egger v. Burlington N.R.R.*, No. CV89-159-BLG-JFB, 1991 U.S. Dist. LEXIS 19240 (D. Mont. Dec. 18, 1991). Generally, the proponent of evidence must demonstrate by a preponderance of the evidence that the evidence in question is admissible. *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. at 1239. Under FED. R. CIV. P. 50, 56, the court must determine the sufficiency of the evidence on a motion for summary judgment, a motion for a directed verdict, or a motion for judgment notwithstanding the verdict. Generally the standard for granting any of the foregoing (for defendant) is that no reasonable juror could find or have found for the plaintiff. *See, e.g.*, *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1555 (D. Colo. 1990) (granting summary judgment to defendants), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

71. FED. R. EVID. 705 provides: "The expert may testify in terms of opinion or inference and give his reasons therefor without prior disclosure of the underlying facts or data, unless the court requires otherwise. The expert may in any event be required to disclose the underlying facts or data on cross-examination."

As a practical matter, the court may have to do some translation of the language of the scientific field or of legal expressions into commonly understood terms. The court may also be guided by court-appointed experts who serve as witnesses or advisors. For an example of court-appointed experts serving as advisors to the judge, see *Renaud*, 749 F. Supp. 1545.

disease or injury. The same tools used to probe the underlying reasoning of the evidence can be used to inquire into its accuracy, but the question of whether the evidence is sufficiently accurate to satisfy legal standards is, of course, a legal question.

It is important to note that active review is not strict scrutiny.⁷² The plaintiff need not show that her evidence is stronger than the defendant's or that it meets some high level of certainty. The plaintiff's scientific evidence need only be such that a rational factfinder could conclude from the testimony that it is more likely than not that the defendant caused the plaintiff's injury.⁷³ Only when the factual basis and reasoning underlying the expert's opinion on causation do not meet that minimum level of rationality and accuracy should the evidence be excluded.

Active review is not tied to any particular formulation of the standards for admissibility of expert testimony. It is, however, more easily related to the "reliability" determination embraced by a number of courts⁷⁴ than it is to the general acceptance rule of *United States v. Frye*.⁷⁵ The *Frye* rule forecloses the occasion for the court to examine the reasoning underlying the expert's method; however, it leaves questions such as the applicability of a generally accepted method to a particular case, the way in which a generally accepted method was carried out in a particular case,⁷⁶ and the sufficiency of the evidence to be addressed under other criteria. Thus, even if the United States Supreme Court upholds the application of the *Frye* rule in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,⁷⁷ it will not eliminate the need for courts to actively review scientific expert testimony.⁷⁸

72. But see, e.g., Peter A. Bell, *Strict Scrutiny of Scientific Evidence: A Bad Idea Whose Time Has Come*, Toxics L. Rep. (BNA) 1014 (1992). A more apt comparison would be to the "hard look" doctrine of administrative law, that is, the view articulated by Judge Leventhal that courts reviewing the decisions of a technical agency, such as the Environmental Protection Agency, should review the evidence on which the agency's decision is based to determine "whether the agency decision was rational and based on consideration of the relevant factors." Ethyl Corp. v. EPA, 541 F.2d 1, 34-36 (D.C. Cir.) (en banc), cert. denied, 426 U.S. 941 (1976). Judge Leventhal's views are not without detractors. See *id.* at 66-67 (Bazelon, C.J.). A non-technical, lay jury's decisions would seem to justify greater scrutiny than those of a regulatory agency with technical expertise.

73. See, e.g., *Renaud*, 749 F. Supp. 1545.

74. See *supra* notes 59-60.

75. See *supra* note 55.

76. Cf., e.g., *United States v. Jacobetz*, 955 F.2d 787 (2d Cir. 1992) (the value of DNA testing depends on whether accepted protocols were followed in the specific case), cert. denied, 113 S. Ct. 104 (1992).

77. 951 F.2d 1129 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992).

78. The *Frye* rule at least creates a threshold for evaluation of the evidence that may serve to curb courts' tendencies to uncritically admit all arguably relevant evidence. The reliability standard nonetheless can serve an appropriate screening function if the court actually conducts a reliability analysis.

B. Active Review and Scientific Reasoning

One of the factors that seems to dissuade courts from scrutinizing scientific evidence more carefully is the belief that the differences between scientific and legal inquiries into causation are such that courts are poorly equipped to examine and evaluate science.⁷⁹ Actually, in determining whether there is a link between an event and a later harm, law and science use identical reasoning processes. Differences between scientific and legal institutions, goals and policies, however, obscure that commonality.

Judge Markey has succinctly stated an essential distinction between science and technology on the one hand, and law on the other:

The differences between the judicial and scientific-technological processes are profound and pervasive. Failure to recognize that difference has led to judicial expressions of frustration and an unfortunate tendency to rest judicial decisions on current, often transient, "truths" and "facts" of science and technology. The purpose of science is to learn physical facts. The purpose and function of technology is to provide a means of using that learning. All that is important and necessary, but that's all it is—learning and using physical facts.

The purpose and function of law is to resolve disputes and to facilitate a structure for the organization of a just society—in a word, to provide justice.⁸⁰

As Markey suggests, science and law do differ in important ways. The culture, institutions and processes by which scientific knowledge is developed and refined are very different from those of law.⁸¹ The development of scientific knowledge involves observation, hypothesis

79. See, e.g., *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984). The *Ferebee* court stated:

Judges, both trial and appellate, have no special competence to resolve the complex and refractory causal issues raised by the attempt to link low-level exposure to toxic chemicals with human disease. On questions such as these, which stand at the frontier of current medical and epidemiologic inquiry, if experts are willing to testify that such a link exists, it is for the jury to decide whether to credit such testimony.

Id. at 1534.

80. Markey, *supra* note 42, at 210, quoted in *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 741 (N.J. 1991).

81. See Sheila Jasanoff, *What Judges Should Know About the Sociology of Science*, 32 JURIMETRICS J. 345 (1992); David Kaye, *Proof in Law and Science*, 32 JURIMETRICS J. 313, 317-18 (1992). The discomfort many scientist-experts experience in the adversarial setting of legal adjudication is largely due to scientists' perception that the law requires unequivocal statements on matters that are not clear cut from the scientist's perspective. Further, they are uncomfortable with the legal system's insistence on decisions, often before adequate evidence is available from a scientific perspective. For further discussion of the differences between the processes of legal and scientific inquiry, see Huber, *supra* note 15, at 739-42 (1992).

building, testing, generalizing, and consensus building.⁸² Legal factfinding, on the other hand, is adversarial, confrontational, and directed toward a definitive result in the case at hand. Concern for consistency from case to case plays a lesser role in law⁸³ than in science.⁸⁴

Unfortunately, these institutional and methodological differences obscure the reality that factfinding, that is, science in its broadest sense, is a necessary part of legal decisionmaking. Legal decisionmaking has additional policy components beyond the purely factual, so that it may attach different consequences to the same facts than would a scientist. Thus, the starting point for the analysis of the relationship between science and law on the issue of causation is a delineation of the factual and nonfactual components of legal concepts of cause.

To be sure, causation issues in tort law have nonfactual, policy-laden elements, as exemplified by the legal concept of proximate cause.⁸⁵ All tort theories include some notion of "cause-in-fact" as a prerequisite to liability,⁸⁶ however, and where cause-in-fact is concerned, science and law are attempting to answer the same questions. Further, law, like science, accepts only rational or reasoned findings of fact.⁸⁷ Most importantly, scientific and legal factfinding employ the same logic.⁸⁸

82. See Black, *supra* note 15, at 615-27.

83. In *Wells v. Ortho Pharmaceutical Corp.*, 615 F. Supp. 262 (D. Ga. 1985), *aff'd in part, modified on other grounds*, 788 F.2d 741 (11th Cir.) (modifying damage award), *cert. denied*, 479 U.S. 950 (1986), the court held that plaintiff had proved that her daughter's birth defects were caused by the mother's prenatal use of a spermicide, despite FDA approval and scientific consensus that spermicides do not cause birth defects. *See id.* at 266. But see *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307, 315 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989), *cert. denied*, 494 U.S. 1046 (1990). In *Brock*, the court expressed the hope that its ruling would have "a precedential effect on other cases pending in this circuit which allege Bendectin as the cause of birth defects." *Id.* at 315.

84. That is not to say that science is fixed and unchangeable. Scientific knowledge is always open to revision as new information comes to light that is inconsistent with previously understanding. The point, however, is that scientific reasoning requires that a scientific explanation accommodate and be consistent with all the available data at any point in time.

85. The concept of proximate cause is generally recognized as encompassing policy questions of how closely the defendant's tortious conduct must be related to the plaintiff's injury for the defendant to be held liable. *See, e.g.*, Richard W. Wright, *Responsibility, Risk, Probability, Causation, Naked Statistics and Proof: Pruning the Bramble Bush by Clarifying the Concepts*, 73 IOWA L. REV. 1001, 1011-12 (1988). Viewed in that light, the proximate cause requirement is a limitation on liability where defendant's conduct was the actual cause of plaintiff's injury. *Id.*

86. *See id.* Of course, the way in which the factual question is framed, as well as the burden of proof and evidentiary standards has policy overtones. *See Eggen, supra* note 35, at 899-904 (suggesting shift of burden of proving causation); Nancy L. Firak, *The Developing Policy Characteristics of Cause-in-Fact: Alternative Forms of Liability, Epidemiologic Proof and Trans-Scientific Issues*, 63 TEMP. L. REV. 311, 313 (1990) (arguing that courts' acceptance of epidemiologic evidence is a policy choice rather than a factual conclusion).

87. *See Wright, supra* note 85, at 1011-12; *see also In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1250 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487

Much of the early commentary about the differences between science and law in toxic torts concerned courts' discomfort with statistical evidence of causation. Commentators have attributed that discomfort in part to courts' preferences for mechanistic causal explanations and their reluctance to rely heavily or entirely on statistical evidence.⁸⁹ Courts and lay persons typically think about causal issues in terms of *how* things happen and statistical evidence does not explain how events occur.⁹⁰

When mechanistic thinking about cause is extended to the area of toxic substance disease causation, it immediately encounters a large, perhaps insurmountable, stumbling block. Scientists know very little about how, in a mechanistic sense, toxic substances cause diseases such as

U.S. 1234 (1988). The ubiquitous legal requirement that there exist a rational or reasonable basis for findings of fact evidences the underlying assumption that reasoning and logic must connect evidence to conclusions.

88. David Kaye has demonstrated that science and law use the same logical rules in proving facts. *See Kaye, supra* note 81. He concludes: "[W]hen it comes to proving facts, the logic of law and that of science are one and the same. At an abstract level, the rules of inference can be given the same formal representation." *Id.* at 317; *see also* Lee Loevinger, *Standards of Proof in Science and Law*, 32 JURIMETRICS J. 323, 328 (1992).

In regard to the role of social science in overturning *Plessy v. Ferguson*, Kenneth B. Clark has stated:

The development of science as an approach to the determination of truth involved the development of methods for the control of errors in human observation, judgment, biases, and vested interests. These were the factors which seemed to have distorted man's concept of, or blocked his contact with, the "truth" or "facts" of experience. When they are operative, man's "common knowledge" becomes inconsistent with "scientific knowledge." When they are controlled or for some other reason non-operative, "common knowledge" and "scientific knowledge" are coincident—both reflecting the nature of reality, truth, or facts, as these are knowable to the human senses and intelligence.

Science is essentially a method of controlled observation and verification for the purpose of reducing human errors of observation, judgment, or logic. Science begins with observation and ends by testing its assumptions against experience. It is not a creation of another order of reality. In a very basic sense there cannot be a "legal fact" or a "fact of common knowledge" which is not at the same time a "scientific fact." Whenever this appears to be true, one or the other type of "fact" is not a fact.

Kenneth B. Clark, *The Desegregation Cases: Criticism of the Social Scientist's Role*, 5 VILL. L. REV. 224, 233 (1959).

89. *See e.g.*, Brennan, *supra* note 15, at 478-91. Brennan refers to mechanistic conceptions of cause as "corpuscularianism," after the writings of various philosophers of science. *See id.* at 478-79.

90. The understanding of how a cause produces an effect makes us more comfortable with the conclusion that causation occurred. Richard Wright puts it this way:

Usually, the issue [of proving causation] is what has happened—including how it happened and who did it—although sometimes the issue is what is expected to happen—for example, the expected reduction in future income as an element of damages. That is, proof generally involves either causal explanation or causal prediction.

Wright, *supra* note 86, at 1049.

cancer or injuries such as birth defects.⁹¹ Nonetheless, they may know a considerable amount about whether toxic substances cause disease or injury through inferences drawn from statistical associations and other indirect means.⁹² Thus, the shift in thinking required for courts to come to grips with current scientific knowledge had more to do with abandoning a felt need for an explanatory process that increases comfort with the causal inference than it did with redefining causation.

Courts' discomfort with statistical evidence has gone beyond the absence of mechanistic explanations, however.⁹³ Statistical evidence by definition provides information only about the incidence of disease in groups. Where there are other possible causes of disease, statistical evidence cannot determine which individuals' diseases within the exposed group were caused by background or other factors.⁹⁴ It can only

91. Brennan, *supra* note 25, at 20-25 (discussing scientific evidence of cancer causation).

92. *Id.*

93. See Black & Lilienfeld, *supra* note 37, at 744-50; Brennan, *supra* note 15, at 483-93. Brennan states that courts' refusal to consider and accept statistical evidence reflects and is consistent with courts' traditional reliance on mechanistic causal explanations. See Brennan, *id.* at 491-92.

94. Extensive or complete reliance on epidemiologic proof and other statistical evidence is not without its detractors. See, e.g., Michael Dore, *A Commentary on the Use of Epidemiologic Evidence in Demonstrating Cause-in-Fact*, 7 HARV. ENVTL. L. REV. 429 (1983); Wright, *supra* note 86, at 1049-67 (arguing that particularistic evidence is required to prove actual causation). Dore reiterates the commonly held view that epidemiologic evidence is proof not of actual, individual causation, but only of risk. See Dore, *supra*, at 435. Regarding the use of epidemiology in proving risk (apparently meant as the ability of a substance to cause harm), Dore states:

Within the limitations just discussed, epidemiologic evidence can demonstrate the relative level of risk to which the defendant's activities exposed the members of the plaintiff's group. This risk, of course, does relate to the individual plaintiff. Courts that fail to distinguish the issue of risk from that of actual causation may accordingly, but erroneously, permit the evidence of risk to establish causation. Epidemiologists do not design their studies to resolve issues of individual biological causation, however, and the courts must strictly limit the use of such studies for this purpose.

The limitations on epidemiology's ability to prove individual causation stem from its general and statistical nature. Epidemiologic studies are general in that they deal with sources of disease in groups of people rather than particular individuals. Being statistical, they quantify the probabilities, or risks, that members of a group will contract certain diseases under certain conditions. The only individual cause-and-effect relationship that epidemiologic evidence can show is that the defendant's conduct increased the plaintiff's risk of injury to some statistically measurable extent. *It cannot answer the critical question whether the defendant's conduct actually injured the plaintiff.*

Id. at 436 (citations omitted). Dore and other detractors of statistical evidence frame the question incorrectly, however. The issue in toxic torts is whether there is evidence from which an inference can be made that it is more probable than not that the exposure caused the plaintiff's disease. As others have pointed out, the statistical evidence provided by epidemiology is probative of that issue. See Black & Lilienfeld, *supra* note 37, at 764-69 (combining relative risk with more-probable-than-not standard of proof); Khristina L. Hall

provide an estimate of the likelihood that an individual's disease was caused by the toxic substance in question.⁹⁵ Thus, courts' concerns are not unreasonable. The more likely than not standard of proof, however, implicitly contemplates the marshaling of facts that ultimately prove liability in terms of probabilities.

Uncomfortable with factual indeterminacy, some courts rejected statistical evidence entirely, demanding evidence that is particular to the plaintiff.⁹⁶ Other courts have accepted statistical evidence on issues such as whether a toxic substance is capable of causing harm, but not on the question of whether it caused the plaintiff's harm.⁹⁷ A number of recent cases, however, have recognized the necessarily statistical nature of proof at all levels in toxic torts, and accepted statistical evidence as probative of individual causation, at least where there is evidence indicating a greater than 50% likelihood that the toxic substance caused the plaintiff's disease.⁹⁸ A number of recent decisions evidence a sophisticated understanding of epidemiologic evidence and its relation to legal standards of proof.⁹⁹

The remaining areas where science seems to fit poorly with legal problems are largely the result of failure to distinguish legal standards of proof from factual issues. Courts are concerned that they must decide cases based on the information available, which may not be complete enough to satisfy the requirements of a particular scientific discipline.¹⁰⁰ Some courts perceive scientists as generally requiring higher levels of certainty than does the law.¹⁰¹ That perception may be correct in some instances, particularly in areas such as epidemiology, where standard

& Ellen K. Silbergeld, *Reappraising Epidemiology: A Response to Mr. Dore*, 7 HARV. ENVTL. L. REV. 441, 445-46 (1983). Indeed, signature diseases, which are usually not perceived as presenting difficult individual causation issues, are simply cases in which the statistical evidence is very persuasive because the background incidence of disease is very low compared to the incidence in the exposed population.

95. See *supra* notes 94 and accompanying text.

96. See Brennan, *supra* note 15, at 492 & nn.114-15.

97. See, e.g., *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992) (reversing the trial court's rejection of opinion testimony on individual causation based on epidemiology).

98. See, e.g., *id.* at 1087.

99. See, e.g., *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 946-56 (3d Cir. 1990) (discussing the statistical significance of epidemiologic data); *Landrigan*, 605 A.2d at 1085-87 (discussing the significance of relative risk and attributable fraction).

100. See, e.g., *Ferebee v. Chevron Chem. Corp.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984). The *Ferebee* court held that a treating physician could testify to his opinion that a cause and effect relationship existed between the insecticide paraquat and Ferebee's pulmonary fibrosis even if such a relationship had not been "clearly established" by animal or epidemiologic studies. *Id.* at 1535.

101. In *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 737, 740-41 (N.J. 1991), the New Jersey Supreme Court made several references to the "extraordinarily high level of proof" required by the scientific method. Defendant's witness apparently played into that concern, however unwittingly. See *id.* at 737.

protocols for statistical analysis of relative risk data typically require a 95% level of certainty that an observed increased in risk is not due to chance.¹⁰² Scientists do not require a high degree of certainty for all purposes, however. Risk assessment for purposes of regulation is based on highly uncertain risk estimates. Additionally, scientists often use highly tenuous or uncertain assumptions in making decisions about further research.¹⁰³

The issue of how much uncertainty is acceptable is a legal requirement to be applied to the evidence once the uncertainty attending the scientific evidence is established.¹⁰⁴ Where the law requires the plaintiff to prove her case by a preponderance of the evidence, current standards permit the plaintiff to win if sufficient evidence is available, but not prevail if it is not available.¹⁰⁵ Scientific evidence can be evaluated against those standards, irrespective of whether the scientific discipline would be satisfied or not with the available level of certainty.¹⁰⁶ Moreover, the fact that scientists may require a different level of certainty is not a good reason to dispense with science's requirement of a reasoned analysis, a requirement common to law and science. Unfortunately, some courts throw the baby out with the bathwater by rejecting scientific reasoning altogether when they perceive scientists' requirements for certainty to be too stringent.¹⁰⁷

102. Black & Lilienfeld, *supra* note 37, at 757 n.104; *see infra* notes 364-67 and accompanying text; *see also* DeLuca, 911 F.2d at 946-49 (discussing statistical significance in epidemiology).

103. For example, as discussed in Wells v. Ortho Pharmaceutical Corp., 615 F. Supp. 262 (D. Ga. 1985), *aff'd in part, modified on other grounds*, 788 F.2d 741 (11th Cir.) (modifying damages), cert. denied, 479 U.S. 950 (1986), the "Oeschli study" raised suspicions about the possibility of an association between spermicides and birth defects and recommended further study. *Id.* at 284. Further studies with greater statistical power failed to confirm that suspicion. *See id.*

104. *See* Black, *supra* note 15, at 600 (discussing reliability as a legal question).

105. From that perspective, science and law seem to have parallel requirements because each refuses to reach an affirmative conclusion that causation exists until an acceptable level of certainty is attained, even though the law and the scientific discipline may require different levels of certainty. The relationship between legal and scientific notions of sufficiency of proof is perhaps less clear, however, than is the identity of the logic employed by each. *See* David Kaye, *On Standards and Sociology*, 32 JURIMETRICS J. 535 (1992); Lee Loevinger, *On Logic and Sociology*, 32 JURIMETRICS J. 527 (1992). *Compare* Kaye, *supra* note 81, *with* Loevinger, *supra* note 88.

106. That is not to say that scientists' perceptions of the appropriate level of certainty should be ignored. Requirements such as epidemiologists' practice of requiring a 95% confidence level often have their roots in years of experience in the discipline. With epidemiologic studies in particular, there may be undetected systematic bias in selection of the comparison groups, including the possibility of undetected confounding factors, that are not taken into account in the statistical analysis. *See* ROTHMAN, *supra* note 44, at 89-96; Black & Lilienfeld, *supra* note 37, at 737-38; *infra* text accompanying notes 346-49.

107. *See, e.g.*, Rubanick v. Witco Chem. Corp., 593 A.2d 733 (N.J. 1991), *discussed* *infra* notes 287-300 and accompanying text.

Courts can and should evaluate the underlying reasoning of scientific evidence and measure its reliability or uncertainty against legal standards of sufficiency to meet the applicable burden of proof. The following part of this article attempts to facilitate that process by explicating the bases on which courts can recognize and reject invalid or unreliable evidence, matters on which the differences between science and law are a matter of degree, not kind. Thus, courts need not fear that delving into science and technology will be entirely a foray into alien territory.

IV. ACTIVE REVIEW OF CAUSATION EVIDENCE IN TOXIC TORTS

A. Validity, Reliability, and the Determination of Probative Value

Whether courts operate under the *Frye* rule, the "reliability" standard of *United States v. Downing*,¹⁰⁸ or some other formulation of the rules governing scientific expert testimony, the question courts must answer when they evaluate scientific evidence is, "How probative is it?"¹⁰⁹ That question includes two subissues, however: validity and reliability.¹¹⁰ Validity is the issue of whether the evidence is capable of producing the kind of information sought; thus, it is essentially equivalent to the concept of relevance as used in the rules of evidence.¹¹¹ Reliability connotes the likelihood of a correct or accurate result,¹¹² and thus encompasses notions of certainty or accuracy.¹¹³ Reliability is

108. 753 F.2d 1224 (3d Cir. 1985).

109. So framed, that question corresponds to the determination of "reliability" under *United States v. Downing*. See *supra* notes 59-63 and accompanying text.

110. See Black, *supra* note 15, at 599-600 (discussing validity as part of the reliability determination).

111. FED. R. EVID. 401.

112. See *supra* note 110.

113. This Article thus adopts and expands on the analytical framework proposed by Black, although it uses the terms validity and reliability in a slightly different way. See generally Black, *supra* note 15.

Black defines validity as "that which results from sound and cogent reasoning," and reliability as meaning "that a successful outcome, or correct answer, is sufficiently probable for a given situation." *Id.* at 599-600. Thus, he frames validity as a scientific question, and reliability as a legal one. *Id.* at 600. Validity is to be determined largely by reference to widespread acceptance in the scientific community of the underlying reasoning. *Id.* at 637-38. Black also recognizes, however, that some aspects of the validity analysis relate to the specifics of a particular case that must be examined apart from the test of general acceptance. See *id.* at 657-58 (discussing *Downing* court's evaluation on remand of the applicability of research on eyewitness identification to the facts at hand).

As defined by Black and as used herein, validity is a subissue of reliability, rather than a separate and independent factor, since invalid reasoning or methodology cannot

therefore the ultimate indicator of the probative value and sufficiency of evidence, either alone or in combination with other evidence, to meet the more probable than not standard of proof.

Consider, for example, a diagnostic blood test for a viral blood disease. Without the blood test, the disease can be diagnosed only by elaborate procedures. A simple test is desired for screening large numbers of blood samples for the presence of the virus. A virologist might speculate about any number of parameters that might be indicative of the presence of the virus. None of the possible indicators could be used as a diagnostic test, however, until validated by testing that demonstrates a correspondence between the indicator (a "positive" test) and the presence of the virus. This example illustrates the more general principle that where the physical connections between observed and inferred facts are hidden from direct observation, it is necessary for the inferred connection (e.g., between the indicator and the virus) to be validated through trials or tests that independently measure the properties or characteristics that are ostensibly connected.¹¹⁴

A valid method may nonetheless be insufficiently reliable for evidentiary purposes; that is, the method may be incapable of producing the desired information to an acceptable level of certainty. Using again the example of a test for an asymptomatic virus, the test might have a high rate of false positives or false negatives, or both. Thus, although persons who are test positive are more likely than those who test negative to actually have the virus in their blood, the test may be too inaccurate or unreliable for the purpose for which it is administered.¹¹⁵ Similarly, if the question of whether someone is infected with the virus were a factual

produce reliable or accurate results. *See id.* at 599-606, 613. Reliability is the criterion that courts tend to apply to expert scientific evidence; thus, the proposed analysis fits within recognized criteria for evaluating scientific evidence. *See supra* notes 59-60 and accompanying text.

The definitional structure used herein departs, however, from the usage of the terms validity and reliability in social science research. In social science disciplines, reliability describes the reproducibility of the results and validity describes the degree to which the phenomenon measured corresponds to the phenomenon sought to be measured. Thus, this Article's use of reliability to encompass the accuracy as well as the reproducibility of an outcome encompasses some issues that would be characterized as validity issues under the social science rubric.

114. In toxic torts, validity issues are present when a physician or other expert witness testifies on causation based on patient examination even though there is no clinical basis for linking individual cases to a particular causative agent. *See infra* text accompanying notes 197-205.

115. If, for example, the test were used to determine whether donated blood is safe for transfusions, a significant rate of false negatives would be of much greater concern than a correspondingly high rate of false positives.

question in a legal setting, our hypothetical test might be insufficiently reliable to satisfy the legal standard of proof.¹¹⁶

Validity or reliability questions may arise when methodology that has proved valid and reliable is applied in new circumstances. Invalid application of valid methodology may result from extending a method or line of reasoning to purposes for which it has not been validated.¹¹⁷ In toxic torts, this question arises in connection with whether the conclusions derived from toxicological research on animals or single-celled organisms are applicable to humans.¹¹⁸

Uncertainty or reliability questions may also result from the improper application of valid and reliable methodology. Failure to properly calibrate an instrument such as a breathalyzer, or other concerns related to how a method is applied in a particular case, may increase the likelihood of erroneous results.¹¹⁹ Assume, for example, that in the hypothetical virus test, the incidence of false negatives increases with the length of time that the patient's blood samples are stored before the laboratory test is run. The inferences drawn from a test run by a laboratory that stores its blood samples longer than the optimum time for the test would be subject to a greater variation and uncertainty than results from a laboratory that runs its tests promptly.

A subset of questions regarding "reliability as applied," particularly where the methodology involves calculations from raw data, concerns the quality and quantity of the underlying data. In toxic torts, the data on which estimates of exposure to a toxic substance are based are often sketchy or subject to large uncertainties. Those uncertainties make the inferences of causation that depend on the exposure data similarly uncertain and unreliable.

116. If the question whether someone is infected with a virus were part of the prosecution's proof in a criminal action, that fact would have to be proven beyond a reasonable doubt, so that any significant rate of false positives would likely render it "unreliable" for that purpose. Proof in a civil action would have to satisfy a more probable than not standard, so that a somewhat higher level of false positives, possibly up to 49%, could be tolerated. Courts sometimes refuse to admit evidence that nominally satisfies the applicable standard of proof, however.

117. The issue of the generalizability of a study is characterized as one of external validity. See ROTHMAN, *supra* note 44, at 95-96 (epidemiologic studies).

118. See, e.g., *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1273 (E.D.N.Y. 1985) (questioning the use of human epidemiologic studies of workplace dioxin exposures and animal studies as evidence of effects in Vietnam veterans), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

119. See MCCORMICK ON EVIDENCE § 209, at 513 (Edward W. Clearly ed., 3d ed. 1972) (discussing factual predicate for admitting chemical testing for alcohol intoxication). Courts differ, however, in their approach to whether the manner in which a method is applied goes to the weight of the evidence rather than its admissibility, and is therefore a jury question. See, e.g., *United States v. Jakobetz*, 955 F.2d 786 (2d Cir.) (DNA testing), *cert. denied*, 113 S. Ct. 104 (1992).

B. Validity and Reliability of Causation Evidence in Toxic Torts

Analysis of the kinds of evidence at issue in the typical toxic torts case illuminates many of the problems with such evidence that tend to be obscured when cases are considered as a whole. The following discussion is intended to suggest some ways of looking at such evidence to distinguish probative evidence from junk science; thus, the discussion deals separately with the subissues of the ability of the toxic substance to cause disease (general causation), exposure, and the causation of the plaintiff's disease (individual causation). It would be well to keep in mind, however, that the ultimate question on causation is whether the evidence allows a reasoned conclusion that exposure to the toxic substance in question, rather than other known or unknown factors that also cause such disease, caused the plaintiff's condition.

As noted previously, the characteristics of toxic tort cases impose limitations on the ability to establish causal connections between exposure and disease. The latency periods typical of toxic tort injuries, the absence in most cases of a unique signature injury associated with a toxic substance, the fact that injury does not occur in every instance of exposure, and the absence of clinical indicators that discriminate among causes of a particular individual's disease all tend to obscure toxic injury causation.¹²⁰ In simple terms, the typical toxic tort case looks something like this: The plaintiff believes she has been exposed to a toxic chemical. She has a disease that is commonplace, or at least not unknown, in the general population. The current progress of the disease bears no relation to the continuation of exposure, and the exposure may have long since terminated. There is no diagnostic or clinical test that can determine what caused her disease.

How can such a plaintiff prove that a toxic substance caused her disease? Because of the absence of clinical indicia of cause, the plaintiff must always make out her case indirectly. First, she needs evidence that the substance can cause the condition from which she suffers and of the circumstances under which disease causation is reasonably likely to occur. The coincidence of exposure and disease in the same individual, while necessary, can never be sufficient to prove the capability of the substance to cause disease. Similar problems attend the use of anecdotal case reports or evidence of clusters of disease that have not been subjected to statistical analysis because a certain amount of coincidence and toxic chemical exposure or even clustering of a disease can occur as the result of random chance.¹²¹

120. See *supra* notes 43-49 and accompanying text.

121. Chance may lead to disease clusters rather than disease uniformly distributed throughout a large population. Anecdotal reports and clusters of disease are important in

Second, she needs to establish that she is within the class of persons to which inferences from the general causation evidence should be applied. This second, particularistic causation component of proof, which is discussed later in this article, usually involves two parts: proof of sufficient exposure to permit the inference that the general causation evidence is applicable to her and a demonstration that other causal explanations, including background causes, are less likely causes than the toxic substance exposure.¹²²

1. ABILITY OF THE TOXIC SUBSTANCE TO CAUSE DISEASE (GENERAL CAUSATION)

a. Causal Inferences from Human Disease

On the issue of general causation, systematic studies that can account for the effects of chance are necessary to allow a causal inference to be drawn from data on exposure and disease incidence in humans. Epidemiologic studies, which involve comparisons of disease incidence in exposed and unexposed human populations,¹²³ are based on this line of reasoning. A higher incidence of disease in the exposed population, if parameters of the study are such that the differential rates of disease are unlikely to be due to chance or other confounding factors,¹²⁴ may be indicative of a causal relationship between exposure to the toxic substance and disease.¹²⁵ Scientists have long accepted epidemiologic studies as indicative of causal relationships and courts have more recently begun to do so. Epidemiologic studies are the basis of causation findings

the identification of possible causal links that should be investigated further, however. See Brennan, *supra* note 25, at 21.

122. The issue of whether other causal explanations are less likely is referred to herein as the issue of individual causation.

123. Other commentators have described epidemiologic studies and their relation to proof of causation of human disease. See Black & Lilienfeld, *supra* note 37. See generally 2 DORE, *supra* note 5, §§ 25.01-05. Epidemiologic studies will be described in more detail in the discussion of distinguishing among causes. See *infra* notes 180-90 and accompanying text; see also *infra* notes 342-58, 364-76 and accompanying text (discussing limitations of epidemiology and statistical significance).

124. See 2 DORE, *supra* note 5, §§ 25.02[4], 25.03.

125. Cause is an inference drawn from epidemiologic studies; the studies themselves can only directly prove an association between exposure and disease incidence. Epidemiologists use the Henle-Koch-Evans postulates or other similar premises as criteria for arriving at biological inferences of causation from epidemiologic studies. Black & Lilienfeld, *supra* note 37, at 762-64. The Henle-Koch-Evans postulates are addressed to the magnitude of the risk elevation in the exposed group and other factors tending to increase the plausibility of a biological relationship between the exposure and disease. See *id.*

in asbestos injury claims, and have served as important evidence of the lack of causation in the Bendectin cases.¹²⁶

Epidemiologic studies are expensive to conduct and are subject to a number of limitations on the size of the effect they can detect.¹²⁷ Thus, it is sometimes argued that case reports and clusters of disease constitute sufficient evidence of the capability of a substance to cause toxic injury.¹²⁸ Case reports and disease clusters are sometimes sufficient to raise suspicions and stimulate investigation of toxic chemicals as causative agents.¹²⁹ Benzene was identified as a leukemogenic agent through clinical studies of case reports beginning in the late 1800s,¹³⁰ and vinyl chloride was more recently recognized as carcinogenic through the appearance of clusters of angiosarcoma of the liver in plant workers in the early 1970s.¹³¹ Those examples, however, are typified, in the case of benzene, by very high exposures and the accumulation of evidence over decades,¹³² or by the unexpected appearance of an otherwise very unusual disease.¹³³ Such identifications through case reports and clusters, however, depend on at least a rough sense that the incidence of the disease in the exposed group exceeds the background rates,¹³⁴ even if the reports of unusually high incidence are not initially subjected to the same rigorous statistical analysis as is typical of an epidemiologic study.¹³⁵ Moreover, those initial clusters or unusual case reports will often suggest other places to look for additional evidence, such as workplace exposures involving the same substance, or other users or consumers of the suspect chemical.¹³⁶ The absence of similarly affected individuals among other

126. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307, 313 (5th Cir.) (plaintiffs could not succeed without epidemiologic evidence), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990).

127. See Brennan, *supra* note 25, at 53 & n.228; Green, *supra* note 65, at 653.

128. See *Rubanick v. Witco Chem. Corp.*, 542 A.2d 975 (N.J. Super. Ct. Law Div. 1988), rev'd, 576 A.2d 4 (N.J. Super. Ct. App. Div. 1990), modified, 593 A.2d 733 (N.J. 1991); *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984).

129. See Brennan, *supra* note 25, at 502.

130. JOHN GRAHAM ET AL., IN SEARCH OF SAFETY: CHEMICALS AND CANCER RISK 119-23 (1988).

131. See David D. Doniger, *Federal Regulation of Vinyl Chloride: A Short Course in the Law and Policy of Toxic Substances Control*, 7 ECOLOGY L.Q. 497, 500 (1978).

132. See *supra* text accompanying note 130.

133. Doniger, *supra* note 131, at 500.

134. See GRAHAM ET AL., *supra* note 130, at 119-23 (discussing relative risk estimates derived from clinical studies and epidemiologic studies of disease in benzene-exposed workers).

135. A causal argument based on observation of an otherwise unknown group of symptoms in breast implant recipients may be possible if recent reports of such symptoms are borne out. See Rigdon, *supra* note 6, at B1.

136. The suspicions aroused by the initial reports of angiosarcoma of the liver in B.F. Goodrich's vinyl chloride plant were quickly confirmed by reports from other companies. See Doniger, *supra* note 131, at 500.

populations with similar exposures would suggest that the cluster is a statistical accident rather than a true cluster.

Case reports and apparent disease clusters are likely to be argued in toxic tort cases in circumstances where they do not have even minimal indicia of reliability. In *Renaud v. Martin Marietta Corp.*,¹³⁷ the plaintiffs argued that the existence of four cases of childhood cancer in Friendly Hills, an area in which only two would have been expected, was evidence that the substances allegedly in their water supply had caused their cancers.¹³⁸ Plaintiffs' experts agreed, however, that the Friendly Hills population was too small to yield meaningful results. Moreover, another expert's opinion was that four cases of childhood cancer was within the expected range for the community.¹³⁹

b. Animal Studies¹⁴⁰

Animal studies, other biological assay methods and chemical structure-activity relationships, all of which are used by toxicologists to estimate human risk from toxic chemicals,¹⁴¹ are much more problematic than epidemiologic studies in the toxic tort context. The use of such methods in risk regulation is based on unproven assumptions about the applicability of the results of such studies to humans, assumptions that are subject to a large degree of uncertainty and in some cases skepticism in the scientific community.¹⁴²

Animal studies are based on the theory that substances that cause harmful effects in animals are likely to cause similar harmful effects in humans.¹⁴³ That thesis is supported by observations that many substances that cause harmful effects in one species also cause harmful effects in other species.¹⁴⁴ All but one of the chemicals identified by epide-

137. 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

138. *Id.* at 1554-55.

139. *Id.* at 1551. Dr. Steven Piantodosi's epidemiological study on the incidence of cancer in children in Friendly Hills indicated that the difference between expected and observed incidence rates was not statistically significant. *Id.*

140. Animal testing has been treated as either admissible or inadmissible. See Jack L. Landau & W. Hugh O'Riordan, *Of Mice and Men: The Admissibility of Animal Studies to Prove Causation in Toxic Tort Litigation*, 25 IDAHO L. REV. 521 (1988-89). As is discussed *infra* notes 151-68 and accompanying text, the issue of animal testing should be addressed as one of how probative is animal testing of causation of human disease, that is, as a question of sufficiency rather than of relevance.

141. Brennan, *supra* note 25, at 21-23, 44. See generally Chemical Carcinogens: A Review of the Science and Its Associated Principles, 50 Fed. Reg. 10,371 (Office of Science & Technology Policy 1985) [hereinafter OS&TP, Chemical Carcinogens].

142. Brennan calls these issues "trans-scientific." See Brennan, *supra* note 25, at 23. Such issues are not always inherently unprovable, although it may be impractical to do so.

143. See Brennan, *supra* note 15, at 504-06.

144. See, e.g., James E. Huff & Joseph K. Haseman, *Exposure to Certain Pesticides May Pose Real Carcinogenic Risk*, CHEMICAL & ENGINEERING NEWS, Jan. 7, 1991, at 33, 34

miologic studies as causing cancer in humans have also proven to be carcinogenic in one or more animal species.¹⁴⁵ Thus, there appears to be some correlation between carcinogenicity in animals and carcinogenicity in humans. Similar observations and findings have been made with respect to other kinds of toxic effects, including teratogenic effects.¹⁴⁶

As any observer of the popular media knows, however, animal testing for diseases such as cancer, which has a long latency periods, and for which even low incidence rates are of concern, are conducted under conditions that are very different from the usual human exposure scenario.¹⁴⁷ Animal studies of carcinogenicity typically utilize doses at or near the maximum level tolerated by the animal.¹⁴⁸ That practice is necessitated by the need to detect effects in relatively small groups of test subjects, in a relatively short period of time. Those same concerns also have led to protocols using animal strains bred for their susceptibility for tumor formation.¹⁴⁹ Additionally, the route of administration may differ from the likely human exposure route.¹⁵⁰

The prediction of effects in humans from animal testing involves a number of extrapolations—from animal species to humans, from one route of administration to another, and most acutely, from a high-dose exposure in which the animals are typically subjected to the maximum dose they can tolerate (the MTD),¹⁵¹ to a low-dose chronic exposure.¹⁵² Each of those extrapolations introduces uncertainty into the predictive value of animal testing in proving causation of human disease.¹⁵³

(reporting that information on carcinogenicity of 8 of 54 known carcinogens was first obtained in animal studies).

145. OS&TP, Chemical Carcinogens, *supra* note 141, at 10,411; J.F. Robens et al., *Methods of Testing Carcinogenicity*, in PRINCIPLES AND METHODS OF TOXICOLOGY 251, 253 (A. Wallace Hayes ed., 2d ed. 1989).

146. See Jeanne M. Manson & L. David Wise, *Teratogens*, in CASARETT AND DOULL'S TOXICOLOGY 226, 240 (Mary O. Amdur et al. eds., 4th ed. 1991) [hereinafter CASARETT & DOULL].

147. Ames, *supra* note 29, at 589; see Landau & O'Riordan, *supra* note 140, at 545.

148. OS&TP, Chemical Carcinogens, *supra* note 141, at 10,377; Bruce N. Ames & Lois S. Gold, *Cancer Prevention Strategies Greatly Exaggerate Risks*, CHEMICAL & ENGINEERING NEWS, Jan. 7, 1991, at 28, 29; see also Kent R. Stevens & Michael A. Gallo, *Chronic Toxicity Studies*, in PRINCIPLES AND METHODS OF TOXICOLOGY, *supra* note 145, at 237, 238-39.

149. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,377.

150. Animal testing may involve skin application, oral gavaging or injection, *see id.* at 10,413-14, rather than the usual human exposure routes of inhalation, ingestion or dermal contact. *See id.*

151. See Ames & Gold, *supra* note 148, at 29. Test animals such as rodents live only one to two years, though they may receive test doses throughout the majority of their lifespans. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,413, 10,414.

152. See Landau & O'Riordan, *supra* note 140, at 543-48.

153. The use of animal test results as proof of toxic effects in humans can be regarded as raising validity issues, because it is questionable whether results in one species can be extrapolated to another at all. See Black, *supra* note 15, at 677-79; Green, *supra* note 65, at 654-56. This issue is treated here as one of uncertainty or inaccuracy, however, because

Differences in species can have a dramatic impact on the effects of a toxic substance,¹⁵⁴ as can routes of administration.¹⁵⁵

The high dose exposure scenario of typical animal testing protocols raises several concerns. One concern relates to the model used to extrapolate the results of high dose exposures to the much lower doses encountered by humans. The lack of a complete mechanistic understanding of cancer causation precludes the adoption of any particular extrapolation model with a high degree of certainty.¹⁵⁶ For example, one

even if validity is assumed, the uncertainty attending extrapolation of results from animal studies to humans will usually render them insufficiently probative to support a plaintiff's verdict. Additionally, most scientists regard animal studies as having some validity in predicting human disease, and such studies are widely used for regulatory purposes. See OS&TP, *Chemical Carcinogens*, *supra* note 141. Animal studies vary in their predictive value for humans, however, depending on the nature of the effects being studied and the number of species in which toxic effects of a substance have been confirmed. Models that may improve predictive capabilities are being developed for quantitative interspecies extrapolations. See Robert A. Scala, *Risk Assessment*, in *CASARETT & DOULL*, *supra* note 146, at 985, 993 (discussing potency correlations of animal and human carcinogens). Thus, it seems appropriate to address the extrapolation of animal test results to humans as an issue of the degree of certainty that attends that extrapolation in a given instance, rather than to assert as a general proposition that animal tests results can or cannot in any instance be validly extrapolated to humans. *Id.*

154. Animal testing is of limited value in a context where false positives are a concern, as is the case with toxic torts, a generalization that cuts across the various types of effects for which such studies are conducted. For example, Manson and Wise report that of 38 substances with positive teratogenic findings in humans, only one was negative in all animal species studied, thus producing a low rate of "false negatives" for human teratogenicity. Manson & Wise, *supra* note 146, at 240. In contrast, of 165 substances studied with no teratogenic finding in humans, only 47, or 29%, were negative in all laboratory animal test species. *Id.* Similar uncertainties occur in animal testing for carcinogenicity. Although all but a few of the 30 or so known human carcinogens (substances or industrial processes) are also carcinogenic in at least one animal species, there are many more substances that have exhibited carcinogenicity in animals that are not known to be human carcinogens. *Id.* It is not uncommon for a substance to exhibit carcinogenicity in one species and not in another. See Scala, *supra* note 153, at 992. For example, in one study of almost 1000 chemicals, only 76% of rat carcinogens were positive in mice, and 70% of mouse carcinogens were positive in rats. *Id.* Studies of carcinogenic potency of the same substances in animals and humans have yielded good correlations for some substances, but animal data overpredicts human response by as much as a factor of 500 for vinyl chloride. See Landau & O'Riordan, *supra* note 140, at 536 (citing Michael D. Hogan and David G. Hoel, *Extrapolation to Man*, in *PRINCIPLES AND METHODS OF TOXICOLOGY*, *supra* note 145, at 879). That difference may be due to incomplete data on human cancer, but it cannot be ignored.

155. For example, EPA's "level of regulatory concern" for inhalation of chromium has been stated as 1.9×10^{-6} mg/day, as compared to 0.1 mg/day for exposure by ingestion. See, e.g., Final Exclusion, 53 Fed. Reg. 29,038, 29,040-41 (EPA 1988) (tbls. 1 & 2) (evaluating petition for delisting of hazardous waste). The inhalation level of regulatory concern was thus set 50,000 times lower than the ingestion level, apparently due to demonstrated respiratory tract carcinogenicity of inhaled chromium as compared to lower risks through other routes of exposure. See Robert A. Goyer, *Regulatory Toxicology*, in *CASARETT & DOULL*, *supra* note 146, at 623, 639.

156. See generally Robert A. Scala, *Risk Assessment*, in *CASARETT & DOULL*, *supra* note 153, at 985, 990-91.

possible set of assumptions is that no dose of a carcinogen is completely risk free and that the disease incidence rate will be directly proportional to the dose. Those assumptions lead to a linear extrapolation model.¹⁵⁷ Another possibility, which apparently applies to some carcinogens, is that at very low levels, a toxic chemical exerts no adverse effects and that such effects appear only when a threshold level of exposure is exceeded.¹⁵⁸ The set of assumptions adopted in a particular instance can lead to vastly different predictions of the effects of low dosage exposures, sometimes as much as several factors of ten.¹⁵⁹

The accuracy of risk extrapolations from exposure of animals to the MTD has recently been called into further question by prominent researchers in the field of carcinogenesis.¹⁶⁰ Bruce Ames, the developer of the "Ames test" for mutagenicity,¹⁶¹ now argues that risk estimates obtained under such circumstances are largely due to toxic effects of the test chemical, rather than factors that might operate at lower doses in human.¹⁶² Thus, the results from animal studies may not be predictive of human carcinogenicity under the usual exposure scenario.

Despite the sparse knowledge of mechanisms of cancer causation, toxicologists have identified a number of steps in the carcinogenesis process, including DNA alteration, DNA expression, and promotion and progression to neoplastic or cancerous tumors. They have also identified two general classes of carcinogens. See Gary M. Williams & John H. Weisburger, *Chemical Carcinogenesis*, in CASARETT & DOULL, *supra* note 146, at 127, 129-31, 170-85. DNA-reactive carcinogens are those that appear to initiate cancer through chemical alteration of DNA. *Id.* at 170. Epigenetic carcinogens, on the other hand, do not necessarily react with DNA and exert carcinogenic effects through other pathways such as by promoting the growth of dormant cancer cells. *Id.* at 185.

157. See Scala, *supra* note 153, at 990-91 (discussing a linearized, multi-stage, nonthreshold model); see also OS&TP, Chemical Carcinogens, *supra* note 141, at 10,438-39. This model assumes that there is no threshold dose below which cancer does not occur, but recognizes the multi-step nature of carcinogenesis. See *id.* Threshold doses are common for acute effects of toxins, i.e., those that occur within a short time after exposure. See Curtis D. Klaasen & David L. Eaton, *Principles of Toxicology*, in CASARETT & DOULL, *supra* note 146, at 12, 38. However, whether carcinogens are subject to thresholds is an unresolved issue. See Scala, *supra* note 153, at 990-91. The choice of model can make a difference of several orders of magnitude in risk levels. As a matter of policy, the conservative linearized multi-stage extrapolation model is often used to estimate the upper bound on risk. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,438-39.

158. See Williams & Weisburger, *supra* note 156, at 154. This assumption appears a likely model for carcinogens classed as promoters rather than as initiators. See David J. Hanson, *Dioxin Toxicity: New Studies Prompt Debate, Regulatory Action*, CHEMICAL & ENGINEERING NEWS, Aug. 12, 1991, at 7, 13 (EPA reconsidering model for dioxin carcinogenicity low-dose extrapolation to allow for threshold).

159. See *supra* note 158; see also OS&TP, Chemical Carcinogens, *supra* note 141, at 10,439.

160. Ames & Gold, *supra* note 148, at 29.

161. See *supra* note 6 and accompanying text.

162. See *supra* note 139 and accompanying text; see also Bruce N. Ames et al., *Ranking Possible Carcinogenic Hazards*, 236 SCIENCE 271 (1987). The authors believe the toxicity effect to be particularly true for carcinogens that are not DNA reactive. See Bruce N. Ames & Lois S. Gold, *Too Many Rodent Carcinogens: Mitogenesis Increases Mutagenesis*, 249 SCIENCE

Quite a number of toxic torts plaintiffs have offered animal studies in support of their contentions that the substances in question can cause harm in humans.¹⁶³ In some cases courts have been willing to entertain such evidence,¹⁶⁴ while others have found it inadmissible¹⁶⁵ or insufficient.¹⁶⁶ The closer examination of the assumptions and methodologies involved in animal testing, however, reveals that the extrapolation of animal test results to humans is too uncertain—the potential for error is too high—for animal testing alone to support an inference that it is more probable than not that a substance causes cancer or birth defects in humans at a specified level of exposure.¹⁶⁷ There is considerable doubt about the inference that an animal carcinogen is a human carcinogen at all. Even if that hurdle is assumed away, however, the uncertainties that attend the interspecies and high-dose to low-dose extrapolations necessary to extend animal test results to human exposure scenarios are simply too large. Policy considerations in the regulatory arena dictate or at least support the use of models that overpredict rather than underpredict risks levels.¹⁶⁸ Risk estimates based on unproven dose-response models, where the choice of model may alter results by a thousand times or more, however, are not consistent with a more likely than not standard of proof.

c. Biological Screening Methods

Even greater problems attend the application of biological screening methods such as short term assays. Short term assays are designed to detect mutagenic effects and cancer-initiating or promoting properties of

970 (1990). Ames and coworkers direct their argument to the allocation of scarce resources for cancer prevention purposes. Their concern about the predictive value of animal testing and other protocols for determining carcinogenicity bears on the general causation issue in toxic torts, however.

163. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D.Colo. 1990), aff'd, 972 F.2d 304 (10th Cir. 1992); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985) (Agent Orange and dioxin), aff'd, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988).

164. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991).

165. See, e.g., *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985), aff'd, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988).

166. See, e.g., *Brock*, 874 F.2d 307.

167. In conjunction with positive epidemiologic studies, positive animal results may be relevant. In such instances, however, some would argue that such evidence is cumulative and of such low probative value as to warrant its exclusion. See, e.g., Landau & O'Riordan, *supra* note 67, at 551-54.

168. Douglas G. Camp et al., *Pesticide Regulation Is Sound*, CHEMICAL & ENGINEERING NEWS, Jan. 7, 1991, at 44, 46.

substances.¹⁶⁹ Mutagenicity,¹⁷⁰ for example, is used as a predictor of carcinogenicity because many known carcinogens are also mutagens.¹⁷¹ Mutagenicity is also considered to be an indicator of potential for causing birth defects.¹⁷² Short term assays' predictive capabilities have been validated by reference to animal carcinogenesis,¹⁷³ however, so they are subject to all of the uncertainties of animal testing, as well as additional uncertainties introduced by the procedure itself.¹⁷⁴ Moreover, because such tests involve single-celled organisms, they are less likely than *in vivo* testing in animals to represent the response of humans.

Another kind of evidence, typically offered as evidence of causation of birth defects or other noncancerous disease or injury, is *in vitro* testing, tests involving exposure of isolated groups of cells or organs¹⁷⁵ to suspect chemicals. To test for teratogenesis (birth defects), fetal cells or embryos may be used. These tests are fraught with uncertainties, however, related to whether and to what degree the chemical in question would reach or react with the sensitive cells or organs in a whole organism.¹⁷⁶ They are also fraught with the same uncertainties relating to interspecies extrapolation as is animal testing generally.

d. Chemical Structure-Activity Analysis

Chemical structure-activity analysis is a kind of scientific reasoning by analogy that is based on the recognition that similarities in chemical structure sometimes correspond to similarities in biological activity.

169. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,403.

170. Mutagenicity refers to the alteration of the genetic material of a cell.

171. Ames, *supra* note 29, at 589.

172. *Id.* at 587.

173. *Id.* at 588.

174. Single-celled organisms are, of course, farther removed biologically from humans than are mammals such as mice and rats which are typically used in animal testing. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,404 (listing commonly used assays). Short-term assays are utilized to select substances for chronic (i.e., long-term) animal testing. See *id.* at 10,408. The OS&TR review states the utility of short-term assays succinctly:

Short-term tests are presently limited in their ability to predict the presence or absence of carcinogenicity and cannot supplant data from long-term animal studies or epidemiologic data since the tests do not necessarily screen for all potential means of cancer induction and do not necessarily mimic all reactions that would occur *in vivo*.

175. *Id.* at 10,376.

176. In *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989), *cert. denied*, 494 U.S. 1046 (1990), the plaintiff's evidence included limb bud tests as evidence of teratogenicity of doxylamine, the active ingredient of Bendectin.

177. See *id.* at 314 (discussing the possibility that doxylamine breaks down in the human body and does not reach limb buds in unaltered form).

Observations based on such reasoning are the impetus of much drug research, as well as research on the hazardous effects of chemicals.¹⁷⁷ The relation of chemical structure to biological activity is highly uncertain, however, at least where the effects of only one or two compounds similar to the one in question are known.¹⁷⁸ No two chemicals have the same structure, so the question is always whether the similarities are more important than the differences in predicting toxicological properties. Structure-activity analysis is used primarily to select candidates for additional study. In most cases, reasoning based on structure-activity relationships will fall far short of the reliability required to satisfy a more probable than not standard of proof.

e. The Insufficiency of Animal Test Results, Short-term Assays,
In Vitro Testing and Structure-Activity Relationships to Prove
General Causation.

As can be seen from the foregoing discussion, animal test results, biological screening methods and chemical structure-activity relationships are insufficiently reliable, even if arguably valid, to permit inferences to be drawn that it is more probable than not that a substance can cause a disease in humans. It is also important to understand that even when all such indicators are positive, they are still insufficient for that purpose, given the present state of science.

Toxicological research into the causes of human disease, when direct evidence in humans is unavailable, proceeds according to a hierarchy of reasoning, from the least costly and time-consuming, and least predictive methods, to the most costly, time-consuming methods

177. See Williams & Weisburger, *supra* note 156, at 156-57 (discussing chemical structure-activity analysis as the first step in carcinogenicity assessment).

178. Saffiotti summed up the state of predictions from structure-activity analysis as follows:

There is a moderately substantial base of empirical data that permits conclusions about carcinogenic potential on the basis of molecular structure, *at least on the basis that* certain groupings of atoms (functional groups) in some molecules may impart carcinogenic properties. *The predictive power of such correlations has, however, been unsatisfactory so far, and the general consensus of the scientific community appears to be that chemical structure has limited value in identifying carcinogens and is to be used in carcinogenic hazard assessment only as corroborative supporting evidence.*

Umberto Saffiotti, *Identification and Definition of Chemical Carcinogens: Review of Criteria and Research Needs*, 6 J. TOXICOLOGY & ENVTL. HEALTH 1029, 1043 (1980).

It is unusual for all chemicals in a class to be carcinogenic, or for all carcinogenic members of a class to be equally potent. When 60 structural analogs of thalidomide were studied for teratogenicity, only three were found to exhibit that property. Manson & Wise, *supra* note 146, at 228. Structure-activity relationships are used in toxicology research primarily to select candidates for short-term assays and other more extensive tests. See, e.g., Williams & Weisburger, *supra* note 156, at 156-67.

that are believed to correspond most closely to human response. Thus, the investigation of the toxicological properties of a chemical is likely to start with the analysis of available information about chemicals with similar structures—chemical structure-activity analysis.¹⁷⁹ The most likely candidates identified by structure-activity analyses are then subjected to biological assays such as mutagenicity testing or *in vitro* testing on cell groups. Lastly, animal testing will likely be conducted on chemicals that exhibit toxic effects in the short-term screening procedures. Structure-activity analysis and short-term screening are not the end points of the evaluation process, even in a regulatory context, because they are recognized as significantly less valid and reliable than animal testing. Whether considered separately or in the aggregate, the methods that do not involve observations of disease in humans are too likely to lead to an erroneous conclusion to satisfy the traditional burden of proof.

2. CAUSATION OF PLAINTIFF'S DISEASE (INDIVIDUAL CAUSATION)

Even where the capability of a substance to cause disease can be shown, the plaintiff will still need to prove that the toxic substance caused his disease. The primary difficulty facing such a plaintiff is the need to differentiate between the exposure and background causes as explanations of the injury. Much speculation masquerading as science appears in connection with this issue.

To understand what kinds of evidence are probative of individual causation, we must first make reference to the kinds of evidence probative of the capability of the substance to cause harm, namely, epidemiologic evidence or possibly other human evidence of sufficient reliability. That evidence will identify one or more diseases that are believed to be causally associated with exposure to a toxic substance. Such studies will also typically be based on or identify certain levels or ranges of exposures. The plaintiff must argue that the association established through the statistical study applies to him and that background and other risk factors are less likely causal explanations.

a. Epidemiologic Reasoning

The best case for the plaintiff is the situation in which an epidemiologic study has identified an association between exposure to a toxic substance and a disease. For example, a number of studies have shown an association between asbestos exposure and lung cancer. The study will produce an estimate of the increased incidence of disease

179. See Williams & Weisburger, *supra* note 156, at 156-57 (describing the decision point approach to carcinogen testing).

associated with the exposure, such as relative risk.¹⁸⁰ Relative risk, illustrated by the formula below, is the ratio of disease incidence in the exposed population to disease incidence in the unexposed population in the study.¹⁸¹

$$\text{Relative Risk} = \frac{\text{incidence in exposed group}}{\text{incidence in unexposed group}}$$

Black provides an example in which the disease rates in exposed and unexposed groups are 50 and 5 per 100,000 population respectively.¹⁸² The relative risk in that case is 50/5 or 10, indicating that the exposure increases the disease rate to ten times that of the background rate.

How can a toxic tort plaintiff use such information? At a minimum, it would seem obvious that the plaintiff must have a disease identified as associated with the toxic chemical exposure. Nonetheless, it is not uncommon for plaintiffs' experts to assert that evidence that a substance causes any cancer is evidence that it can and has caused other cancers.¹⁸³ Although substances that are discovered to cause one type of cancer may cause other types of cancer as well, that possibility does not permit a prediction of what those other cancers, if any, are likely to be.¹⁸⁴ The

180. See Black & Lilienfeld, *supra* note 37, at 758.

181. *Id.* at 758 & n.105. Relative risk estimates can also be generated from case control studies, which compare the incidence of exposure in cases and controls that do not exhibit the disease in question. See ROTHMAN, *supra* note 44, at 63-64.

182. See Black & Lilienfeld, *supra* note 37, at 758 & n.105.

183. In *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988), the opt-out plaintiffs alleged, based in part on animal tests, that Agent Orange exposure had caused a number of different kinds of cancer, including Hodgkin's disease and cancer of the ileum, chronic skin rashes, infertility, *id.* at 1239, 1252-54, gastrointestinal disorders, miscarriages and other birth defects, *id.* at 1231, various behavioral disorders, including memory loss, increased irritability, anger, depression and others, weight loss, various liver disorders, including abnormal liver function, hepatitis and cirrhosis, and elevated triglycerides and cholesterol, *id.* at 1235-36. Plaintiffs relied on animal and workplace exposure studies to link Agent Orange or its contaminant, dioxin, to their injuries. *Id.* at 1236. The court noted, however, that plaintiffs' liver injuries differed "substantially" from those reported in the studies. *Id.* at 1236. Plaintiffs' reports of skin rashes or chloracne many years after exposure were also inconsistent with the immediate and transient relationship of chloracne and dioxin exposure. *Id.* at 1260.

Renaud v. Martin Marietta Corp., 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992), also typifies such cases. Plaintiffs alleged that a number of injuries, including childhood cancers (one case of leukemia), kidney cancer, seizure disorders, and congenital heart defects resulted from exposure primarily to hydrazines, also classified as animal carcinogens. *Id.* at 1547. Plaintiff also alleged exposure to several other chemicals. *Id.*

184. Nor does it suggest what the relative risk ratios would be for this possible, but unproven disease. This question is thus one of both general and individual causation.

problem is compounded when the plaintiff's general causation evidence is not based on human evidence, but on animal studies or other less reliable methods that only generally suggest a possible carcinogenic effect;¹⁸⁵ in such cases, the evidence may not allow an identification of any particular disease associated with the substance.

If plaintiff has a disease associated with the toxic substance exposure, demonstrating that it is more likely than not that plaintiff's condition was caused by exposure usually will require the plaintiff to demonstrate that her exposure was in the range found to be associated with an increased risk of disease.¹⁸⁶ Alternatively, plaintiff might be able to show that the results of the study could be extrapolated to lower doses.¹⁸⁷

If the plaintiff can demonstrate sufficient exposure to argue that the relative risk factor in the study applies to him, he still must differentiate background causes where the disease occurs in the background population. Under the traditional more likely than not rule, the plaintiff will prevail if the relative risk identified in the epidemiologic study is greater than two. That is because relative rate greater than two corresponds to more than doubling of the disease rate, permitting the inference that more than half of the cases of the disease in the exposed population were caused by the exposure. When applied to the plaintiff, the inference can be made that it is more likely than not that the exposure caused the

185. Such was the problem in *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991), discussed *infra* text accompanying notes 250-85. See also *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733 (N.J. 1991), discussed *infra* text accompanying notes 287-300.

186. Exposures are often at issue. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). Much of the evidence at issue in *Paoli* concerned whether plaintiffs' PCB exposures had exceeded normal background levels in the general population. See *id.* at 860-61; see also *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1555 (D. Colo. 1990) (dismissing suit because of plaintiffs' inability to present evidence of exposure establishing a *prima facie* case), aff'd, 972 F.2d 304 (10th Cir. 1992). Proof of exposure is discussed *infra* notes 214-49 and accompanying text.

187. It is widely accepted that dose-response relationships exist for toxic substances, that is, that disease incidence increases with increased exposure. The existence of a dose-response relationship is considered to be evidence that the association of toxic substance and disease incidence is causal. See *Black & Lilienfeld*, *supra* note 37, at 762-63 (discussing the Henle-Koch-Evans postulates).

High-dose to low-dose extrapolations based on epidemiologic data are subject to some of the same limitations as those discussed in connection with animal studies. See *supra* notes 151-59 and accompanying text. At some point, the application of epidemiologic studies to persons whose exposure levels were very different from those in the study raises a validity issue. This issue arises in cases where plaintiffs offer epidemiologic evidence based on relatively high occupational exposures as probative of the effects of lower level environmental exposures. See, e.g., *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1241 (E.D.N.Y. 1985) (rejecting epidemiologic studies based on industrial exposures), aff'd, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988).

plaintiff's disease.¹⁸⁸ Put another way, when the relative risk is greater than two, the fraction of all disease in the exposed group attributable to the exposure is greater than 50%.¹⁸⁹ The causal connection inferred from the presence of a signature disease represents the application of this principle when the relative risk is very large and the corresponding risk attributable to the exposure may be ninety percent or more.¹⁹⁰

b. Mechanistic Explanation

When the plaintiff's exposure evidence is weak, and when general causation evidence is based on animal studies, *in vitro* testing and the like, experts may attempt to bolster the plaintiff's case through speculation in the guise of a mechanistic explanation of causation.¹⁹¹ For example, an expert may offer testimony about the "one hit" theory of causation, namely, the theory that cancer can be caused by only one molecule of the

188. See Black & Lilienfeld, *supra* note 37, at 767-69. Relative risks greater than two are the basis of causation findings in cases involving claims for lung cancer from asbestos exposure. See *infra* notes 212-15 and accompanying text. The principle has been cited with seeming approval in a number of cases. See, e.g., DeLuca v. Merrell Dow Pharmaceuticals, Inc., 911 F.2d 941, 958-59 (3d Cir. 1990).

189. Attributable risk can be viewed as the proportion of a disease that is statistically attributable to a risk factor. Black & Lilienfeld, *supra* note 37, at 760-61. The attributable risk in an exposed population can be calculated as follows:

$$\text{Attributable Risk} = (\text{Relative Risk} - 1) / (\text{Relative Risk})$$

Where the relative risk is 2.0, attributable risk or attributable fraction is $(2.0-1.0)/2.0$, or 0.5. See Black & Lilienfeld, *supra* note 37, at 761 & n.123.

The reasoning illustrated in the text accompanying this note can result in recovery by 100% of exposed persons with the disease where the relative risk is greater than two, even though up to 50% would almost certainly have contracted the disease without exposure. Further, when there is a relative risk less than 2.0 but greater than 1.0, no plaintiffs will recover even though the epidemiologic study indicates causation of a group constituting less than half the cases. The unfairness of such results has led commentators to suggest that when an epidemiologic study indicates any increased risk, all exposed individuals with the indicated disease should recover proportionately to the magnitude of the increased risk. See, e.g., David Rosenberg, *supra* note 32; cf. Robinson, *Probabilistic Causation*, *supra* note 50, at 783 (recommending compensation for risk of future disease); Gregory L. Ash, Note, *Toxic Torts and Latent Diseases: The Case for an Increased Risk Cause of Action*, 38 KAN. L. REV. 1087, 1102-03 (1990).

190. For example, the relative risk of mesothelioma for asbestos exposure is on the order of 46, see Brennan, *supra* note 25, at 39 n.166 (citing A.D. McDonald & J.C. McDonald, *Malignant Mesothelioma in North America*, 46 CANCER 1650 (1980)), resulting in an attributable risk of over 97%.

191. In Brock v. Merrell Dow Pharmaceuticals, Inc., 874 F.2d 307 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989), *cert. denied*, 494 U.S. 1046 (1990), Dr. McBride, one of plaintiff's experts on the issue of whether Bendectin causes limb defects, testified about his theory of how Bendectin could cause such defects. *Id.* at 314-15. The court characterized the doctor's theory of causation as "nothing more than unproven medical speculation lacking any sort of consensus." *Id.*

toxic substance acting on only one cell, which then becomes cancerous.¹⁹² The witness may then explain that cancerous changes are thought to involve chemical alteration of DNA, the genetic material of cells, alterations that can be brought about by interaction with toxic chemicals.¹⁹³ Sometimes the witness expounds on the one hit theory by explaining that even very low concentrations of toxic chemicals involve exposure to trillions of molecules, and thus many opportunities for cancerous or mutagenic changes.¹⁹⁴ This kind of argument has much superficial appeal because it represents a common line of reasoning in cancer research or risk assessment.¹⁹⁵ When offered in proof of the likelihood that a low-level exposure, rather than background factors, caused the plaintiff's disease, however, the one hit theory and its corollaries amount to nothing more than speculation. The validity of the theory in a given case will rarely have been tested. Moreover, unless a mechanistic explanation offers a way to distinguish between causation by the toxic substance and causation by background factors,¹⁹⁶ it adds nothing to the proof of the plaintiff's case. Mechanistic explanations are not a substitute for statistical information such as epidemiologic studies when the background rate of the plaintiff's disease is significant.

192. See, e.g., *Peteet v. Dow Chem. Co.*, 868 F.2d 1428, 1433 (5th Cir.), *cert. denied*, 493 U.S. 935 (1989). A variant of this line of reasoning, mechanistically related to it, is the assertion that there is no safe level of a carcinogen, that is, that there is no threshold dose below which cancer does not occur. *See supra* notes 157-59 and accompanying text; *see also* *Sterling v. Velsicol Chem. Corp.*, 647 F. Supp. 303, 482-83 (W.D. Tenn. 1986), *aff'd in part, rev'd in part*, 855 F.2d 1188 (6th Cir. 1988); *Rubanick v. Witco Chem. Corp.*, 576 A.2d 4, 12 (N.J. Super. Ct. App. Div. 1990), *modified*, 593 A.2d 733 (N.J. 1991);. The U.S. Supreme Court discussed the one-hit theory and its relation to the issue of threshold exposures in *Industrial Union Department, AFL-CIO v. American Petroleum Institute*, 488 U.S. 607, 636 & n.41 (1979).

193. See, e.g., 2 Transcript of Hearing at 191-95 (testimony of Dr. Marvin Legator), *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (Civ. A. No. 87-Z-42), *aff'd*, 972 F.2d 304 (10th Cir. 1992); *id.* at 275-79 (testimony of Dr. David Ozonoff).

194. *Id.*

195. See generally OS&TP, Chemical Carcinogens, *supra* note 141, at 10,387-88, 10,438; *see also* *supra* notes 157-59. Cancer researchers recognize that DNA alteration represents only one mode of chemical carcinogenesis. Some carcinogens have been demonstrated not to react with DNA. These are generally grouped under the general heading of epigenetic carcinogens, which include promoters, that is, agents that facilitate the growth of dormant cancer cells into tumors. *See Williams & Weisburger, supra* note 156, at 185-86. These carcinogens typically require high doses and sustained exposure to exhibit carcinogenicity. *Id.* at 185.

The "one hit" theory is consistent with the linear extrapolation model or no-threshold model discussed *supra* notes 157-59 and accompanying text. The appeal of this kind of reasoning is also fundamentally related to the desire to understand *how* causation occurs.

196. Nor is even a proven mechanism likely to provide much information on the likelihood that an exposure caused disease unless the mechanistic explanation leads to a clinical test capable of distinguishing toxic chemical exposure from background or other causes.

c. Medical Opinion Evidence

Plaintiffs often offer medical opinion evidence on individual causation; indeed, courts sometimes express a preference for testimony by a treating or examining physician.¹⁹⁷ Such testimony may be essential to establish the diagnosis of the plaintiff's disease, his medical history, and the presence or absence of other possible risk factors for the disease.¹⁹⁸ The problems with medical opinion evidence arise when epidemiologic evidence of general causation is weak or absent¹⁹⁹ and the evidence of the capability of a substance to cause harm consists only of animal studies, mutagenicity testing, *in vitro* studies, or chemical structure-activity relationships.²⁰⁰ Those kinds of evidence are a weak basis for concluding that the toxic substance causes any human disease at all.²⁰¹ They are an extremely uncertain basis for making quantitative predictions that would allow a comparison of exposure risks and background risks. Indeed, the cases that have approved such evidence as sufficient to support a plaintiff's verdict have tended to ignore the absence of evidence²⁰² that would permit the plaintiff to distinguish background risks.

Plaintiffs who lack an epidemiologic basis for their proof of causation nonetheless frequently offer medical testimony from a treating physician or other expert who opines that the plaintiff's disease was caused by toxic substance exposure.²⁰³ This form of opinion is evident in

197. See *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984); *Wells v. Ortho Pharmaceutical Corp.*, 615 F. Supp. 262, 272-73 (N.D. Ga. 1985), *aff'd in part, modified in part*, 788 F.2d 741 (11th Cir.), *cert. denied*, 479 U.S. 950 (1986); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1235 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

198. See *infra* notes 206-12 and accompanying text.

199. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 862 (3d Cir. 1990) (opinion evidence based on, *inter alia*, animal tests and industrial exposures should not have been excluded), *cert. denied*, 111 S. Ct. 1584 (1991); *Ferebee*, 736 F.2d at 1535 (treating physicians testified).

200. See *supra* note 179 and accompanying text.

201. *Id.*

202. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d at 862 (approving proffered medical causation evidence). Such evidence is not a valid means of distinguishing other causal explanations when the causes of a majority of background cases are unknown. See *Rubanick v. Witco Chem. Co.*, 593 A.2d 733, 735-36 (N.J. 1991) (animal studies of PCBs, personal history, and higher than expected incidence of cancer at place of work admissible on causation of colon cancer); *infra* notes 264-76, 293-300 and accompanying text.

203. See, e.g., *Eggar v. Burlington N.R.R.*, No. CV89-159-BLG-JFB, 1991 U.S. Dist. LEXIS 19240 (D. Mont. Dec. 18, 1991); *In re Joint E. and S. Dist. Asbestos Litig. (Maiorana)*, 758 F. Supp. 199 (S.D.N.Y. 1991), *rev'd*, 964 F.2d 92 (2d Cir. 1992); *Wells v. Ortho Pharmaceutical Corp.*, 615 F. Supp. 262 (N.D. Ga. 1985), *aff'd in part, modified in part*, 788 F.2d 741 (11th Cir.), *cert. denied*, 479 U.S. 950 (1986); *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992).

The problems associated with medical testimony or disease causation have been discussed at length in Black, *supra* note 15, at 659-81. Depending on the circumstances, courts or the parties may espouse the view that a physician's testimony is preferred on the

*Renaud v. Martin Marietta Corp.*²⁰⁴ and other recent cases. When there is no clinical test that establishes a cause or distinguishes among possible causes, however, such "intuition" can only be characterized as speculation. This kind of speculation could easily be unmasked by inquiring into the reasoning behind the witness's opinion.²⁰⁵

d. Differential Diagnosis

The problem of distinguishing other causes takes a somewhat different form when there are other known risk factors for the plaintiff's disease. The effort to distinguish and eliminate other known risk factors is sometimes called "differential diagnosis."²⁰⁶ Diseases such as cancer that can result from toxic chemical exposure may also be associated with other identified risk factors, such as smoking, diet, lifestyle, as well as undifferentiated background risk associated with radiation and biological processes. Thus, the obvious question, particularly when the plaintiff does not exhibit other known risk factors such as diet, smoking, or a family history of the same cancer, is whether the absence of other risk factors increases the likelihood that the plaintiff's disease was caused by exposure to the toxic substance.²⁰⁷

issue of causation. The District of Columbia Circuit affirmed a jury's finding of liability in *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529, 1535-36 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984), based on the testimony of treating physicians. Plaintiffs often offer such testimony.

Interestingly, defendants sometimes object when a treating physician does not testify on causation. *See, e.g., Landrigan*, 605 A.2d at 1083 (trial court ruled that an epidemiologist could not testify on individual causation, nor could a nontreating physician offer an opinion based on epidemiologic evidence). Medical evidence, including testimony of a treating physician, may provide evidence of diagnosis of plaintiff's disease or injury or of the existence of other risk factors for the disease. *See id.*

204. 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

205. *See Ferebee v. Chevron Chem. Corp.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984); *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733 (N.J. 1991). The *Ferebee* court based its affirmation of the plaintiff's verdict on the treating physicians' use of "tissue samples, standard tests, and patient examination." 736 F.2d at 1536. The court did not appear to consider whether those tests were in any way indicative of the cause of *Ferebee's* condition.

206. This usage of the term seems to be a misnomer. STEDMAN'S MEDICAL DICTIONARY (25th ed. 1990), defines differential diagnosis as "the determination of which of two or more diseases with similar symptoms is the one from which the patient is suffering, by a systematic comparison and contrasting of the clinical findings." *Id.* at 428. In toxic torts, the term is applied to the determination of which of two or more factors caused the plaintiff's disease, the diagnosis of which is not in question.

207. "Differential diagnosis" is an argument that cuts both ways. Plaintiffs are likely to make a differential diagnosis argument that the toxic exposure is the likely cause of the plaintiff's disease because other known risk factors are absent. *See, e.g., Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992) (discussing risk factors for colon cancer). Defendants are likely to point out that plaintiff has failed to eliminate other known risk factors that are applicable to the plaintiff. *See, e.g., In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1253 (E.D.N.Y. 1985) (plaintiff's experts "fail to show how the myriad illnesses at issue are more likely to have been caused by Agent Orange than by something

Plaintiffs often argue and courts sometimes accept the notion that the absence of other risk factors increases the likelihood that the plaintiff's disease was caused by the toxic exposure at issue.²⁰⁸ The validity of that kind of reasoning, however, rests on two unstated, and usually untested, assumptions. First, such reasoning treats toxic exposure and the other risks as alternatives. In other words, it assumes that the disease was caused by the toxic exposure *or* some other cause, such as the other identified risk factors.²⁰⁹ Second, it assumes that most causes of the disease in question are known; otherwise, the elimination of other risk factors would not significantly increase the likelihood that the toxic exposure was the cause of the plaintiff's disease.

The assumption that risk factors, including the toxic exposure, represent alternative causes is true only if the various risks are additive. Additivity is only one of several ways in which risk factors for the same disease may relate. The combined effects may be the same, greater, or less than the sum of the effects as measured separately.²¹⁰ Additive effects represent the absence of interaction between risk factors.²¹¹ Thus, each factor adds an incremental level of risk to the background risk that is independent of the presence or absence of other risk factors. Additive

else"), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988). In actuality, the issue of other causes is present in every case that involves a disease with a significant background risk. Almost certainly, there are as yet unidentified risk factors that affect background disease incidence. See ROTHMAN, *supra* note 44, at 12. The assumption of uniform risk factors in the background population reflects ignorance of what those factors are. *Id.* Only with signature diseases such as mesothelioma and clear cell adenocarcinoma that are rare in the absence of exposure to an identified carcinogen can this issue be avoided. See *supra* note 190 and accompanying text.

208. See generally Landrigan v. Celotex Corp., 605 A.2d 1079, 1087-88 (N.J. 1992); Rubanick v. Witco Chem. Corp., 542 A.2d 975 (N.J. Super. Ct. Law Div. 1988), *rev'd*, 576 A.2d 4 (N.J. Super. Ct. App. Div. 1990), *modified*, 593 A.2d 733 (N.J. 1991).

209. Attribution of causation in an illness with several possible causes is often a complex task. As the reader will recall from the discussion at the beginning of this section, the relative risk data from epidemiologic studies can be used to determine the fractions of cases of a disease in an exposed population that are attributable to the exposure and to the background causes. See *supra* notes 187-90 and accompanying text.

210. See ROTHMAN, *supra* note 44, at 311-26. Combined effects that are less than additive are considered antagonistic, while combined risks that are greater than the sum of the separate effects are considered synergistic. *Id.* at 318-20.

211. *Id.* at 313-15. Rothman states that although some epidemiologists treat the foregoing definitional scheme as arbitrary, the use of an additivity assumption for independent action has practical consequences in interpreting and utilizing epidemiologic data. *Id.* at 316-17. For example, the combined relative risks of oral contraceptives and hypertension for thrombotic stroke are greater than the sum of the relative risks of each. *Id.* at 316. When these greater than additive risks are considered to be synergistic, the practical conclusion is that a woman should consider her blood pressure history in deciding whether to use oral contraceptives, a result that seems intuitively correct. *Id.* at 316-17.

risks are properly treated as alternative risks in a causation analysis.²¹² Often, however, the information necessary to support that assumption is not available.

Risk factors whose combined effects are greater than additive are considered interactive or synergistic.²¹³ In this situation, each risk factor enhances the risk contributed by the other factor so that the total incidence of the disease is greater than the sum of the incidence attributable to each factor separately, sometimes approaching a multiplicative effect. Perhaps surprisingly, the presence or absence of other risk factors that are multiplicative does not increase or decrease the fraction of disease attributable to the toxic exposure. Thus, when the causes are synergistic, as with smoking and asbestos and lung cancer,²¹⁴ it is incorrect to pose the question as one of whether the disease was caused by one factor or another.²¹⁵

212. For example, smoking increases the incidence of lung cancer by a factor of about 10 and occupational exposure to asbestos increases risk by about a factor of 5. U.S. SURGEON GEN., *supra* note 47, at 216-17. If the background incidence of lung cancer among nonsmokers who do not have occupational exposures to asbestos is normalized to 1.0, nonsmoking asbestos workers would have a lung cancer incidence of 5.0, while smokers would have a lung cancer rate incidence of 10.0. For smoking asbestos workers, if the risks were noninteractive (and therefore additive), the lung cancer incidence would be about 14, which is the sum of the background case, the four cases added by asbestos exposure, and the nine cases expected to be added as a result of smoking. In this situation, smoking and asbestos would represent alternative causes, since any one plaintiff's case would probably be caused by one factor or the other, not by both acting together. Thus, while a nonsmoking asbestos worker could argue that the probability that asbestos caused his cancer was 80% (4/5), the smoking asbestos worker could point to only a 28% probability (4 cases out of 14 total) that his lung cancer was caused by asbestos (the other 10 cases being the result of background causes or cigarette smoking).

213. ROTHMAN, *supra* note 44 at 315-16.

214. *Id.* at 312.

215. Because the effects of smoking and asbestos are multiplicative for lung cancer, the population of smoking asbestos workers described in *supra* note 212 is expected to have lung cancer incidence of 5 times 10, or 50, rather than the 15 cases predicted by adding the separate risks. See U.S. SURGEON GEN., U.S. DEP'T OF HEALTH & HUMAN SERVS., *supra* note 47, at 216-17. The fraction attributable to asbestos in this case is 40/50 (or 0.8, subtracting the 10 cases that would have occurred as the result of smoking alone or without exposure to either factor from the 50 total cases), the same attributable fraction obtained when the effects of asbestos alone are considered. Counterintuitively, the fraction of lung cancer cases among smoking asbestos workers for which smoking can be considered causative is 45/50 (or 0.9, obtained by subtracting the 5 cases that would have occurred due to asbestos alone or in the absence of either exposure, from the total cases). Those results do not mean, however, that arguments cannot be made that the plaintiff's smoking constituted contributory negligence or an intervening cause that should reduce the asbestos manufacturer's liability or eliminate it altogether.

Defendants sometimes make such arguments. See, e.g., *In re Brooklyn Navy Yard Asbestos Litig.* (Joint E. & S. Dist. Asbestos Litig.), 971 F.2d 831 (2d Cir. 1992); *In re Manguno* (Acosta v. Babcock & Wilcox), 961 F.2d 533 (5th Cir. 1992); *Borman v. Raymark Indus.*, 960 F.2d 327 (3d Cir. 1992) (upholding plaintiff's verdict and affirming trial court's refusal to charge the jury on apportionment of damages between smoking and asbestos). In *Manguno*, the jury returned a verdict for defendant asbestos manufacturers. *Id.* at 534.

There are several scenarios under which this issue could arise, but the actual cases tend to be grouped into two extremes. Where epidemiologic data are available that address the contributions of both the toxic substance and other causal factors, the plaintiff's attributable risk and probability of causation by the toxic substance can be determined by calculating attributable fractions, whether the risks are additive, multiplicative or antagonistic. From those calculations, it can be determined whether the plaintiff can satisfy the more likely than not standard of proof on causation.²¹⁶

The other extreme is represented by toxic tort cases where multiple risk factors are treated in a vague or qualitative fashion and data are not available to support a quantitative analysis. The plaintiff's expert may opine that because other known risk factors are absent in plaintiff's case, it is the expert's opinion that plaintiff's disease was caused by the toxic substance exposure.²¹⁷ This argument has superficial appeal. It can only be valid, however, when risk factors that account for most cases of the plaintiff's injury and their interactions are understood. Although some risk factors for cancers and birth defects have been identified, the causes of background incidence of most birth defects and cancers remain unknown. Even if the identified risk factors are alternative, independent causes, as this line of analysis assumes, the expert's opinion distinguishes among factors that make up only a small part of the total picture, while ignoring the probability that the plaintiff's injury may stem from the same unidentified factors that are responsible for most cases of the injury. Whether the differential diagnosis argument is made on behalf of plaintiff or defendant, it adds little to the resolution of the case when it is based on vague, qualitative assumptions about alternative causes.²¹⁸

The defendants argued, and the jury apparently accepted, that plaintiff's smoking cast sufficient doubt on the proposition that asbestos caused the plaintiff's lung cancer. The Fifth Circuit reversed on the basis of improper jury instructions and remanded for a new trial. *Id.* at 535-36.

216. Often there will not be sufficient data to determine whether the risk factors are additive, synergistic, or antagonistic. In such cases the point is arguable, but the better approach would seem to be to ignore other risk factors whose interactions with the toxic substance exposure are unknown and assume that the relative risk associated with the toxic substance exposure applies whether the other risk factors are present or not.

217. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 735-36 (N.J. 1991).

218. The defendant's argument that the existence of known risk factors dilutes the likelihood of causation by the chemical exposure has a point, however, when the plaintiff's general causation evidence is based on medical opinion, animal studies, cellular assays and other such evidence. In *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 839-40 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991), the plaintiffs claimed that a variety of commonplace ailments were due to PCB exposure. Available epidemiologic data failed to demonstrate any connection between PCBs and the claimed physical injuries. The defendants' objection that plaintiffs failed to rule out the known causes of such commonplace ailments as high cholesterol and high blood pressure should have been well taken by the court. See *id.* at 861-62.

3. PROOF OF EXPOSURE

a. Inferences from Similar Circumstances.

While the foregoing discussion has focused on proving the harmful effects of exposure, the issues of whether an exposure occurred and if so, of what magnitude, are often present in toxic tort cases.²¹⁹ In several recent cases, the duration and magnitude of the plaintiffs' exposure was subject to a great deal of uncertainty.²²⁰ How can the plaintiff prove exposure? In instances where the plaintiff has a signature disease, the disease itself constitutes strong evidence of exposure because it rarely occurs in the absence of exposure.²²¹ More commonly, however, the plaintiff's injury is not so distinctive that it can be reliably attributed to a toxic substance exposure, even where the substance is known to increase the risk of the condition. Where the claimed injury is one that is not attributable almost solely to a toxic substance, exposure data is meaningful only if it is quantitative. The plaintiff must prove the magnitude of her exposure to a degree of certainty that supports the inference that the evidence linking a toxic chemical to disease is applicable to her.

Some cases, particularly those involving workplace exposures, involve situations that are known to involve exposure to a toxic substance.²²² The severity of the plaintiff's exposure can be inferred from the length of time he was present in the environment. The situation would be similar where there is an ongoing exposure, such as a drinking

219. See, e.g., *Paoli R.R. Yard*, 916 F.2d 829 (evidentiary issue of the degree of plaintiffs' exposure to PCBs), cert. denied, 111 S. Ct. 1584 (1991); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (case dismissed because of plaintiffs' inability to make out a prima facie case of exposure), aff'd, 972 F.2d 304 (10th Cir. 1992).

220. See *Christopherson v. Allied-Signal Corp.*, 939 F.2d 1106 (5th Cir. 1991) (en banc), cert. denied, 112 S. Ct. 1280 (1992). In *Christopherson*, plaintiff's husband worked in a plant that produced nickel/cadmium batteries. *Id.* at 1108. He did not work in the production area, however, but visited the area intermittently. *Id.* There also appears to have been no direct evidence on the nature of fumes to which Christopherson was exposed during those visits. *Id.* at 1113. Apparently a fellow employee's affidavit alleged that Christopherson was exposed to airborne particles of nickel and cadmium, but it is not clear that the employee could have known the chemical composition of the fumes. See *Christopherson v. Allied-Signal Corp.*, 902 F.2d 362 (5th Cir. 1990), rev'd, 939 F.2d 1106 (5th Cir. 1991) (en banc), cert. denied, 112 S. Ct. 1280 (1992).

221. In signature disease cases, the existence of the disease may suffice to prove that sufficient exposure occurred and that the exposure caused the plaintiff's injury. Cf. *Renaud*, 749 F. Supp. at 1553 (plaintiffs could have attempted to prove exposure through epidemiologic study).

222. Occupational exposure of insulation installers to asbestos is an example of that situation. See, e.g., *In re Joint E. & S. Dist. Asbestos Litig.* (Johns-Manville Corp.), 129 B.R. 710, 868 (E.D. & S.D.N.Y. 1991) (Weinstein, J.) (discussing proof of asbestos exposure in occupational and other settings), vacated, 982 F.2d 721 (2d Cir. 1992).

water exposure, that can be measured. In such cases, inferences about past conditions can be drawn from the present ones.²²³

Sometimes there is clinical evidence of the toxic substance that will suffice to prove the plaintiff's exposure.²²⁴ Substances such as asbestos²²⁵ and PCBs²²⁶ remain in tissues indefinitely, and can be detected by appropriate analytical tests. Other substances may result in subclinical changes that can be detected through appropriate testing.²²⁷

b. Inferences from Modeling

The reasoning underlying the proofs of exposure outlined above is readily apparent and is of the kind already familiar to courts in other contexts. Cases involving possible past exposures where contemporaneous measurements cannot be taken or inferred, and for which there are no available analytical tests, present a much more difficult case. In such cases, experts may use various models to estimate exposure. Models are mathematical formulas that are designed to provide estimates of facts that cannot be measured directly.²²⁸ They range from simple formulas such as the model discussed below that purports to describe the ratio of blood or serum PCB levels to adipose tissue levels, to complex computer programs used to model groundwater contaminant migration and air pollutant dispersion.

Models, at least conceptually, sometimes begin with theories or hypotheses about how different kinds of data might be related. A scientist would be unlikely to propose a model that was not at least plausible, based on her understanding of how the phenomena in question

223. See, e.g., *Sterling v. Velsicol Chem. Corp.*, 855 F.2d 1188 (6th Cir. 1988) (chlorinated hydrocarbons detected in 12 to 15 drinking water wells).

224. Clinical evidence will not necessarily be sufficient to prove that the exposure caused injury, however, because exposure does not always result in disease. See *supra* note 47 and accompanying text.

225. See, e.g., *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992) (clinical data, such as asbestos in or near a tumor, may support a finding of specific causation).

226. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 841 & n.10 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). Dioxin is also retained in fat. Brennan, *supra* note 25, at 51 (citing Peter C. Kahn et al., *Dioxins and Dibenzofurans in Blood and Adipose Tissue of Agent Orange-Exposed Vietnam Veterans and Matched Controls*, 259 JAMA 1661 (1988)).

227. See Robert R. Lauwerys, *Occupational Toxicology*, in CASARETT & DOULL, *supra* note 148, 947, 954-66; see also OS&TP, *Chemical Carcinogens*, *supra* note 141, at 10,386, 10,441 (discussing DNA alteration and relation to carcinogenesis); Brennan, *supra* note 15, at 502 n.174. Clinical testing for exposure-induced damage may become more important as knowledge accumulates regarding molecular alterations from chemical exposures. See OS&TP, *Chemical Carcinogens*, *supra* note 141, at 10,441.

228. For a discussion of groundwater contaminant migration modeling, see Allen Kezsombi & Alan V. Goldman, *The Boundaries of Groundwater Modeling Under the Law: Standards for Excluding Speculative Expert Testimony*, 27 TORT & INS. L.J. 109 (1991). See also Itzchak E. Kornfeld, *Comment to the Boundaries of Groundwater Modeling Under the Law: Standards for Excluding Speculative Expert Testimony*, 28 TORT & INS. L.J. 59 (1993).

are connected. Plausibility alone, however, like the mechanistic theories of cancer causation, is often insufficient to eliminate other plausible but untested models or theories. Thus, a model must be validated before an expert or a court should assume that it has any probative value.

Validation of a model involves testing the model's predictive capability in circumstances where the expected results can be independently measured.²²⁹ In the case of the blood/adipose tissue partition model or the much more complex groundwater contaminant migration modeling, actual concentrations can be measured and compared with values predicted by the model. From such data, it can be determined whether the model has any predictive value and how accurate or reliable those predictions are. If the model proves valid and reasonably accurate, then it is reasonable to apply it to other situations similar to the one for which the model has been tested.

A number of cases, however, have involved modeling of exposures where the models have not been subjected to the most cursory validation, sometimes in the face of data that contradict the model. In *In re Paoli Railroad Yard PCB Litigation*,²³⁰ one of the ways the plaintiffs attempted to prove exposure to PCBs was by showing that their PCB levels were elevated above background levels. Because PCBs accumulate in fatty tissue and are not quickly eliminated from the body, it is possible to measure PCB levels in blood or tissue samples from individuals and compare them to norms for the general population.²³¹ The Agency for Toxic Substances and Disease Registry (ATSDR) had done a study on blood PCB levels in Paoli residents and concluded that the residents' serum PCB levels did not differ significantly from those of the general population.²³² Plaintiffs contended that the ATSDR study's conclusions regarding background PCB levels in the general population were erroneous. Their experts sought to show that the plaintiffs' adipose or fat PCB levels exceeded norms found in the Environmental Protection Agency's National Human Adipose Tissue Study.²³³ Because most of the plaintiffs' PCB levels were determined by blood tests alone, plaintiffs'

229. See Kezsbom & Goldman, *supra* note 228, at 117 (noting that the plaintiffs in *Sterling v. Velsicol* never verified their model against real-world data). Kornfeld, who has a contrary view of *Sterling v. Velsicol*, states that models must be scrutinized to determine whether they replicate real-world data and have been calibrated. See Kornfeld, *supra* note 228, at 68.

230. 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991).

231. See *id.* at 839-41. PCBs are polychlorinated biphenyls, once commonly used in electrical transformers.

232. Nor did they differ significantly among the residents according to higher soil concentrations in the residents' yards, number of years in the vicinity, or residence in more or less highly contaminated areas. *In re Paoli R.R. Yard PCB Litig.*, 706 F. Supp. 358, 364-65 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991).

233. *Id.* at 371. The study was conducted between 1970 and 1983.

expert used his own formula to calculate their adipose tissue PCB levels, which he then compared with results reported in the national study.²³⁴

As the trial court recognized, however, neither of the plaintiffs' experts on this issue²³⁵ cited any basis for the claimed relationship between PCB concentrations in blood and adipose tissue.²³⁶ Moreover, where blood and adipose tissue PCB levels were measured in the same plaintiffs, the results did not bear out the ratios asserted by plaintiffs' experts.²³⁷ Thus, although a relationship between blood and adipose tissue levels of PCBs is plausible, the direct ratio posited by plaintiffs' witnesses was not validated and was demonstrably inaccurate as indicated by comparison of predicted levels with actual measurements in plaintiffs who had both tests. Conclusions based on models that have not been validated by actual measurement,²³⁸ or worse, which are contradicted by actual measurements, are based on invalid reasoning and should be rejected by the courts.

Use of unvalidated models is not the only concern about models. Models inherently involve approximations—generalizations about physical phenomena and estimations that must be made because the actual situation cannot be studied directly.²³⁹ Those approximations inevitably introduce inaccuracies into the model's predictions. At some point, the uncertainty or inaccuracy may become so large that the model's results are too unreliable to prove the fact on which they are offered.

Modeling of groundwater and surface water contamination was at issue in *Renaud v. Martin Marietta Corp.*,²⁴⁰ in which the plaintiffs claimed that contamination of the public water supply had caused childhood cancers and other diseases.²⁴¹ Plaintiffs' case on exposure involved the issue of whether contaminants released at the Martin Marietta facility had reached their taps through the Denver culinary water distribution system. Because the circumstances that created the discharges at issue had changed in the years preceding the suit, the plaintiffs relied on

234. *Id.* at 370-71. The Third Circuit opinion implies that the NHATS study used the formula that relates blood and adipose tissue levels. *See Paoli*, 916 F.2d at 841 n.10.

235. Dr. Ian C.T. Nesbit, a Ph.D. physicist and consultant, and Dr. Robert K. Simon, an industrial hygienist, toxicologist, and forensic analytical chemist, testified on the same issue. *See Paoli*, 916 F.2d at 840, 847.

236. *Paoli*, 706 F. Supp. at 372..

237. *Id.*

238. In *Paoli*, one of plaintiffs' experts, Dr. Herbert Allen, purported to calculate the levels of airborne PCBs to which plaintiffs had been exposed based on soil PCB concentrations, using a formula, i.e. a model, he had devised. 916 F.2d at 839. Dr. Allen's predictions, however, were higher than the measurements actually taken. *See infra* text accompanying notes 257-59.

239. *See Kezsomb & Goldman, supra* note 228, at 109, 116-19.

240. 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

241. *Id.* at 1547.

hydrological modeling of ground and surface water movement as proof that contaminants had reached the water distribution system.²⁴²

Defendants argued that there were serious flaws in the modeling, most notably that plaintiffs had erred by failing to consider all relevant factors when deriving the decay coefficient for the contaminants.²⁴³ Further, the experts had not taken into account the possibility that chlorination at the water intake plant had destroyed or greatly reduced the concentration of contaminants.²⁴⁴ Although the court stated that “[t]he issues of which factors should have been considered and what impact each should have been given” were questions for the jury, it seems clear that the factors that plaintiffs’ experts ignored would have had a large impact on the concentrations predicted by the fate and transport modeling.²⁴⁵ Models are always subject to dispute over the factors that are included or excluded, and thus cannot be judged by too rigorous a standard. Nonetheless, where experts have excluded significant factors that would tend to produce results at odds with their conclusions, the court should exclude the modeling results unless there is some more direct way to demonstrate the model’s validity and accuracy.²⁴⁶

Another concern with modeling is that uncertain input data affect the reliability of the results of all kinds of exposure modeling. If the input data are very limited, there will be uncertainty about how representative those data are, and those uncertainties will produce corresponding uncertainties about the modeling results, no matter how good the model is. The greatest uncertainty about the exposure modeling in *Renaud* resulted from such a scarcity of data.²⁴⁷ In that case, the plaintiffs’ exposure estimate, obtained through the transport modeling described above, was based on a single loading concentration for the contaminants. The court recognized that the single data point on which the modeling was based could not be said to be representative of the 11-year period

242. *Id.* at 1549.

243. *Id.* at 1552. A decay coefficient was necessary because the chemicals at issue were known to undergo degradation in the environment. *See id.* at 1549.

244. *See Letter from Dr. Hannah Pavlik, Ebasco Environmental, to Judge Zita Weinshienk, U.S. District Court* (Aug. 29, 1990). Dr. Pavlik, a geochemist and hydrogeologist, was retained as a court-appointed expert witness. *Renaud*, 749 F. Supp. at 1553. Her report also indicated a number of other methodological and factual flaws in the hydrological modeling. *See Letter from Dr. Hannah Pavlik, supra*, at 9.

245. *Renaud*, 749 F. Supp. at 1552; *see Letter from Dr. Hannah Pavlik, supra* note 244, at 16.

246. The *Renaud* plaintiffs’ experts could have conducted experiments on the effects of chlorination on the chemicals in question. The only available information on that issue appeared to be the defendants’ own tests, however, which were consistent with their position. *See Letter from Dr. Hannah Pavlik, supra* note 244, at 9.

247. Data that is based on unsound assumptions, unverifiable assumptions, or erroneous input results in a manifestation of the “garbage in, garbage out” phenomenon. *See Kezsomb & Goldman, supra* note 228, at 116.

over which releases occurred.²⁴⁸ It found the single data point to be a fatal flaw in the plaintiffs' exposure case.²⁴⁹

V. DIVERGENCE OF OPINION

A. Deferential Review and the Accumulation of Errors

The foregoing Part has examined separately the problems associated with proof of exposure, capability of the substance to cause harm, and distinguishing among causes and concluded that the evidence deemed acceptable in many toxic tort cases is often grossly inadequate to prove the propositions on which it is offered. It is also important to examine how those issues are brought together in real cases, keeping in mind that the ultimate causation issue is whether exposure to a toxic substance caused the plaintiff's disease. This part discusses several recent cases that are particularly troubling when viewed as a whole, because the plaintiff's cases can, at best, be characterized as consisting of possibilities and speculation strung together in ways that fall far short of the legal requirements of proof.

The Third Circuit's decision in *In re Paoli Railroad Yard PCB Litigation*²⁵⁰ represents one of the most troubling decisions on the admissibility and sufficiency of challenged scientific evidence. *Paoli* involved an action by 38 neighbors and employees of an electric railcar maintenance facility contaminated by PCBs.²⁵¹ The action, which was brought against owners and operators of the site, and suppliers of PCBs and transformers, made claims for various injuries and for medical monitoring costs necessary to protect against latent disease.²⁵²

Defendants' motions to exclude plaintiffs' evidence of exposure to PCBs and other causation evidence, and for summary judgment were granted by the trial court.²⁵³ The Third Circuit, however, reversed, finding that the trial court had improperly excluded sufficient evidence to survive summary judgment.²⁵⁴ The case represents a virtually complete catalog of the unprobative and insufficient kinds of proof identified in this article.

Two factual issues on which scientific evidence was crucial were: (1) whether plaintiffs received any exposure above background levels of

248. *Renaud*, 749 F. Supp. at 1552-53.

249. *Id.*

250. 916 F.2d 829 (3d Cir. 1990), *rev'd* 706 F. Supp. 358 (E.D. Pa. 1988), *cert. denied* 111 S. Ct. 1584 (1991).

251. *Id.* at 832.

252. *Id.* at 849.

253. *Id.* at 835.

254. *Id.* at 862.

PCBs that could be attributed to the Paoli railyard; and (2) whether PCBs are capable of causing the ailments of which plaintiffs complained or of which they believed they were at risk.²⁵⁵ The primary problem with plaintiffs' evidence on exposure involves unvalidated methodology through which plaintiffs attempted to show that their exposure to PCBs exceeded background levels.

Plaintiffs offered several forms of evidence in addition to the blood/adipose tissue calculations discussed in the preceding Part,²⁵⁶ in support of their contentions of higher than background levels of PCB exposure. First, they offered the testimony of Dr. Herbert Allen, an environmental chemist who used a formula of his own devising to calculate airborne exposure levels from PCB levels measured in neighborhood soils.²⁵⁷ Nothing in either the district court's or the appellate court's opinions, however, suggests that Dr. Allen's formula had been tested, that is, that its validity had been demonstrated by comparing the airborne concentrations predicted by the formula and actual, measured air concentrations.²⁵⁸ In fact, the district court opinion states that the "actual measurements that were taken showed an amount much lower than [Allen] calculated."²⁵⁹

Plaintiffs also offered the testimony of Dr. Deborah Barsotti, a toxicologist employed by the Agency for Toxic Substances and Disease Registry, who claimed to have correlated gas chromatography tracings of PCBs in the plaintiff's blood with tracings from soil samples from the Paoli railyard.²⁶⁰ Dr. Barsotti, however, was apparently unable to support her general statements with reference to any specific plaintiffs' blood samples or soil samples.²⁶¹ Later, she apparently conceded that the equipment she had used was not capable of yielding the results she claimed.²⁶²

Plaintiffs' evidence on general causation and on distinguishing background causes was hardly more probative. On the question of whether PCBs are capable of causing the kinds of illnesses complained of, the plaintiffs were faced with various studies that had failed to find a correlation between PCB exposure and significant human disease. One such study was the ATSDR's study, *Toxicological Profile for Selected*

255. See *id.* at 860-62.

256. See *supra* notes 230-38 and accompanying text.

257. *Paoli*, 916 F.2d at 838-39.

258. *Id.* at 839, 842; *In re Paoli R.R. Yard PCB Litig.*, 706 F. Supp. 358, 370 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), *cert. denied*, 111 S. Ct. 1584 (1991).

259. *Paoli*, 706 F. Supp. at 370.

260. *Paoli*, 916 F.2d at 839.

261. *Id.* at 842.

262. *Id.*

PCBs.²⁶³ As summarized in the Foreword, the ATSDR study found that only skin lesions and liver effects that were not associated with "clinically detectable disease" had been observed in PCB-exposed workers. The study also concluded that adverse effects had not been observed in persons with non-occupational exposures.²⁶⁴

The Third Circuit, however, found various plaintiffs' witnesses' testimony sufficient to create a jury question.²⁶⁵ Some of this testimony can only be described as conclusory.²⁶⁶ Several of plaintiffs' experts relied on animal studies and on studies involving incidents of accidental ingestion of a mixture of PCBs and PCDFs,²⁶⁷ the "Yusho" and "Yu Cheng" incidents that occurred respectively in Japan in 1968, and Taiwan in 1979.²⁶⁸ Lastly, the plaintiffs offered the testimony of Dr. William J. Nicholson, a professor of community medicine who had performed a "meta-analysis" of existing (negative) epidemiologic studies²⁶⁹ and concluded that his analysis showed that PCBs can cause liver, biliary tract and gall bladder disorders.²⁷⁰

The evidence on general and individual causation was clearly insufficient to permit the inference that would satisfy the more likely than not standard of proof. Animal studies, as discussed earlier, are at best subject to large uncertainties when extrapolated to humans, particularly for effects of chronic, low-level exposures.²⁷¹

The use of the Yusho and Yu Cheng studies were not challenged as invalid,²⁷² but their use for the purposes of proving that PCBs cause significant adverse effects in humans involves invalid reasoning.²⁷³ The

263. *Paoli*, 706 F. Supp at 365 (citing Toxicological Profile for Selected PCBs (draft Nov. 1987)).

264. *Id.*

265. *Paoli*, 916 F.2d at 862. Dr. Barsotti testified that exposure to PCBs at Paoli was a substantial factor in causing plaintiffs' elevated triglycerides, cholesterol, and liver enzyme levels. *Id.* at 839. According to the court, she based her conclusions on her inspection of the Paoli railyard and her review of various reports and studies, and of soil samples from the railyard. *Id.*

266. Dr. Barsotti's affidavits on causation for each plaintiff were identical for the first fourteen pages, with only two additional paragraphs listing the alleged injuries and concluding they were caused by PCBs. *Id.* at 842-43. Thus, she never explained her reasoning beyond the fact that plaintiffs were exposed to the railyard PCBs, studies have indicated possible injuries from PCBs, and plaintiffs have injuries.

267. PCDFs are polychlorinated dibenzofurans, chemically related to PCBs. *See id.* at 839.

268. *Id.* at 840.

269. *Id.* at 841. A meta-analysis combines the results of epidemiologic studies to increase the total sample size and reanalyzes the data. *Id.*

270. *Id.*

271. *See supra* notes 140-68 and accompanying text.

272. *See In re Paoli R.R. Yard PCB Litig.*, 706 F. Supp. 358, 366 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), *cert. denied*, 111 S. Ct. 1584 (1991).

273. *Id.* at 368.

studies identified harmful effects from two incidents involving the ingestion of oil mixtures containing PCBs and PCDFs. Assuming that the conclusions regarding the Yusho and Yu Cheng incidents were accurate, the results can at most be said to prove the following proposition: The harmful effects observed in the Yusho and Yu Cheng incidents were caused by either: (1) PCBs; (2) PCDFs; or (3) PCBs and PCDFs in combination.²⁷⁴ The studies do not allow the conclusion that PCBs alone can cause the effects observed in the study. Additionally, because PCDFs are regarded as more toxic than PCBs,²⁷⁵ the proposition that plaintiffs argued from the study is not supported by the study.²⁷⁶

The meta-analysis of epidemiologic studies presents a somewhat different problem. Meta-analyses are not outside the scope of recognized scientific methodology.²⁷⁷ The defendants did raise questions, however, about the way in which Dr. Nicholson analyzed the existing data, contending that he omitted data that were inconsistent with his conclusions.²⁷⁸ Thus, the trial court could have examined the bases on which Dr. Nicholson included and excluded data to determine whether

274. This analysis assumes that there is a basis for concluding that the effects were not caused by other unnamed constituents.

275. See *In re Paoli R.R. Yard Litig.*, 916 F.2d 829, 839, 843 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). PCDFs are polychlorinated dibenzofurans. *Id.* at 839.

276. The Third Circuit noted that Dr. Herbert Allen testified that activities at the railyard such as welding and cutting contaminated equipment could have converted PCBs into dioxins and PCDFs. *Id.* at 839. Although the court characterized this testimony as "particularly significant." *Id.* It appears to offer only the most tenuous evidence of an unquantifiable possibility, since there is no indication that plaintiffs tested for or offered physical evidence of dioxin or PCDF contamination or exposure.

277. See *id.* at 857 (citing C. David Naylor, *Two Cheers for Meta-analysis: Problems and Opportunities in Aggregating Results of Clinical Trials*, 138 CAN. MED. ASS'N J. 891, 894 (1988)).

278. *Id.* at 845. The Third Circuit regarded the issue of how the meta-analysis was conducted as one of credibility, and therefore for the jury, although it qualified that conclusion with the statement that the meta-analysis would be excludable if no reasonable person could believe the study. *Id.* at 858. The district court appears to have excluded the meta-analysis primarily on relevance grounds, although it first discussed the standards for admitting novel scientific evidence in connection with the meta-analysis. *In re Paoli R.R. Yard Litig.*, 706 F. Supp. 358, 372-73 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). It deemed the study irrelevant because the diseases the study attributed to PCB exposure were not those claimed by plaintiffs. *Id.* at 373. The Third Circuit rejected that rationale, however, because it also held that plaintiffs were entitled to proceed under a medical monitoring claim; it deemed the meta-analysis relevant to a determination of the risks of future disease to which the plaintiffs were exposed.

The issue of when evidence based on a technique that may have been improperly applied may be excluded, as opposed to submitted to the jury, is a current controversy in evidence law. If, however, it was apparent that the meta-analysis was a result-oriented manipulation of data from other studies, the trial court would have been justified in excluding it, even under the Third Circuit's standard. See *infra* notes 357-66 and accompanying text (discussing "data dredging" in connection with meta-analysis and reanalysis of epidemiological studies).

there were logical criteria, systematically applied, in combining and evaluating the data from previous studies.

Paoli also involved questions related to expert testimony on individual causation.²⁷⁹ Several witnesses appear to have asserted that based on test results indicating the presence of PCBs in the railyard and surrounding area, the presence of PCBs in plaintiffs' blood, plaintiffs' medical records, and the literature on effects of PCBs, they could state "to a reasonable medical certainty"²⁸⁰ that plaintiffs' various ailments were caused by exposure to PCBs.²⁸¹

Reference to the evidence on which these conclusions were purportedly based reveals that those conclusions amounted to nothing more than speculation. Animal studies and studies of incidents involving several chemicals provide only uncertain evidence that the substance will cause human disease at all. Animal studies simply do not produce results that permit reliable conclusions about the likelihood of human disease at particular exposure levels. The Yusho and Yu Cheng studies involved ingestion of much larger quantities of PCBs than the *Paoli* plaintiffs were exposed to, and that exposure involved another, probably more toxic chemical. In fact, *Paoli*, like other recent cases,²⁸² involves a scenario in which plaintiffs complaining of a variety of common ailments²⁸³ attempt to attribute those ailments to exposure to a toxic chemical for which human effects have not been demonstrated or are different from those

279. *Paoli*, 916 F.2d at 862. In some cases, the witnesses seemed to combine the question of whether PCBs can cause disease with the question of whether the assumed exposure caused a particular plaintiff's disease. See, e.g., *id.* at 839 (testimony of Dr. Deborah Barsotti); *id.* at 840 (testimony of Dr. Harry Shubin).

280. *Id.* at 851.

281. In some cases, the diagnoses themselves (not only their causes) were at issue. For example, Dr. Arthur Zahalsky opined that the plaintiffs suffered from immune system injuries. *Id.* at 840. At his deposition, however, Zahalsky admitted that testing (of his own design) required to validate his opinion had not been carried out. *Id.* at 843. The speculative nature of Zahalsky's contentions regarding whether PCBs can cause immune system damage is illustrated by his admission that he had not tested any of the plaintiffs. *Id.* at 843. The court also quotes Zahalsky as stating that "if his tests should support such a conclusion, 'then I will have done something with the clinical immunologists that has not yet been done.' " *Id.* (quoting Zahalsky).

Diagnoses relating to increased fear of illness and emotional distress seem particularly lacking in support. Dr. Deborah Barsotti apparently offered her opinion of this issue without having met or examined any of the plaintiffs, having spoken to only one plaintiff on the telephone. *Id.*

282. See, e.g., *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (childhood cancer, including leukemia; kidney cancer; seizure disorders), *aff'd*, 972 F.2d 304 (10th Cir. 1992); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985) (plaintiffs alleged infertility, miscarriage, birth defects, cancer, emotional disturbance, and other commonplace ailments), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

283. *Paoli*, 916 F.2d at 839 (elevated blood pressure, triglycerides, and cholesterol, elevated liver enzymes, and emotional distress).

complained of by the plaintiffs. Even if one assumes that some level of exposure above background has occurred, a proposition that appears doubtful in *Paoli* and other cases,²⁸⁴ the question remains as to whether the individual plaintiffs' conditions were caused by the exposure or by the more commonplace causes of such diseases in the general population, whether known or unknown, a question that cannot be answered without evidence demonstrating that a substance can cause the plaintiff's disease and indicating increased disease incidence at the levels to which plaintiffs were exposed.²⁸⁵ Animal testing and studies involving high level exposure to several toxic substances simply cannot provide that information.

The logical extension of the plaintiffs' position is that virtually anyone with one or more of a whole host of commonplace ailments who may have come into contact with toxic substances should be able to recover from the entity responsible for the toxic substance. Such a proposition clearly goes too far; yet it is difficult to draw any principled distinctions about who should or should not recover if the reasoning of the *Paoli* plaintiffs and their experts is accepted.²⁸⁶

Concerns about the evidentiary basis of causation in toxic torts are not limited to cases involving large numbers of plaintiffs and an array of alleged injuries. Cases involving one or a few plaintiffs may raise similar concerns. Further, a court may tend to view such cases in isolation, even though extension of the case's underlying logic may lead to results similar to those in cases such as *Paoli*, namely, that there is no principled way to

284. See *supra* notes 260-89 and accompanying text.

285. On remand, the district court held a series of evidentiary hearings in accordance with the Third Circuit's directive. The rulings on the summary judgment motion and other evidentiary rulings can be found at *In re Paoli R.R. Yard PCB Litig.*, No. 86-2229, 1992 U.S. Dist. LEXIS 16287 & 18427-37 (E.D. Pa. Oct. 21, 1992) (numerous rulings); *In re Paoli R.R. Yard Litig.*, No. 86-2229, 1992 U.S. Dist. LEXIS 17602 (E.D. Pa. Nov. 13, 1992) (defendant granted summary judgment with respect to plaintiffs' property damage claims). The plaintiffs presented a somewhat different group of witnesses than it had prior to the first summary judgment and appeal. Judge Kelly analyzes the testimony of each of those witnesses in separate opinions. See *Paoli*, 1992 U.S. Dist. LEXIS 18427-37.

286. Similar concerns undoubtedly underlay Judge Weinstein's opinion in the Agent Orange "opt-out" cases, discussed *infra* text accompanying notes 319-21. The plaintiffs sought recovery for many commonplace ailments including infertility, birth defects, miscarriages, liver disorders, skin rashes, and gastrointestinal disorders. It seems obvious that many instances of those conditions would have occurred among Vietnam veterans and their families without any causal relation to their service in Vietnam. As Judge Weinstein observed: "There are roughly 7500 new cases of Hodgkin's Disease per year in the United States. The fact that seventeen of these persons happen to be Agent Orange plaintiffs proves nothing about the origin of their condition." *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. at 1253 (citation omitted). Judge Weinstein goes on to state: "[Plaintiffs' experts Doctors Singer and Epstein] fail to show how the myriad illnesses at issue are more likely to have been caused by Agent Orange than by something else." *Id.*

distinguish persons whose disease was caused by a toxic substance exposure from those whose diseases were not so caused.

The New Jersey Supreme Court's decision in *Rubanick v. Witco Chemical Co.*,²⁸⁷ another case involving injuries claimed to have resulted from PCB exposure, is an example of that scenario. Plaintiffs, the relatives of two Witco employees who had died of colon cancer, claimed that their decedents' cancers were caused by PCB contamination at the site.²⁸⁸ Their expert on causation was Dr. Balis, a Ph.D. biochemist with extensive research experience on colon cancer.²⁸⁹ Dr. Balis cited studies indicating that PCBs cause cancer in animals, reports on the effects of PCBs on animals and humans, a high rate of cancer at Witco "during the relevant period," and the personal history of one of the deceased, indicating the absence of other risk factors.²⁹⁰ The difficulties of drawing inferences from animal studies and studies of the effects of PCBs in humans paralleled the difficulties in *Paoli*.²⁹¹ Further, the absence of other risk factors is of little probative value where the causes of most instances of a disease are unknown.²⁹²

The New Jersey court focused disapprovingly on the trial court's finding that Dr. Balis' theory of causation was not generally accepted in the scientific community.²⁹³ Citing the need for a more liberal standard for determining the reliability of scientific theories of causation in toxic torts cases,²⁹⁴ the court held that "a scientific theory of causation that has not yet reached general acceptance may be found to be sufficiently reliable if it is based on sound, adequately-founded scientific

287. 593 A.2d 733 (N.J. 1991).

288. *Id.* at 735.

289. *Id.*

290. *Id.* at 735-36.

291. See *supra* notes 260-89 and accompanying text.

292. There also seems to have been considerable doubt about the extent of Rubanick's exposure to PCBs. The New Jersey Supreme Court quotes Dr. Balis' summary of the evidence of exposure as follows:

that there was some thirty-five thousand parts per million PCBs in the soil around there, that he would come home covered with this stuff and the material was oozing out of his clothes, according to I guess it was his wife's testimony, it was something, and I think that report that he lifted these heavy drums and slopping around in this muddy PCB mix, and you also showed me some document about the State of New Jersey, some agency complaining about contamination from that stuff.

Rubanick, 593 A.2d at 736 (quoting testimony of Dr. Balis). One of the defendant's witnesses stated, "[T]here is no evidence that I have seen to this date that would definitively suggest that the individual actually did have extensive exposure to PCBs." *Id.* (quoting testimony of Dr. Fahey).

293. The trial court also found that although Dr. Balis was qualified to offer an opinion on human carcinogenesis generally, he was not a physician and thus was not qualified to offer an opinion on the cause of a specific person's cancer. *Id.* at 737.

294. *Id.* at 740.

methodology involving data and information of the type reasonably relied on by experts in the scientific field."²⁹⁵ The court went on to state that the theory must be offered by an expert with "a demonstrated professional capability to assess the scientific significance of the underlying data and methodology, and to explain the bases for the opinion reached."²⁹⁶

Remanding the case for further proceedings on the admissibility of Dr. Balis' testimony, the court admonished the trial court not to scrutinize the expert's methodology itself to determine its soundness, however, but to refer to the opinions of comparable experts in the field.²⁹⁷ The problem with that recommendation is that it was unclear that the expert used any methodology at all. The trial court's opinion makes it clear that Dr. Balis testified in terms of possibilities, not probabilities.²⁹⁸ Moreover, the trial court had the benefit of the testimony of the defendant's witnesses, who opined that the scientific literature did not support the plaintiffs' expert's opinion that PCBs can cause cancer in humans.²⁹⁹ The court itself had read the articles cited by the plaintiff's expert and concluded that they "do not say what plaintiff's expert concludes."³⁰⁰

Rubanick, like *Paoli*, is a case in which an appellate court appears to approve of toxic tort causation testimony based on uncertain exposure levels, animal testing and other indicators of possible carcinogenicity, despite the absence of any evidence suggesting a connection between PCB exposure and the plaintiff's specific disease. The court cites the appropriate criteria, including reliability, but neither conducts nor allows the trial court to conduct a reliability analysis. Rather than accepting the trial court's findings, which were based on testimony of other experts, the appellate court interjects its own assessment. In reality, the New Jersey Supreme Court's rationale is based almost entirely on the qualifications of the expert, since it harks back to the witness's qualifications as a point of

295. *Id.* at 747-48.

296. *Id.* at 748. The New Jersey Supreme Court emphasized the need for the court to scrutinize the expert's "status" and to direct the jury's attention to "factors that bear relevantly on the expert's credibility." *Id.* at 750.

297. *Id.* at 747-48.

298. When plaintiff's counsel questioned Dr. Balis on the issue of whether the scientific community accepts PCBs as human carcinogens, Balis replied that the people he contacted would not say "probable," but rather that it was a "high possibility." *Rubanick v. Witco Chem. Corp.*, 542 A.2d 975, 983 (N.J. Super. Ct. Law Div. 1988), *rev'd*, 576 A.2d 4, 14 (N.J. Super. Ct. App. Div. 1990) (discussing Dr. Balis' comment), *modified*, 593 A.2d 733 (N.J. 1991).

299. *Id.* at 983.

300. *Id.*

reference for determining the reliability of the data on which the expert relies as well as his methodology.³⁰¹

The problems of deferential review are not restricted to environmental exposure cases; they also occur in products liability cases where the costs of erroneous findings of liability in terms of withdrawal of useful products and disincentives to new product development are perhaps more apparent. *Ferebee v. Chevron Chemical Co.*,³⁰² perhaps the paradigm decision involving deferential review, involved injuries allegedly caused by a pesticide.³⁰³ Plaintiff's causation case was essentially based on the testimony of treating or examining physicians who claimed to have seen a few similar cases. It therefore illustrates the pitfalls of medical testimony based on the coexistence of exposure and disease.

The Bendectin cases are based on a more complex assemblage of evidence and testimony, consisting of chemical structure-activity analysis, *in vitro* testing, animal studies and purported reanalyses of existing epidemiologic studies.³⁰⁴ A review of one of the early cases decided in favor of plaintiffs, *Oxendine v. Merrell Dow Pharmaceuticals, Inc.*³⁰⁵ reveals much of the same evidence that plaintiffs have argued in other cases alleging birth defects caused by Bendectin. The chemical structure-activity analysis consisted of the observation that one of Bendectin's ingredients is an antihistamine and that some antihistamines are teratogenic.³⁰⁶ The *in vitro* and *in vivo* animal test results cited by plaintiffs' witness Dr. Done are subject to the same concerns for high rates of false positives that were discussed previously. The remaining evidence

301. *Rubanick*, 576 A.2d at 7-8, 14. The New Jersey Supreme Court's opinion in *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992), appears to have tightened the standard for admission of expert testimony. In regard to the issue of whether epidemiologic studies could provide the basis for an expert's opinion on causation, the court stated that such studies "must have been 'soundly and reliably generated' and be 'of a type reasonably relied on by comparable experts in the particular field.' " *Id.* at 1087 (quoting *Rubanick*, 593 A.2d 733). The court went on to state: The court must also examine the manner in which experts reason from the studies and other information to a conclusion [T]hat conclusion must derive from a sound methodology that is supported by some consensus of experts in the field. *Id.* Nonetheless, the court remanded the case for further proceedings on the issue of whether the plaintiff could satisfy the more-probable-than-not standard of proof despite relative risk data that showed less than a doubling of background risk. *Id.* at 1088.

302. 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984).

303. *Id.* at 1531-32.

304. See generally Sanders, *supra* note 17.

305. 506 A.2d 1100 (D.C. 1986).

306. *Id.* at 1104.

consisted primarily of Dr. Done's unpublished reanalysis of a previous epidemiologic study, which involved selective elimination of data.³⁰⁷

As will be discussed in more detail in Section VI.B.2 of this Article, reanalyses of epidemiologic studies are particularly susceptible to manipulation to achieve a preconceived result. The reanalyses offered by plaintiffs stand in contrast to a large body of epidemiologic evidence that has failed to confirm a statistically significant association between Bendectin and birth defects. Further, the fact that the studies offered by plaintiffs have been unpublished and therefore not subjected to peer review,³⁰⁸ lends further support to other courts' decisions to exclude them.³⁰⁹

B. Active Review Exemplified

In contrast to the deferential, uncritical review accorded expert testimony in *Paoli, Rubanick and Oxendine*, there are a growing number of decisions that utilize active review to make discerning judgments about scientific evidence. Judge Weinstein has been widely criticized for his exclusion of plaintiffs' evidence in the Agent Orange "opt out" cases,³¹⁰ but basically, he got it right. Plaintiffs claimed a wide variety of commonplace ailments, cancers and birth defects as injuries due to Agent Orange or more specifically, the contaminant dioxin.³¹¹ Their evidence consisted of animal studies, and workplace exposure studies that apparently did not indicate an association of dioxin or Agent Orange exposure with the diseases complained of.³¹² There was simply no evidence from which a fact-finder could conclude that any of the plaintiffs suffered from conditions attributable to Agent Orange rather than the causes of such disease in the general population, a fact recognized by the court.³¹³ That conclusion is valid even without taking into account the many studies of

307. See *id.* at 1107-08; see also *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 946-49, 954-57 (3d Cir. 1990) (discussing the failure of Done's reanalysis of other studies to meet traditional statistical significance criteria).

308. See, e.g., *Daubert v. Merrell Dow Pharmaceuticals*, 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992); *Lynch v. Merrell-Nat'l Lab.*, 830 F.2d 1190, 1194-96 (1st Cir. 1987) (discussing reanalyses by Dr. Done and Dr. Shanna Swan).

309. At least one published reanalysis of epidemiologic data has found no association between Bendectin and birth defects. See *infra* notes 373-76 and accompanying text.

310. See, e.g., *Brennan, supra* note 25, at 9 n.40, 53-56; *Green, supra* note 65.

311. See *Green, supra* note 65, at 659.

312. See *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1236 (E.D.N.Y. 1985) (liver disorders in animal and industrial studies differed from those reported by plaintiffs), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

313. *Id.* at 1239.

Vietnam veterans put before the court that failed to show any increased incidence of serious disease.³¹⁴

The Bendectin litigation has also produced opinions that discerningly review scientific evidence; two such cases are *Brock v. Merrell Dow Pharmaceuticals, Inc.*³¹⁵ and *Lynch v. Merrell-National Laboratories*.³¹⁶ Bendectin has been the subject of over 2000 suits for birth defects allegedly caused by in utero exposure to the anti-nausea drug. Plaintiffs have sought recovery for a variety of malformations, but a number of the cases have involved limb reduction defects.³¹⁷ In *Brock*, the court based its reversal of a jury verdict for plaintiffs on the absence of any statistically significant epidemiologic evidence of an association between Bendectin and birth defects.³¹⁸ Both the *Brock* and the *Lynch* courts concluded that the plaintiffs' *in vitro* testing and animal studies evidence was insufficient, the *Lynch* court noting particularly the inability of *in vivo* and *in vitro* animal studies to prove causation in humans "in the absence of confirmatory epidemiologic data," which it contrasted with a number of studies that failed to find an association between Bendectin and birth defects.³¹⁹

In what has become one of the more controversial aspects of the Bendectin litigation, both courts rejected reanalyses of existing epidemiological studies that purported to show an association between Bendectin exposure and birth defects. The *Lynch* court noted the plaintiffs' failure to file any description of the expected testimony of Dr. Shanna Swan, whose reanalysis of epidemiologic data was offered by plaintiffs. The court went on to examine the basis of Swan's opinion from testimony in other litigation, observing that Swan's control group consisted of children with genetic birth defects, a group that had a lower than background risk for certain types of birth defects, raising the

314. More recent information concerning the hazards of dioxin does not significantly change the analysis. New data on chemical workers exposed to dioxin for more than a year, with more than twenty years' latency, has indicated a 46% increase in all cancers (although not any individual cancer). David J. Hanson, *supra* note 158, at 7, 10. According to Marilyn Fingerhut, the author of the study, the serum levels of dioxin correlated well with duration of exposure. *Id.* at 10. The Fingerhut study is consistent, however, with the Ranch Hand study of Vietnam veterans, which has not detected an increase in cancer at the lower exposure levels experienced by veterans, *id.* at 9, although it has recently revealed significant increases in body fat and diabetes that correlated with dioxin concentration, *id.* at 9. This information does not appear to provide a basis for distinguishing background causes from dioxin for most of the ailments claimed to result from dioxin in the Agent Orange litigation.

315. 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990).

316. 830 F.2d 1190 (1st Cir. 1987).

317. See *Sanders, supra* note 17, at 398-99.

318. See *Brock*, 874 F.2d at 314-15.

319. *Lynch*, 830 F.2d at 1194.

question whether genetic defects might make that control group less susceptible to non-genetic defects such as limb reduction. A lower susceptibility in the control group would skew the relative risk observed for the Bendectin-exposed group.³²⁰

In contrast, the *Brock* court's rationale focused on the fact that the elevated risk found in the reanalysis conducted by Dr. Jay Glasser lacked statistical significance.³²¹ As will be discussed in more detail in the following Section, reanalyses of epidemiological data are susceptible to inadvertent and inadvertent introduction of bias. Although the statistical significance point is arguable, the reanalyses were unpublished and therefore lacked the safeguards against biased or result-oriented data selection that peer-reviewed publication would have provided.

VI. ACTIVE REVIEW: THE ANTIDOTE FOR JUNK SCIENCE

As the foregoing Sections have demonstrated, active review of scientific evidence and expert testimony can go far to eliminate the arbitrary and unfair results that can result from the acceptance of junk science in toxic torts cases. Courts nonetheless cite a number of reasons for deferential review of scientific evidence, including the lack of any special expertise and, perhaps more significantly, the belief that traditional tort law, with its typical reliance on established science, is inadequate to redress toxic injuries of the industrial age. Those reasons, however, do not stand up to careful examination.

A. Courts' Ability to Review Scientific Evidence

As noted previously, one of the concerns regarding scrutiny of scientific evidence is the belief that courts lack the ability to understand scientific evidence and therefore should not deprive the jury of the opportunity to consider possibly relevant and probative evidence. It should be evident from the foregoing discussion, however, that courts are quite capable of determining whether there is a reasoned basis, grounded in fact, for expert opinion, as well as a level of reliability consistent with the applicable standard of proof.

The Third Circuit and the New Jersey Supreme Court, who authored the *Paoli* and *Rubanick* decisions respectively, have demonstrated their understanding of complex scientific evidence. In *DeLuca v.*

320. *Id.* The court also addressed a reanalysis of epidemiologic studies by Dr. Alan Done. Neither the Swan nor the Done study had been published. The court also questioned the bases for exclusion of certain data in each. *Id.* at 1194-96.

321. The Glasser study yielded a relative risk of 1.49, with a confidence interval of 0.17 to 3.0. *Brock*, 874 F.2d at 312. Because the confidence interval included the value 1.0, which represents no increased risk, the result did not satisfy statistical significance criteria. See *infra* notes 367-82 and accompanying text (discussing statistical significance).

Merrell Dow Pharmaceuticals, Inc.,³²² the Third Circuit discussed the epidemiologic evidence on Bendectin and birth defects. At issue was the admissibility of a meta-analysis of existing epidemiologic studies. The meta-analysis in question³²³ did not meet the level of statistical significance³²⁴ typically required for epidemiologic studies.³²⁵ The court's view of the meta-analysis was overly generous,³²⁶ but the opinion demonstrates the court's understanding of the concepts of statistical significance and the effects of bias in epidemiologic studies.³²⁷ Similarly, when confronted with a causation issue on which epidemiologic evidence was offered, the New Jersey Supreme Court also evidenced a sophisticated understanding of such evidence. In *Landrigan v. Celotex Corp.*,³²⁸ the court discussed the proffered epidemiologic evidence and the causal inferences to be drawn from it, as well as the concept of attributable risk.³²⁹

322. 911 F.2d 941 (3rd Cir. 1990).

323. One issue concerning admissibility is publication of studies; one meta-analysis of Bendectin epidemiologic studies has been published, unlike the study offered in DeLuca by Dr. Alan Done. See Thomas R. Einarson et al., *A Method for Meta-Analysis of Epidemiological Studies*, 22 DRUG INTELLIGENCE & CLINICAL PHARMACY 813 (1988); Sanders, *supra* note 17, at 341 n.182.

324. See *DeLuca*, 911 F.2d at 946-48.

325. *Id.* at 955-56.

326. On remand, the district court once again dismissed the plaintiffs' case on summary judgment. See *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 791 F. Supp. 1042 (D.N.J. 1992). The court excluded Dr. Alan Done's reanalysis of epidemiologic studies, finding that Done's calculations and presentation of his results contained numerous errors and his methodology could not be discerned or replicated by the other experts of either plaintiffs or defendants. *Id.* at 1047-48. Plaintiffs' other expert was Dr. Shanna Swan, who has also appeared in other Bendectin cases, including *Daubert v. Merrell Dow Pharmaceuticals, Inc.* In *DeLuca* on remand, she apparently commented on the Done reanalysis based on Done's representations of how it was performed. *Id.* at 1047.

327. The Third Circuit reversed the trial court's summary judgment for defendants and directed the trial court to evaluate the reliability of the proffered evidence "with an eye to all the risks of error posed" by it. *DeLuca*, 911 F.2d at 955. The court further stated, "The root issue . . . is what risk of what type of error the judicial system is willing to tolerate." *Id.* The court suggested that additional expert testimony on statistical significance would be helpful, but also stated a preference for admitting evidence with probative value and "dealing with the risk of error through the adversary process." *Id.* at 956. The ultimate issue, however, is whether that evidence would be sufficient to support a jury finding that Bendectin "more likely than not caused the [plaintiff's] birth defects." *Id.* at 958. The court characterized that requirement as requiring a relative risk greater than two in the exposed population. *Id.*

328. 605 A.2d 1079 (N.J. 1992).

329. *Id.* at 1085-88. The *Landrigan* court, however, refused to hold that a relative risk of 2.0 or greater is required to prove that individual causation was more probable than not. *Id.* at 1087. The court's assumption appears to have been that by ruling out other causes, plaintiff's expert could opine that causation of plaintiff's disease was more probable than not, a proposition that depends on how well developed the evidence is on other risk factors and the relationships among them. *Id.*

No doubt there are times when expert witness testimony and scientific evidence are obscure. The details of statistical significance calculations could undoubtedly lose all but the most mathematically inclined and dedicated lay observer. But judges need not examine expert testimony and scientific evidence at that level of detail. Courts can and should, however, require the proponent of such evidence to demonstrate to the court that the evidence is valid and reliable, that is, that it makes sense and is sufficiently likely to produce an accurate result.

B. Overcompensating for the Deficiencies and Inequities of the Tort System

Another reason courts cite for lenient review of expert testimony is perceived inequities and deficiencies of the tort system. The Sections below examine several aspects of the perception and show that there, too, the cited reasons do not justify the remedy.

1. THE BELIEF THAT MOST CANCERS AND BIRTH DEFECTS ARE CAUSED BY TOXIC PRODUCTS AND ENVIRONMENTAL POLLUTANTS.

Rubanick and *Paoli* illustrate the perception that many, if not most, cancers and birth defects are caused by toxic substances introduced into the environment in products or as waste injuries that they believe will go uncompensated if traditional evidentiary standards are applied. The New Jersey Supreme Court's opinion in *Rubanick* states that concern explicitly:

There are undeniable indications that persons do in fact suffer grave and lethal injury as a result of the wrongful or tortious exposure to toxic substances. Those indications do not spring simply from conjecture; they conform to our common experience and informed intuition. Judge Petrella noted in his opinion below that "[i]t has been widely considered that PCBs are a carcinogenic substance." Our common sense, with some empirical support, tells us of the deleterious effects of PCBs.³³⁰

The Third Circuit expressed similar beliefs in *Paoli*.³³¹ Those perceptions appear to be based on widely quoted statements that most

330. *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 747 (N.J. 1991) (citations omitted).

331. In addressing whether Pennsylvania would recognize a medical monitoring claim, the court noted the need to "accommodate a society with an increasing awareness of the danger and potential injury caused by the widespread use of toxic substances." *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 850 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). The court went on to note:

The necessity of addressing problems of toxic exposure becomes particularly important with the continued widespread use of chemicals in American industrial and agricultural development. One commentator has pointed out that there are approximately 50,000 hazardous waste sites nationwide. In

cancers are caused by environmental factors.³³² A more careful reading of the sources of such sweeping statements, however, reveals that the

all, over 65,000 chemicals are in commercial use today which have not been tested for their effects on human health or the environment. According to varying estimates, workplace exposure to hazardous substances alone accounts for from five percent to as much as thirty-eight percent of all cancers.

Id. at 850 n.22 (quoting Leslie Gara, Note, *Medical Surveillance Damages: Using Common Sense and the Common Law to Mitigate the Dangers Posed by Environmental Hazards*, 12 HARV. ENVTL. L. REV. 265 (1988)).

332. See *id.* In 1978, David Doniger wrote the following:

From comparisons of different rates of different cancers throughout the world, the World Health Organization and other prominent institutions and individual experts have concluded that 60 to 90 percent of all human cancers are caused by exposure to chemical substances (and, to a lesser extent, radiation) present in our air, workplaces, food, water, and the rest of our environment.

Doniger, *supra* note 131, at 509.

Doniger was not alone in his concern about chemicals and carcinogenesis. *See id.* and references cited therein. Professor Bruce Ames, the developer of the "Ames" mutagenicity test, expressed similar concerns in a 1979 publication recommending mutagenicity assays as methods for identifying mutagens and carcinogens:

A variety of data supports the hypothesis that environmental factors are a major cause of cancer. Epidemiologic studies show different rates of incidence for certain types of cancer in different parts of the world. For example, in Japan there is an extremely low rate of breast and colon cancer and a high rate of stomach cancer, whereas in the United States the reverse is true. When Japanese immigrate to the United States, within a generation or two they show the high colon and breast cancer rates and low stomach cancer rates characteristic of other Americans. Known environmental mutagens that can cause human cancer include cigarette smoke tar, ultraviolet light, x-rays, and asbestos, and the list of human chemical carcinogens is steadily lengthening.

Ames, *supra* note 29, at 587 (citations omitted). Those concerns were prompted in part by the rapid increase in production and exposure of the workforce and the general public to synthetic chemicals. These concerns were summarized by Ames:

Clearly, many more chemicals will be identified as human mutagens and carcinogens. Currently over 50,000 synthetic chemicals are produced and used in significant quantities and close to 1000 new chemicals are introduced each year. Only a small fraction of these were tested for carcinogenicity or mutagenicity before their use. In the past this problem was largely ignored, and even very high-production chemicals with extensive human exposure were produced for decades before adequate carcinogenicity or mutagenicity tests were performed. Such chemicals now known to be both carcinogenic and mutagenic include vinyl chloride (produced at a rate of about 6 billion pounds per year in the United States in 1977) and 1,2-dichloroethane (ethylene dichloride, about 10 billion pounds per year) and a host of high-production pesticides.

The increase in production and use of chemicals has been particularly great since the mid-1950's This flowering of the chemical age may be followed by genetic birth defects and a significant increase in human cancer during the 1980 decade (because of the 20- to 30-year lag) if many of these

environmental factors encompassed by such statements include commonplace causative factors such as background radiation and probably biological processes such as aging, that are largely beyond human control, as well as cigarette smoking, alcohol consumption and dietary factors such as a high fat diet that are the result of lifestyle, not industrial pollutants.³³³

The fraction of cancers and other diseases that could be prevented by reducing or eliminating exposure to man-made toxic chemicals is still in dispute. Studies that have attempted to estimate the fraction of cancers caused by environmental pollution have placed the figure at about six percent, up to as much as fifteen percent.³³⁴ Other exposures and industrial products are thought to add an additional four to five percent, perhaps as much as ten percent.³³⁵

The debate about the role of synthetic chemicals in cancer causation has occurred against a backdrop of increasing cancer rates.³³⁶ The meaning of the data is unclear, however, because most if not all of the increase can be attributed to increases in smoking-related cancers and

chemicals with wide-spread human exposure are indeed powerful mutagens and carcinogens.

Ames, *supra* note 29, at 587-88.

333. Richard Doll and Richard Peto have noted that the phrase "environmental factors" has been "misinterpreted by many people to mean only 'man-made chemicals,' which was certainly not the intent of the WHO committee." Richard Doll & Richard Peto, *The Causes of Cancer: Quantitative Estimates of Avoidable Risk of Cancer in the United States Today*, 66 J. NAT'L CANCER INST. 1192, 1197 (1981). The Doll and Peto article was commissioned as a report to the Office of Technology Assessment of the U.S. Congress. *Id.* at 1193. For a discussion of various avoidable risks, including those of smoking, alcohol use, diet, and other causes, see *id.* at 1220-56.

334. *Id.* at 1256. The six percent figure is the sum of the percentages attributed to occupation, pollution and industrial products. The 15% figure is the sum of the high end of the ranges estimated for each of those sources, which is likely an overestimate because it is the sum of worst case estimates and because the contribution of risk factors is not necessarily additive. See *supra* notes 209-22. This analysis omits medical sources, which include diagnostic X-rays.

335. Doll & Peto, *supra* note 333, at 1256. These figures contrast markedly with a much-cited report filed with the Occupational Safety and Health Administration, which asserted that up to about 40% of all cancers in the U.S. might be occupationally related. NATIONAL CANCER INSTITUTE ET AL., ESTIMATES OF THE FRACTION OF CANCER IN THE UNITED STATES RELATED TO OCCUPATIONAL FACTORS 1 (1978). Doll and Peto point out errors in the methodology of the OSHA report, however, which they believe resulted in overestimation of the proportion of cancers attributable to occupational exposures. Doll & Peto, *supra* note 333, at 1240-41. Causes are not mutually exclusive, however, as the example of asbestos and smoking indicates. See *supra* note 222 and accompanying text. Thus, fractions of total cancer death attributable to various causes could exceed 100%. Doll & Peto, *supra* note 333, at 1219-20. Consequently, attribution of a large fraction of cancers to occupational factors would not necessarily be inconsistent with attributing a similarly large proportion to other factors, such as smoking and alcohol consumption. See ROTHMAN, *supra* note 44, at 14.

336. See, e.g., Earl S. Pollack & John W. Horm, *Trends in Cancer Incidence and Mortality in the United States, 1969-76*, 64 J. NAT'L CANCER INST. 1091 (1980); Smith, *supra* note 33, at 998.

aging of the population.³³⁷ Even if the more pessimistic experts are correct in their conclusion that age adjusted rates are increasing for some cancers,³³⁸ the data do not support the proposition that most cancers are caused by toxic pollution or toxic products (other than cigarettes). Thus, there is no factual basis for a presumption that environmental pollutants cause most cancers.

2. THE BELIEF THAT SCIENCE'S ABILITY TO IDENTIFY CAUSES IS TOO LIMITED.

A second factor, which is related to the belief that toxic substance exposures are causing large amounts of disease, is courts' frustration over the limitations inherent in science's ability to identify causes. Unwilling to accept those limitations, the *Ferebee* court stated:

A cause-effect relationship need not be clearly established by animal or epidemiologic studies before a doctor can testify that, in his opinion, such a relationship exists. As long as the basic methodology employed to reach such a conclusion is sound, such as use of tissue samples, standard tests, and patient examination, products liability law does not preclude recovery until a "statistically significant" number of people have been injured or until science has had the time and resources to complete sophisticated laboratory studies of the chemical. In a courtroom, the test for allowing a plaintiff to recover in a tort suit of this type is not scientific certainty but legal sufficiency; if reasonable jurors could conclude from the expert testimony that paraquat more likely than not caused Ferebee's injury, the fact that another jury might reach the opposite conclusion or that science would require more evidence before conclusively considering the causation question resolved is irrelevant.³³⁹

Not surprisingly, *Ferebee* is widely quoted, particularly by courts that are disposed to admit purported scientific evidence without scrutiny of the underlying reasoning.³⁴⁰ Indeed, the premise of *Ferebee*, namely that the law does not in general require statistical evidence of causation, is hardly subject to dispute. *Ferebee* also appeals to fairness by appearing to

337. Cancer death rates for males from lung cancer have increased dramatically since 1930, while death rates from stomach cancer have steadily declined. See Eliot Marshall, *supra* note 33, at 901. Trends for other common cancers in males are less pronounced. There has been considerable debate about the inferences drawn from the data. A number of statisticians and epidemiologists argue that once the data are adjusted for age and the effects of smoking, the overall incidence of cancer is decreasing. See Smith, *supra* note 33, at 998. Other factors that make interpretation difficult are the effects of increased accuracy of diagnosis and disagreement over the significance of increasing cancer rates among the aged. See Marshall, *supra* note 33, at 901-02.

338. See Marshall, *supra* note 33, at 901.

339. *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529, 1535-36 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984).

340. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733 (N.J. 1991).

correct the imbalance that disfavors toxic torts plaintiffs, created by the unavailability, high costs, and insensitivity of epidemiologic studies required to link toxic substance exposures to latent injuries.³⁴¹

The inability of epidemiologic studies to detect small increases in risks has been a major concern in the debate over toxic tort causation evidence.³⁴² The power of an epidemiologic study to identify a small increase in risk is a function of the size of the study groups and the background rate of disease, with larger study groups corresponding to greater statistical power.³⁴³ Meta-analysis, in which the data from a number of smaller studies are combined and reanalyzed, can enhance the likelihood of detecting an effect, if one exists.³⁴⁴ Meta-analysis can also provide the opportunity to refine the selection of data included in the analysis to address potential bias in sample selection, as can reanalysis of a single study.³⁴⁵

Systematic error, of which bias is one form, can be introduced into epidemiologic studies in a number of ways, including the failure to control for causal factors other than the factor under study and the failure to accurately delineate exposed and unexposed populations.³⁴⁶ One of the potential sources of bias in the Bendectin studies is recall bias, the possibility that mothers of children born with defects will be more likely to recall drug use during pregnancy than mothers of normal infants. Such recall bias will tend to result, in some kinds of studies, in an overestimation of the effect of the drug.³⁴⁷ Another concern with inaccurate recall is that the "unexposed" group will, in fact, have some individuals who were exposed and who exhibit effects caused by the exposure.³⁴⁸ If there is an effect, part of that effect will be attributed to the

341. Epidemiologic studies require considerable time and money to conduct, often beyond the means of the toxic tort plaintiff. Further, epidemiologic studies are a crude method for detection of small increases in disease that have long latency periods and significant background risks. See Brennan, *supra* note 25, at 54; Doll & Peto, *supra* note 333, at 1219; see also Ames, *supra* note 29, at 587 (advocating the use of short-term assays to identify mutagens and carcinogens). Further, epidemiologic studies are often designed to detect a relative risk of two or more. See M.J. Adams Jr. et al., *The Use of Attributable Fraction in the Design and Interpretation of Epidemiologic Studies*, 42 J. CLINICAL EPIDEMIOLOGY 659, 659 (1989).

342. Green and Brennan have argued that such insensitivity requires that plaintiffs be allowed to resort to other kinds of toxicological evidence to prove their cases. See Brennan, *supra* note 25, at 56; Green, *supra* note 65, at 680-81.

343. See ROTHMAN, *supra* note 44, at 79-80. Power is "the probability of detecting (as 'statistically significant') a postulated level of effect." *Id.* at 79.

344. See Naylor, *supra* note 277, at 892.

345. Both meta-analyses and reanalyses of existing studies have been at issue in the Bendectin litigation. See *supra* notes 65-70 and accompanying text.

346. See ROTHMAN, *supra* note 44, at 82-94.

347. *Id.* at 85. This possibility applies to case control studies. *Id.*

348. See *id.* at 85-87.

unexposed group, tending to diminish the magnitude of the observed effect.³⁴⁹

To counter the negative epidemiologic evidence that predominates in the published literature concerning Bendectin, several plaintiffs have variously offered meta-analyses or reanalyses by one or both of two expert witnesses, Dr. Alan Done, a professor of pediatrics and pharmacology, and Dr. Shanna Helen Swan, an epidemiologist and chief of the a unit of the California Department of Health Services.³⁵⁰ Meta-analyses are subject to questions about the propriety of combining data from studies in which the original criteria for selection of subjects and controls differed.³⁵¹ Both meta-analyses and reanalyses involve selection of data for inclusion and exclusion, which create the opportunity for "data dredging" that may turn up statistically significant correlations that are actually due to chance.³⁵² The methodology by which data were selected for inclusion and exclusion in meta-analyses and reanalyses should therefore be carefully scrutinized.

A number of objections can be made to the reanalyses and meta-analyses offered by various Bendectin plaintiffs. In the case of Dr. Done's reanalysis at issue in *Lynch*, the basis of the data selection seems less than clear,³⁵³ although the *Oxendine* opinion indicates that in the reanalysis offered by Done in that case, some pairs of exposed and unexposed children were eliminated because Done considered the risk of recall bias to be especially high among Canadian subjects who could have purchased Bendectin without a prescription.³⁵⁴ Dr. Shanna Swan's methodology is explained more completely in *Lynch*; it involved the reanalysis of data previously analyzed by four members of the Center for Disease Control.³⁵⁵ All of the subjects in the original group had involved abnormal children. Swan reanalyzed the data, using only children with genetic abnormalities as the control group so that the control group's abnormalities could not have resulted from Bendectin.³⁵⁶ Her reanalysis

349. See *id.*

350. See *Lynch v. Merrell-Nat'l Lab.*, 830 F.2d 1190, 1194-95 (1st Cir. 1987). In *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990), plaintiffs offered a reanalysis by Dr. Jay Glasser. *Id.* at 312.; see *supra* note 321 and accompanying text.

351. See Naylor, *supra* note 277, at 893. Although Naylor is discussing the aggregation of data from clinical trials rather than retrospective exposure cases, the argument still applies.

352. See *id.*, at 894.

353. *Lynch*, 830 F.2d at 1196.

354. *Oxendine v. Merrell Dow Pharmaceuticals, Inc.*, 506 A.2d 1100, 1107-08 (D.C. 1986).

355. See *Lynch*, 830 F.2d at 1195; see also Eliot Marshall, *Supreme Court to Weigh Science*, 259 SCIENCE 588, 590 (1993).

356. *Lynch*, 830 F.2d at 1195.

concluded that Bendectin is associated with an increased risk of birth defects. Swan's reanalysis raises questions because the control group for her reanalysis was acknowledged to have only a 0.57 relative rate (i.e., a 40% lower rate) for certain categories of birth defects. As the First Circuit noted, "Swan made no allowance for the possibility that the very fact of having such a severe genetic deficiency as Down's Syndrome might operate to make other rare deficiencies such as limb reduction less likely," thus skewing the apparent differences between the exposed and control groups.³⁵⁷ The possibility that both Done's and Swan's reanalyses were based on result oriented "data dredging" or other inadvertent introduction of bias cannot be ignored; further, none of the reanalyses or meta-analyses has been published in peer-reviewed scientific journals, although ample time has elapsed for review and publication.³⁵⁸ The failure of either to publish their results leaves courts without any reassurance that concerns about bias are unwarranted.

In any event, the insensitivity of epidemiologic studies in the case of Bendectin is probably an overrated concern. Although the studies cannot be said to eliminate all possibility that Bendectin is teratogenic, they at least indicate that if Bendectin is a teratogen, it is a weak one.³⁵⁹ Moreover, the insensitivity of epidemiologic studies does not improve the probative value of other evidence. Animal studies, mutagenicity testing and structure-activity relationships do not become more persuasive because of the absence of other kinds of proof.

3. THE BELIEF THAT SCIENTISTS REQUIRE TOO MUCH CERTAINTY.

A third argument courts cite for abandoning scientific criteria for proof of causation is the perception that scientists require too great a degree of certainty before they will accept a factual proposition as established. *Rubanick v. Witco Chemical Corp.*³⁶⁰ makes numerous references to the high level of proof required by scientists³⁶¹ and concludes that "the scientific method . . . fails to address or accommodate the needs and goals of the tort system."³⁶² That scientists may require a higher level of certainty than the legal system may in some instances be true. In part, the mismatch between expert testimony and legal requirements is the result of the failure of lawyers and courts to articulate legal requirements

357. *Id.*

358. A meta-analysis offered by Dr. Done in another case also failed to satisfy statistical significance criteria. See *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 954-57 (3d Cir. 1990); *Lynch*, 830 F.2d at 1195-96.

359. See *Sanders*, *supra* note 17, at 348.

360. 593 A.2d 733 (N.J. 1991).

361. See *id.* at 737, 739-41.

362. *Id.* at 741.

of proof to scientists. Unless the examining attorneys explore what a scientist-expert witness means by "proof," the risk that the expert understands such terms differently from their legal meaning will always exist.³⁶³

The appropriate level of certainty is particularly an issue when epidemiologic evidence that does not meet epidemiologists' criteria for certainty is offered. Epidemiologists typically are unwilling to conclude that increased disease incidence in an exposed population is associated with a toxic substance exposure unless a statistical analysis of the data shows that the probability of a false positive is 5% or less.³⁶⁴ That requirement represents a 95% confidence level,³⁶⁵ a level that is considerably higher than the more probable than not standard would seem to suggest.³⁶⁶ Moreover, the 5% cutoff for statistical significance is

363. In *Rubanick*, it was clear that the plaintiffs' expert was discussing possibilities and was unable to state that it was more probable than not that PCBs caused the decedents' colon cancers. *Rubanick v. Witco Chem. Corp.*, 576 A.2d 4 , 14-15 (N.J. Super. Ct. App. Div. 1990), modified, 593 A.2d 733 (N.J. 1991); see *supra* note 305 and accompanying text. The mismatch is also due to mechanical application of the *Frye* rule. When the *Frye* general acceptance test is applied to an expert's opinion on whether a toxic substance can cause a particular disease, the test incorporates scientists', rather than the legal system's, standards of proof.

364. See *supra* note 103 and accompanying text.

365. The statistical analysis sometimes focuses on the calculation of a "p-value," which represents the probability that the relative risk produced by the study is due to random variability or chance. See *ROTHMAN, supra* note 44, at 115-19. Often, an upper limit for the p-value is selected as $p = 0.05$; if the p-value of the study falls at or below the cutoff, the results of the study are said to be "statistically significant." A p-value of 0.05 corresponds to a five percent chance that an increase in relative risk is actually a false positive, described as a type I error or alpha-error. *Id.* Alternatively, the statistical analysis may be used to generate confidence intervals, that is, ranges of relative risk that are associated with a specified level of confidence. A 95% confidence interval is the range in which the relative risk would be expected to fall 95% of the time if the study were repeated (hence, a 95% confidence level). *Id.* at 119-20. The confidence level is equal to one minus the probability of type I error; thus, a 95% confidence level corresponds to a statistical significance cutoff value of p equal to 0.05. See *id.* at 119. In *Brock v. Merrell Dow Pharmaceuticals Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990), the court discussed epidemiologic data for which the statistical analysis was expressed in terms of confidence intervals. See *id.* at 312. The court recognized that where a confidence interval includes a relative risk of 1.0 (which represents no effect from the exposure), the study could not be said to demonstrate a statistically significant increased risk of limb defects associated with exposure. *Id.* at 312-13. Confidence intervals are usually calculated for a predetermined confidence level, usually 90% to 95% but occasionally lower. *ROTHMAN, supra* note 44, at 119.

366. This issue was explicitly raised in *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941 (3d Cir. 1990), in which the plaintiffs sought to present Dr. Done's reanalysis of epidemiologic data as evidence of an increased relative risk associated with *in vitro* Bendectin exposure. Dr. Done's reanalysis did not satisfy statistical significance criteria. *Id.* at 955.

arbitrary, and has been used to meet epidemiologists' perceived needs for certainty.³⁶⁷

Some commentators have questioned whether statistical significance is relevant to the more probable than not standard of proof.³⁶⁸ Green suggests that focus on the relative risk found in a study is more appropriate.³⁶⁹ That approach seems untenable, however, because it fails to distinguish the issue of whether exposure to the toxic substance causes any effect at all, which is the function of statistical significance testing, from the issue of the likelihood that a particular plaintiff's case resulted from the exposure rather than background or other causes, a conclusion that is inferred from the magnitude of the relative risk.

Relative risk greater than 1.0 in an exposed population is sufficient evidence of an association of disease with exposure only if we can be reasonably certain that the unequal distribution of disease in exposed and unexposed populations is not due to chance. Ignoring the possibility that an increased incidence of disease is due to chance leads to the obviously absurd result that a disease cluster, no matter how small, could be argued as sufficient evidence of an association between an exposure and disease, a result that is indefensible. The evaluation of the role of chance in an epidemiologic study is thus an essential part of determining the probative value of the evidence.

The appropriate confidence level is a more difficult question, however. At a minimum, the more probable than not standard of proof would seem to tolerate epidemiologic data on the issue of general causation if there is less than a 50% probability that the result is due to chance, a confidence level far lower than the 95% level typically employed by epidemiologists. Additionally, Green and others have noted that typical statistical significance testing is concerned only with the risk of false positives, that is, the risk that an effect will be inferred when there is actually no effect.³⁷⁰ The legal system is also concerned, however, with the risk of false negatives, namely, in toxic torts the risk

367. ROTHMAN, *supra* note 44, at 118-19. The use of $p = 0.05$ lessens the possibility that an effect will be assumed when, in fact, there is no association between exposure and disease incidence. The use of low p -values, however, increases the probability that no association will be assumed when, in fact, there is an association. *Id.* The considerations that have led epidemiologists to require 95% confidence interval as a cut-off for statistical significance are not necessarily appropriate for tort law. Commentators have been unable to agree on appropriate alternatives, however.

368. See Green, *supra* note 65, at 682, 687; David H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1334 (1986) (decrying the mechanical application of statistical significance criteria without explanation and suggesting confidence interval testing as more useful).

369. See Green, *supra* note 65, at 647.

370. See *id.* at 683.

that no effect will be detected when there actually is an effect.³⁷¹ Decreasing the risk of false positives tends to increase the risk of false negatives, though not in a straightforward way.³⁷² Thus, there is an argument that in some instances, epidemiological studies should be admitted with less stringent significance criteria than are typically applied. Before such a rule, which significantly lowers the standard of acceptability of epidemiologic evidence of increased risk, is adopted, however, it would be well to consider other sources of error. Epidemiologic studies are plagued to a greater or lesser degree with other, nonrandom sources of error. Exposure data can be highly uncertain. There is always the possibility that there are unknown confounding causes that are not randomly distributed between the exposed and control populations. Although statistical testing usually does not address nonrandom error, the possibility of other confounding factors may have a great deal to do with the high confidence levels that epidemiology has typically required to minimize the risk of error due to chance.

It may be instructive to consider Bendectin because it has been the subject of over thirty epidemiologic studies and at least one published meta-analysis of those studies.³⁷³ If statistical significance criteria are indeed too stringent, causing scientists to miss a real effect, one would expect to see relative risks from the various studies falling above 1.0 more often than below that number. In other words, the results should vary around the "true" relative risk even if no single study qualifies as statistically significant.³⁷⁴ In his comprehensive study of the Bendectin litigation, Sanders notes that of twenty-six studies from which he was able to extract a value indicative of relative risk, thirteen reported a value greater than one, twelve reported values less than one, and one study reported a value of exactly one.³⁷⁵ That result is roughly consistent with a published meta-analysis of seventeen prior studies that concluded that

371. David Kaye has analyzed the preponderance of the evidence rule as having the effect of minimizing erroneous verdicts. See David H. Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation*, 1982 AM. B. FOUND. RES. J. 487, 496-503. In the statistical analysis of epidemiologic studies, the assumption that there is no effect where there is, in fact, an effect (i.e., a false negative) is referred to as type II or beta-error. ROTHMAN, *supra* note 44, at 117-18.

372. See ROTHMAN, *supra* note 44, at 117-18.

373. See Sanders, *supra* note 17, at 341 n.182.

374. This is a commonsense application of the rationale behind the meta-analysis of existing studies, in which smaller studies are combined to obtain larger sample and control populations. Meta-analysis runs the risk of comparing populations that differ in nonrandom ways, however, and thus some caution is warranted in drawing conclusions in the casual manner suggested in the text. See also ROTHMAN, *supra* note 44, at 334-36 (discussing trend estimation based on differing exposure levels even where individual studies do not satisfy statistical significance criteria).

375. Sanders, *supra* note 17, at 340-41.

Bendectin is not associated with human birth defects.³⁷⁶ If statistical significance criteria were lowered to a 50% confidence level, one is left to wonder whether both plaintiffs and defendants would be offering "statistically significant" evidence, respectively, that Bendectin causes and prevents birth defects. Thus, it is not clear without further evaluation that scientific confidence level criteria are too stringent where epidemiologic evidence is concerned.

A more basic concern with courts' perceptions that scientists require too much certainty is that such views seem to form the basis for rejection of scientific reasoning altogether. The problem with *Ferebee* and its progeny is that they fail to recognize that in most cases,³⁷⁷ there are no alternative proofs available that amount to anything more than speculation or estimation with a great deal of uncertainty.³⁷⁸ Courts' unwillingness to scrutinize testimony on disease causation leaves the door open to the self-validating experts who can be found to testify to virtually any proposition.³⁷⁹ Even the courts that have deemed such evidence admissible have recognized the hazards of their approach.³⁸⁰ Nonetheless, they are willing to risk that kind of error because scientific evidence is unavailable to satisfy traditional standards of proof.³⁸¹ The irony of that rationale is that it rests on courts' and commentators' acceptance and even distortion of scientific speculation that widespread dissemination of new chemicals might result in increases in cancer, birth defects and other disease. Having accepted scientific speculation, they then reject the cautionary statements of scientists who want greater certainty before they reach conclusions.

C. The Costs of Overcompensation

The position taken herein runs counter to the views of several recent commentators. Troyan Brennan has urged courts to admit and consider all the kinds of evidence that toxicologists bring to bear on the question of whether a substance causes disease, including animal studies, short term

376. See *id.* at 341 & n.182.

377. See *supra* note 57 and accompanying text.

378. See *supra* notes 64-67 and accompanying text (discussing structure-activity relationships, short-term testing, and animal studies).

379. See *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1242 (E.D.N.Y 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. Ct. 1234 (1988).

380. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 744 (N.J. 1991) ("There are, assuredly, genuine concerns engendered by a test of reliability of complex scientific theories of causation that does not fully embrace the views of a dominant or of a significant segment of the scientific community.").

381. See, e.g., *id.* at 745 (other courts' demands for "near-scientific certainty are unrealistic" because the level of scientific proof is unavailable).

assays and structure-activity relationships.³⁸² Michael Green goes even further, urging courts to approve of all of the foregoing and even individual case reports as a sufficient evidentiary basis for plaintiffs' verdicts.³⁸³ Moreover, those commentators do not significantly disagree with this author about the uncertainty inherent in those kinds of evidence.³⁸⁴ They do, however, differ on the conclusions reached in the face of those uncertainties.

Brennan's primary suggestion is to propose that questions involving significant scientific uncertainty be resolved by referring those questions to court-appointed experts or science panels.³⁸⁵ There are obviously cases, however, that are not significant enough to warrant science panels, or perhaps even court-appointed experts. Moreover, Brennan does not really come to grips with how evidence with such uncertain probative value can satisfy the more probable than not standard of proof, whether reviewed by a science panel or a lay jury. He recognizes that the acceptance of evidence associated with a high degree of uncertainty is a policy question, but does not provide a rationale for such a radical change in policy.³⁸⁶

Green, on the other hand, recognizes that difficulty. His solution is equally troubling: He states that "plaintiffs should be required to prove causation by a preponderance of the *available evidence*."³⁸⁷ This proposal is at least directly addresses the problem with animal studies and other, even more uncertain kinds of proof. The problem that Green's and Brennan's proposals present, however, is that they create potentially unlimited and ultimately arbitrary liability for cancer, birth defects, and other diseases that lack definitive causal explanations. Rare will be the cancer victim who cannot find some arguably toxic exposure, whether it be the pesticide application on the neighbor's lawn, pumping her own gas at the gas station or other such cause. Rarer still will be the plaintiff who cannot find a treating physician or other expert who is willing to state that based on past experience and review of the literature, that a particular toxic substance exposure is consistent with the plaintiff's disease and that the plaintiff lacked other predisposing factors. Reliance on the available evidence when such evidence suggests only the

382. See Brennan, *supra* note 25, at 21-26.

383. See Green, *supra* note 65, at 646, 674-75.

384. See Brennan, *supra* note 25, at 21-26 (discussing the kinds of uncertainty associated with animal tests, short term assays, and epidemiologic evidence); Green, *supra* note 65, at 680-81 (discussing animal testing, *in vitro* testing, short-term assays, structure-activity analysis, and case studies).

385. See Brennan, *supra* note 25, at 62-71. He suggests that science panels and lists of potential experts be coordinated under a federal science board.

386. See Brennan, *supra* note 15, at 523-32.

387. Green, *supra* note 65, at 680 (emphasis added).

possibility, not the probability, of causation suggests that plaintiffs would do well to proceed to court when the evidence on whether a substance can cause disease is in an unformed stage. Such plaintiffs apparently will not have to contend with the messy questions of distinguishing background risk or other known risks that become issues when epidemiologic evidence is available. Indeed, they would have no basis for making such distinctions.

If there were a way to ease plaintiffs' evidentiary burdens without opening the door to arbitrary and potentially devastating liability for defendants, it would undoubtedly garner considerable support. The zone of uncertainty about the role of toxic chemicals in the causation of many diseases is simply too wide however, to suggest a reasonable way to split the difference.

It must be noted that courts' concerns are not all scientific. Other policy concerns, sometimes unspoken but often implied, seem to underlie courts' willingness to entertain unfounded and poorly reasoned evidence. Those concerns are the indignation and outrage felt by the public in general and plaintiffs in particular over exposure to contaminants or products involving substances suspected of causing harm or whose properties are simply unknown.³⁸⁸ In many of the environmental exposure cases, the exposures or the contamination that could have led to exposure occurred without the plaintiffs' knowledge or consent.³⁸⁹ In the case of potentially toxic products such as breast implants, the exposures have occurred with implicit or explicit assurances that the products were safe.

Traditional tort doctrines, however, do not provide for compensation for egregious conduct without causally related physical injury unless it rises to the level of intentional infliction of emotional distress.³⁹⁰ Commentators have suggested creation of causes of action based on creation of risk,³⁹¹ and a limited number of courts have adopted

388. Studies of risk perception have documented the phenomenon that public acceptance of risk is adversely influenced by the involuntariness of the risk. See Paul Slovic, *Perception of Risk*, 236 SCIENCE 280, 283 (1987).

389. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 835 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991); *Sterling v. Velsicol Chem. Corp.*, 855 F.2d 1188 (6th Cir. 1988); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990), aff'd, 972 F.2d 304 (10th Cir. 1992).

390. See generally 1 DORE, *supra* note 5, §§ 4.01-05, 7.01-08. Recovery based on negligent infliction of emotional distress has been traditionally limited to cases involving physical impact or injury. 1 *id.* § 7.02[2], at 7-3. The limitations of this doctrine have been mitigated somewhat by courts' relaxation and broadening the notion of physical impact to include exposure or subclinical changes. See 1 *id.* at 7-4 to 7-5.

391. See, e.g., *Robinson, Probabilistic Causation*, *supra* note 50, at 783.

such theories.³⁹² Those theories are implicitly and sometimes explicitly premised on assumptions that some significant level of risk can be proved,³⁹³ assumptions that in many cases would be erroneous.³⁹⁴

In any event, the tort system is probably not the best forum for addressing public concerns over uncertain risk. The inability of toxic tort claimants to prove causation has been one of the more important rationales for environmental regulation.³⁹⁵ Indeed regulation is an area in which risk is explicitly recognized as a basis for restricting the dissemination of a substance in products or in the environment. Regulation does not compensate those who are injured despite regulation or by unregulated risks, but it has an important role to play in minimizing risks.

However desirable it might be to have the tort system fill all the gaps where toxic injury occurs, the current state of knowledge simply does not permit the necessary causal connections to be made. Given that state of affairs, what is at stake is whether the "more probable than not" standard of proof will continue to apply to toxic torts. Whether that burden should be lessened or even shifted to defendants are policy issues of the greatest importance. They should be addressed directly and changes, if any, should be based on their fullest consideration of the implications. To effect a reallocation of burdens of proof under the pretext of admitting reliable evidence which is in fact not probative, is not the appropriate way to bring about a change in such a fundamental principle of tort law.

392. The claims have been variously cast as claims for emotional distress, increased risk of future injury, and medical monitoring costs. *See generally* 1 DORE, *supra* note 5, §§ 7.01-08.

393. *See 1 id.* § 7.07, at 7-16.5, 7-27.6 (citing cases refusing to recognize claims based on unquantified risk of injury).

394. *See supra* notes 140-79 and accompanying text for a discussion of uncertainty in risk estimation from nonepidemiologic evidence.

395. The breast implant controversy, however it is ultimately resolved, represents a holdover from a period in which medical devices did not require approval by the Food and Drug Administration, a situation that does not apply to new devices.

ARTICLE

ANTITRUST AND INTERNATIONAL COMPETITIVENESS: IS ENCOURAGING PRODUCTION JOINT VENTURES WORTH THE COST?

DONALD K. STOCKDALE, JR.[†]

Table of Contents

I.	INTRODUCTION	270
II.	CURRENT PROPOSALS TO CHANGE THE ANTITRUST LAWS	272
III.	THE POTENTIAL BENEFITS AND COSTS OF RESEARCH JOINT VENTURES AND PRODUCTION JOINT VENTURES	274
	A. The Inefficiencies in Market Generated R&D	274
	B. The Social Benefits and Costs of RJs	276
	C. The Lesser Benefits and Greater Costs of PJs	280
	D. Organizational Difficulties Associated with All Joint Ventures	286
IV.	CURRENT LAW AND ENFORCEMENT POLICY CONCERNING PRODUCTION JOINT VENTURES AND THEIR EFFECT ON SUCH VENTURES	289
	A. Courts Judge PJs Under the Rule of Reason	290
	B. Current Antitrust Enforcement Policy Is Hospitable Towards Legitimate PJs	293
	C. The Threat of Private Antitrust Suits Is Exaggerated	294
	D. Current Antitrust Law Has Not Deterred the Formation of an Increasing Number of PJs	296

© 1993 Donald K. Stockdale, Jr.

[†] Assistant Professor, College of Business & Management, University of Maryland. Ph.D. 1989, Yale University; J.D. 1980, Yale Law School; B.A. 1976, King's College, Cambridge University; B.A. 1974, Yale University.

V. ARGUMENTS FAVORING THE PROPOSED SPECIAL TREATMENT OF DOWNSTREAM CONSORTIA ARE FLAWED.....	297
A. The High Costs and Risks of Commercializing and Producing Innovative Products Do Not Exceed the Capacity of Most Individual Firms	298
B. The Cyclical Nature of the Development Process and Shorter Product Lives Does Not Necessitate Cooperation in Both R&D and Production.....	299
C. International Competition Will Not Eliminate the Danger of Collusion Among the Joint Venture Participants	300
D. Use of Joint Ventures in Japan and Europe Does Not Require that the U.S. Encourage Domestic Joint Ventures	302
VI. WEAKNESSES IN THE PROPOSED LEGISLATION.....	309
VII. CONCLUSION	314

I. INTRODUCTION

For more than a decade, slower productivity growth, persistently large trade deficits, and the apparent decline of the international competitiveness of U.S. firms have concerned policy-makers, business leaders and academicians.¹ In analyzing the causes of these ominous trends, many have questioned whether the U.S. antitrust laws have unduly disadvantaged domestic firms relative to their foreign competitors.

In the late 1970s, many commentators began suggesting that cooperative research and development (R&D) warranted special treatment under the antitrust laws.² Congress responded by passing the National Cooperative Research Act of 1984 (NCRA).³

1. See, e.g., *Competitiveness and Antitrust: Hearings Before the Senate Comm. on the Judiciary*, 100th Cong., 1st Sess. (1987) [hereinafter 1987 Senate Hearings]; MARTIN N. BAILEY & ALOK K. CHAKRABARTI, INNOVATION AND THE PRODUCTIVITY CRISIS (1988); WILLIAM J. BAUMOL ET AL., PRODUCTIVITY AND AMERICAN LEADERSHIP: THE LONG VIEW (1989); MICHAEL L. DERTOZOS ET AL., MADE IN AMERICA: REGAINING THE PRODUCTIVE EDGE (1989); INTERNATIONAL TRADE ADMIN., U.S. DEP'T OF COMMERCE, AN ASSESSMENT OF U.S. COMPETITIVENESS IN HIGH TECHNOLOGY INDUSTRIES (1983); NATIONAL RESEARCH COUNCIL, TECHNOLOGY, TRADE, AND THE U.S. ECONOMY (1978); PRESIDENT'S COMM'N ON INDUS. COMPETITIVENESS, GLOBAL COMPETITION: THE NEW REALITY (1985).

2. See, e.g., *Japanese Technological Advances and Possible U.S. Responses Using Research Joint Ventures: Hearings Before the Subcomm. on Investigations and Oversight and the Subcomm. on Science, Research and Technology of the House Comm. on Science and Technology*, 98th Cong., 1st Sess. (1983); *The National Productivity and Innovation Act and Related Legislation: Hearings Before the Senate Comm. on the Judiciary*, 98th Cong., 1st & 2d Sess. (1983 & 1984); INDUSTRIAL RESEARCH INST., INSTITUTIONAL AND LEGAL CONSTRAINTS TO COOPERATIVE

Recently, however, a number of academicians and business leaders have suggested that the NCRA did not go far enough. They argue that in order to improve the international competitiveness of domestic firms, Congress should enact further legislation to encourage joint ventures in downstream activities, such as production and even distribution and marketing.⁴ Responding to such arguments, members of the 101st and 102d Congresses introduced bills which, in various ways, would relax the antitrust laws for production joint ventures (PJs) and, in some cases, for distribution and marketing joint ventures.⁵

This Article argues that such proposals are misguided, and that, if implemented, they would likely undermine American competitiveness and impose significant costs on U.S. consumers. More specifically the Article contends that: (1) the potential social benefits are lower and costs higher for PJs, in comparison with research joint ventures (RJs); (2) the

ENERGY R&D (Technical Advisory Bd., U.S. Commerce Dep't, No. PB-240-929, 1975); NATIONAL RESEARCH COUNCIL, ANTITRUST, UNCERTAINTY, AND TECHNOLOGICAL INNOVATION (1980).

3. The NCRA mandated that research joint ventures (RJs), as defined in the Act, should not be deemed illegal per se, but rather should be evaluated under the rule of reason. It further provided that RJ participants would be liable in private actions for only single, rather than treble, damages if they filed a notification with the Antitrust Division and the Federal Trade Commission. Finally, it enabled RJ participants which had been sued by private plaintiffs to recover attorneys' fees and costs under certain conditions, regardless of whether the RJ had filed a notification. 15 U.S.C. §§ 4301-4305 (1988).

4. See, e.g., *Legislation Concerning Production Joint Ventures: Hearing Before the Subcomm. on Antitrust, Monopolies and Business Rights of the Senate Comm. on the Judiciary*, 101st Cong., 2d Sess. (1990) [hereinafter 1990 Senate Hearing]; *The Government Role in Joint Production Ventures: Hearing Before the Subcomm. on Science, Research and Technology of the House Comm. on Science, Space, and Technology*, 101st Cong., 1st Sess. (1989) [hereinafter 1989a House Hearing]; *Production Joint Ventures Antitrust Legislation: Hearings Before the Subcomm. on Economic and Commercial Law of the House Comm. on the Judiciary*, 101st Cong., 1st Sess. (1989) [hereinafter 1989b House Hearing]; *High Definition Television: Hearing Before the House Comm. on Science, Space, and Technology*, 101st Cong., 1st Sess. (1989) [hereinafter 1989c House Hearing]; Thomas M. Jorde & David J. Teece, *Innovation, Cooperation and Antitrust*, 4 HIGH TECH. L.J. 1 (1989) [hereinafter Jorde & Teece (1989a)]; Thomas M. Jorde & David J. Teece, *Competition and Cooperation: Striking the Right Balance*, 31 CAL MGMT. REV. 25 (1989) [hereinafter Jorde & Teece (1989b)]; Thomas M. Jorde & David J. Teece, *Innovation and Cooperation: Implications for Competition and Antitrust*, 4 J. ECON. PERSP. 75 (1990) [hereinafter Jorde & Teece (1990)].

5. In the 101st Congress the following bills were introduced: S. 952, 101st Cong., 1st Sess. (1989); S. 1006, 101st Cong., 1st Sess. (1989); H.R. 423, 101st Cong., 1st Sess. (1989); H.R. 1024, 101st Cong., 1st Sess. (1989); H.R. 1025, 101st Cong., 1st Sess. (1989); H.R. 2264, 101st Cong., 1st Sess. (1989). The House of Representatives eventually passed H.R. 4611, 101st Cong., 2d Sess. (1990). See also H.R. REP. NO. 516, 101st Cong., 2d Sess. (1990). No bill was passed in the Senate, however.

In the 102d Congress, new bills were introduced that would provide similar relief. See S. 479, 102d Cong., 1st Sess. (1991); H.R. 1604, 102d Cong., 1st Sess. (1991). See also S. REP. NO. 146, 102d Cong., 1st Sess. (1991), reprinted in 61 Antitrust & Trade Reg. Rep. (BNA) 347 (1991); H.R. REP. NO. 972, 102d Cong., 2d Sess. (1992).

antitrust laws currently permit procompetitive PJs, and, in fact, the wide employment of these joint ventures renders further relaxation unnecessary; (3) further relaxing the antitrust laws for downstream joint ventures may encourage the formation of production consortia having substantial market power; and (4) even if antitrust relief were warranted for production consortia in certain strategically important high-technology industries, none of the current legislative proposals is specifically tailored to that goal.

The Article is organized as follows. Part II describes the specific legislation proposed. Part III compares the potential social costs of RJs and downstream JVs and suggests that for PJs the potential benefits are more limited, while the potential costs are much higher. This Part further argues that production consortia in particular tend to impose significant social costs. Part IV examines existing antitrust precedents and antitrust enforcement policy and contends that, with the possible exception of joint ventures possessing substantial market power, current law does not pose an obstacle to joint venture activity. Part V addresses and criticizes certain specific arguments that have been raised in favor of relaxing the antitrust laws for joint ventures in high-technology industries. Part VI evaluates the specific legislative proposals that have been introduced and suggests that they are unlikely to achieve their purported goals.

II. CURRENT PROPOSALS TO CHANGE THE ANTITRUST LAWS

During the 102d Congress,⁶ the Judiciary Committees of the House and the Senate approved and sent to their respective floors bills that would extend the NCRA to cover production joint ventures.⁷ Although the Senate passed a slightly modified version of the bill, the House

6. In the 101st Congress, members introduced several bills that would have amended the antitrust laws to provide various protection for production joint ventures. Basically, the bills adopted one or more of the following four approaches: (1) extending the notification procedures and protections of the NCRA to joint ventures involving production (and in some cases marketing), *see H.R. 1025, 101st Cong., 1st Sess. (1989); H.R. 2262, 101st Cong., 1st Sess. (1989); S. 1006, 101st Cong., 1st Sess. (1989)*; (2) codifying in detail the substantive law applicable to innovative joint ventures, *see H.R. 1024, 101st Cong., 1st Sess. (1989); S. 2322, 101st Cong., 1st Sess. (1989)*; (3) establishing a safe harbor for PJs whose participants lack market power, *see H.R. 423, 101st Cong., 1st Sess. (1989)*; and (4) establishing a certification procedure under which joint ventures, reviewed and approved by the relevant antitrust authorities, would be exempt from any antitrust penalty or damage liability, *see H.R. 1024, 101st Cong., 1st Sess. (1989); S. 2322, 101st Cong., 1st Sess. (1989)*. See generally H.R. REP. NO. 516, *supra* note 5; Joseph F. Brodley, *Antitrust Law and Innovation Cooperation*, 4 J. ECON. PERSP. 97, 104 (1990). The House ultimately passed H.R. 4611, which adopted the first approach, but no bill reached the floor of the Senate during that Congress.

The bills introduced in the 102d Congress adopted only the first approach.

7. See H.R. 1604, 102d Cong., 1st Sess. (1991); S. 479, 102d Cong., 1st Sess. (1991).

adjourned without acting. Nevertheless, the provisions of the bills remain significant because the next Congress will likely introduce similar legislation.

Under both bills, the NCRA's definition of joint venture would expand to include "the production of a product, process or service" in addition to covering research and development activities.⁸ Thus, PJVs that qualify under the bills would receive rule of reason analysis if challenged under the antitrust laws. In addition, qualified production ventures that file a notification⁹ with the antitrust authorities would be liable only for actual, not treble, damages in actions filed by private plaintiffs. Finally, regardless of whether the venture files a notification, it would be able to recover attorneys' fees and costs if it were named a defendant in an antitrust suit and the court finds the claim was "frivolous, unreasonable, without foundation, or in bad faith."¹⁰

Both bills specifically prohibit the joint marketing of any products jointly produced by the venture.¹¹ At the same time, however, neither bill requires that a production joint venture engage in any joint R&D activities to qualify for protection.

Both bills would also add a new section directed specifically at PJVs. The new section in the House bill would exclude a PJV from protection of the Act "if at any time more than 30 percent, in the aggregate, of the beneficial ownership of the voting securities and equity of such joint venture is controlled by foreign entities." The section would also require that any facilities operated by the venture be located in the United States or its territories.¹² The Senate bill establishes two different conditions for a PJV to qualify under the Act: first, the venture must provide "substantial benefits" to the U.S. economy (such as "increased skilled job opportunities," "investments in long-term production facilities," or "participation by United States entities in the venture"); second, the production facilities of the venture must be located in the United States or in a country that accords "national treatment" to American participants in PJVs.¹³

8. H.R. 1604, *supra* note 7, § 2(b)(4); S. 479, *supra* note 7, § 2(2)(c).

9. The information required to be provided in a notification is limited. For example, the House bill only requires that the joint venture provide the identities of the participants and a brief description of the nature and objectives of the venture. *See, e.g.*, H.R. REP. NO. 516, *supra* note 273, at 19.

10. *See* S. REP. NO. 146, *supra* note 5, at 23.

11. H.R. 1604, *supra* note 7, § 2(c)(3)(B); S. 479, *supra* note 7, § 2(2)(G).

12. H.R. 1604, *supra* note 7, § 2(f). According to the Committee Report, the section is intended to "stimulate more collaborative activity by American-owned firms." H.R. REP. NO. 516, *supra* note 5, at 15.

13. S. 479, *supra* note 5, § 2(10). According to the Senate Report, the requirements are intended to ensure that the act benefits American workers. S. REP. NO. 146, *supra* note 5, at 7.

The Senate bill contains two additional provisions not found in the House version. First, the Senate bill requires that, if a joint venture uses existing facilities, those facilities must produce a "new product or technology."¹⁴ Second, the Senate bill imposes new reporting requirements on the Federal Trade Commission and the Department of Commerce.¹⁵

III. THE POTENTIAL BENEFITS AND COSTS OF RESEARCH JOINT VENTURES AND PRODUCTION JOINT VENTURES

The rationale for giving special treatment to cooperative research stems principally from certain market failures associated with market generated R&D. These market failures can create inefficiencies in the level of R&D investment, the allocation of R&D expenditures, and the dissemination of the R&D results.¹⁶ Before comparing the potential benefits and costs of R&Js relative to downstream joint ventures, it is useful to review these market failures associated with R&D.

A. The Inefficiencies in Market Generated R&D

The special problems connected with R&D activities result principally because the product of R&D activities is *information* or *knowledge*. Information resembles a public good, in that (1) the acquisition of the information by one party need not reduce its availability to others, and (2) the cost of transferring the information to others is often, though not always, low. The public good nature of information creates problems both for private firms engaging in R&D and for society as a whole.

The most widely recognized inefficiency of privately funded R&D is the generation of positive externalities: that is, the benefits of the R&D

14. S. 479, *supra* note 5, § 2(2)(G).

15. Specifically, the bill requires the FTC to prepare an annual report listing the joint ventures that had filed under the Act and any enforcement actions that had been brought by the Department of Justice against ventures filing under the Act. The bill requires the Department of Commerce to prepare triennial reports which describe the "technologies most commonly pursued by joint ventures" (and assess the competitiveness of U.S. industry in those technologies), describe the areas of production most commonly engaged in by P&Js, and review foreign laws concerning joint R&D and production. See S. 479, *supra* note 5, § 2(10).

16. See generally Gene M. Grossman & Carl Shapiro, *Research Joint Ventures: An Antitrust Analysis*, 2 J.L. ECON. & ORGANIZATION 315 (1986); Michael L. Katz, *An Analysis of Cooperative Research and Development*, 17 RAND J. ECON. 527 (1986); Michael L. Katz & Janusz A. Ordover, *R&D Cooperation and Competition*, 1990 BROOKINGS PAPERS ON ECON. ACTIVITY: MICROECONOMICS 139; Janusz Ordover & William Baumol, *Antitrust Policy and High-Technology Industries*, 4 OXFORD REV. ECON. POL'Y 13 (1988); Janusz A. Ordover & Robert D. Willig, *Antitrust for High-Technology Industries: Assessing Research Joint Ventures and Mergers*, 28 J.L. & ECON. 311 (1985).

frequently spill over from the researching firm to others. Because firms cannot appropriate the full rewards or benefits of their investment in R&D, they will tend to invest less than the socially optimal amount.¹⁷ The severity of the inappropriability and underinvestment problems increases with more basic research.¹⁸ In addition, the lumpiness of R&D inputs and economies of scale and scope in R&D may exacerbate this underinvestment.¹⁹

The patent system and trade secrecy laws are intended to alleviate this appropriability problem by assigning and enforcing property rights in the information produced by R&D.²⁰ Unfortunately, these mechanisms for increasing appropriability create other problems.

First, these mechanisms result in an inefficient *ex post* dissemination of the knowledge produced by R&D. That knowledge or information can be used simultaneously by others at little or no extra cost suggests that society should encourage its widest possible dissemination. By utilizing exclusion to increase the appropriability of knowledge, society creates inefficiencies in its *ex post* dissemination.²¹

17. In other words, the private return on investment in R&D will be less than the social return. See Kenneth J. Arrow, *Economic Welfare and the Allocation of Resources for Invention*, in THE RATE AND DIRECTION OF INVENTIVE ACTIVITY 609, 619-622 (R. Nelson ed., 1962); Richard R. Nelson, *The Simple Economics of Basic Scientific Research*, 67 J. POL. ECON. 297, 302 (1959). In addition, because a firm can gain from the R&D of others, it reduces the competitive risk of failing to conduct independent R&D. Katz & Ordover, *supra* note 16, at 39.

18. See Partha Dasgupta, *The Welfare Economics of Knowledge Production*, 4 OXFORD REV. ECON. POL'Y 1, 4 (1988); Katz, *supra* note 16, at 537; Nelson, *supra* note 17, at 302-04.

19. R&D inputs are said to be lumpy because large minimum expenditures are often required before any R&D can be performed or before such R&D can yield any useful results. See WILLIAM D. NORDHAUS, *INVENTION, GROWTH, AND WELFARE* 36 (1969); JEAN TIROLE, *THE THEORY OF INDUSTRIAL ORGANIZATION* 414 (1988); see also Partha Dasgupta, *The Economic Theory of Technology Policy: An Introduction*, in *ECONOMIC POLICY AND TECHNOLOGICAL PERFORMANCE* 9 (Partha Dasgupta & Paul Stoneman eds., 1987). In addition, there is evidence that R&D frequently exhibits significant economies of scale and scope. See NORDHAUS, *supra*, at 414; Grossman & Shapiro, *supra* note 16. Finally, imperfections in the capital markets may limit the availability of firms to obtain outside funding for R&D investment. See Paul Stoneman & John Vickers, *The Assessment: The Economics of Technology Policy*, 4 OXFORD REV. ECON. POL'Y i, viii (1988). These facts suggest that the market may not yield an efficient investment in R&D, and more particularly, that the level of investment necessary for the efficient performance of certain types of R&D may exceed the financial resources of smaller firms..

20. As many have shown, however, the patent system and trade secrecy laws in general fail to eliminate all spillovers. See, e.g., Richard C. Levin et al., *Appropriating the Returns from Industrial Research and Development*, 1987 BROOKINGS PAPERS ON ECON. ACTIVITY 783 (1987); Edwin Mansfield et al., *Social and Private Rates of Return from Industrial Innovation*, 91 Q.J. ECON. 221 (1977).

21. If the information generated is valuable enough, it may confer market power on the innovating firm, which can lead to higher prices and reduced output. This will generate the static, deadweight loss associated with monopoly. See Arnold C. Harberger, *Monopoly and Resource Allocation*, 44 AMER. ECON. REV. 77, 78 (1954); Richard R. Nelson & Sidney Winter, *The Schumpeterian Tradeoff Revisited*, 72 AMER. ECON. REV. 114, 116 (1982).

In addition, to the extent that the patent laws enable an innovator to capture a significant proportion of the social benefits in the form of profits, a race to be first may result in too many firms engaging in duplicative R&D. As a result, the patent laws may create inefficient and possibly excessive investments in R&D.²²

B. The Social Benefits and Costs of RJs

The three above-mentioned inefficient aspects of market generated R&D in turn suggest the three most significant potential benefits of cooperative research. First, RJs can help internalize the externality caused by the inappropriability of R&D and can thus increase R&D investment incentives. This internalization occurs because the RJV compels the participants to commit to sharing costs before the research is conducted and hence before any spillovers can occur. This benefit is likely to be greatest when the RJV is directed at basic research²³ or at research involving areas of limited commercial importance, such as that directed to environmental, health and safety problems, because the

In addition, restricting dissemination of information concerning the most efficient technology can raise the average production cost in the industry over that which would result with widespread use of the new technology. *Id.* Finally, denying competitors the right to use new technological information can induce them to spend research funds on inventing around the patent or on developing new technologies that are less efficient than the current best, but inaccessible technology. See Donald K. Stockdale, Jr., *Three Essays on Antitrust and Innovation* 66 (1989) (unpublished Ph.D. dissertation, Yale University). See generally Katz & Ordover, *supra* note 16, at 145.

22. This excessive investment in a particular area of research has been termed the "common pool" problem. See Partha Dasgupta & Joseph Stiglitz, *Industrial Structure and the Nature of Innovative Activity*, 90 ECON. J. 266, 279 (1980); Partha Dasgupta & Joseph Stiglitz, *Uncertainty, Industrial Structure, and the Speed of R&D*, 11 BELL J. ECON. 1, 3 (1980); J. Hirshleifer & John G. Riley, *The Analytics of Uncertainty and Information*, 17 J. ECON. LIT. 1375, 1404 (1979); Pankaj Tandon, *Rivalry and the Excessive Allocation of Resources to Research*, 14 BELL J. ECON. 152 (1983); see also Partha Dasgupta & Eric Maskin, *The Simple Economics of Research Portfolios*, 97 ECON. J. 581 (1987).

23. The National Science Foundation has defined the various categories of research and development as follows:

Basic Research—Original investigations for the advancement of scientific knowledge not having specific immediate commercial objectives, although such investigations may be in fields of present or potential interest to the . . . company.

Applied Research—Investigations directed to the discovery of new scientific knowledge having specific commercial objectives with respect to products or processes . . .

Development—Technical activities of a nonroutine nature concerned with translating research findings or other scientific knowledge into products or processes . . .

NATIONAL SCIENCE FOUND., RESEARCH AND DEVELOPMENT IN INDUSTRY: 1987, at 2 (1989).

results of such research are the least appropriable.²⁴ In addition, benefits increase when a high percentage of firms in an industry participate in basic or externalities research.²⁵

Second, by providing access to all participants, an RJV may improve the *ex post* dissemination of the information produced by the RJV.²⁶

Third, by replacing a number of independent and competing research centers with a joint facility, an RJV may reduce excessive R&D expenditures associated with a race to be first. More importantly, the RJV may eliminate wasteful duplication in research and hence use research expenditures more efficiently.²⁷

In addition to the benefits that arise from the special characteristics of R&D, RJVs also encompass the more traditional benefits of joint ventures. Like other joint ventures, RJVs permit the participants to share risks and costs, combine complementary skills and resources, and take advantage of economies of scale and scope.²⁸ These advantages are likely to prove especially appealing for smaller firms that lack the skills or resources to conduct R&D on their own.

RJVs may also impose social costs, however, by adversely affecting R&D activity and by reducing other forms of competition. These potential costs may be grouped into three types.

First, if the RJV participants are competitors in the downstream product market, the RJV may reduce the expected return to each participant and, hence, total investment in R&D. Because all cooperating firms have equal and simultaneous access to the results of the R&D, no firm will enjoy a temporary monopoly return resulting from the innovation; rather, price competition among the participants will dissipate any excess profits from the innovation, with the surplus going to consumers. Accordingly, the RJV participants may cut back on R&D investment.²⁹ This reduction is less likely to occur, however, if: (1) the

24. See Grossman & Shapiro, *supra* note 16, at 332-33; Katz, *supra* note 16, at 537; Nelson, *supra* note 17, at 302-04; see also P.S. JOHNSON, CO-OPERATIVE RESEARCH IN INDUSTRY (1973).

25. Increasing the percentage of firms participating in the RJV will reduce the externality by reducing the number of free-riders and by committing the beneficiaries of the research-generated information to share its costs *ex ante*. See Grossman & Shapiro, *supra* note 16, at 321; Stockdale, *supra* note 21, at 67. But this result will not necessarily hold for RJVs directed to applied research and development. See *infra* text accompanying notes 29-32.

26. See, e.g., Grossman & Shapiro, *supra* note 16, at 323.

27. See *id.* at 322; Katz, *supra* note 16, at 528; Ordover & Baumol, *supra* note 16, at 27.

28. 1989a House Hearing, *supra* note 4, at 68-69 (statement of Claude E. Barfield, American Enterprise Institute); DAVID C. MOWERY & NATHAN ROSENBERG, TECHNOLOGY AND THE PURSUIT OF ECONOMIC GROWTH 239 (1989); Grossman & Shapiro, *supra* note 16, at 321-22; Katz, *supra* note 16, at 528-29.

29. See Katz & Ordover, *supra* note 16, at 152, 156; Katz, *supra* note 16. The RJV may also be used to suppress innovation, where implementation of the innovation would

research has no immediate commercial objective (such as basic research), (2) the participants operate in separate downstream product markets,³⁰ (3) there are strong nonparticipants performing competing research, or (4) the RJV agreement permits participants to continue their independent R&D efforts.³¹ Finally, besides possibly reducing investment in R&D, an RJV may also reduce the productivity of the R&D performed by limiting the diversity of approaches to a research problem. This would tend to offset the gains described above.³²

Second, an RJV may reduce competition in the downstream product market(s), which will generate social costs when firms limit output and raise prices to consumers. Participants have a clear incentive to maximize joint returns to any innovation generated by the joint venture, either by cooperating in production³³ or by employing ancillary restraints, such as field of use or geographic restrictions, to restrain product market competition.³⁴ Participants may use the RJV as a forum for exchanging price and cost data in order to collude in the downstream product markets.³⁵ In addition, the RJV may serve as a means for extending cooperation or collusion into other product areas.³⁶

Third, by denying access to the RJV or to its research results, participants may disadvantage, and possibly drive from the market, actual and potential competitors. Although this is most likely to occur

impose significant costs on the participants or destabilize the industry, *see Johnson, supra* note 24, at 84, or where the RJV is used to delay meeting government environmental or safety regulations. *See United States v. Automobile Mfrs. Ass'n*, 1969 Trade Cas. (CCH) ¶ 72,907 (C.D. Cal. 1969) (consent decree), *modified*, 1982-1983 Trade Cas. (CCH) ¶ 65,088 (C.D. Cal. 1982); *LAWRENCE A. SULLIVAN, HANDBOOK OF THE LAW OF ANTITRUST* 301-03 (1977).

30. *See Katz, supra* note 16, at 529.

31. *See id.* at 542; *Grossman & Shapiro, supra* note 16, at 324.

32. CARMELA S. HAKLISCH ET AL., *TRENDS IN COLLECTIVE INDUSTRIAL RESEARCH* 153 (2d ed. 1986); MOWERY & ROSENBERG, *supra* note 28, at 240; Ordover & Baumol, *supra* note 16, at 27; *see also* John T. Scott, *Diversification Versus Cooperation in R&D Investment*, 9 *MANAGERIAL & DECISION ECON.* 173 (1988) (suggesting that NCRA may have reduced diversity and productivity of R&D effort); *cf. JOHN JEWKES ET AL., THE SOURCES OF INVENTION* 221 (1960); RICHARD R. NELSON & SIDNEY G. WINTER, *AN EVOLUTIONARY THEORY OF ECONOMIC CHANGE* 366 (1982) (discussing possible insufficient diversification of research efforts under monopoly).

33. *See infra* text accompanying note 58.

34. *See Grossman & Shapiro, supra* note 16, at 325; Katz & Ordover, *supra* note 16, at 156; *cf. F.M. SCHERER & DAVID ROSS, INDUSTRIAL MARKET STRUCTURE AND ECONOMIC PERFORMANCE* 625 (3d ed. 1990) (discussing use of cross-licensing agreements and patent pools to facilitate collusion and exclude competitors); George L. Priest, *Cartels and Patent License Agreements*, 20 *J.L. & ECON.* 309, 356-77 (1978).

35. *See Katz & Ordover, supra* note 16, at 156. The NCRA attempts to alleviate this danger by specifically excluding the exchange of cost and price data from the protection of the Act. 15 U.S.C. § 4302(b) (1988); *see also* S. REP. NO. 427, 98th Cong., 2d Sess. 15-16 (1984) (explaining reasons for exclusions from protection).

36. *See Katz & Ordover, supra* note 16, at 156; Stockdale, *supra* note 21, at 71.

where horizontal competitors make up the RJV, it may also occur where dominant firms in different markets cooperate to produce a new product or process.³⁷ This exclusionary behavior will only create significant social costs when both upstream and downstream markets are concentrated with high barriers to entry and reentry,³⁸ and when the participants can successfully collude with respect to price, R&D, and other dimensions of competition. This is frequently difficult to achieve.³⁹

Thus, RJVs may be used for anticompetitive purposes and may impose net social costs under certain circumstances. These anticompetitive effects are most likely to occur where: (1) the cooperation extends downstream to areas of competitive concern, (2) the relevant markets are concentrated and exhibit barriers to entry, (3) the combined market power of the participants is significant, and (4) collateral restraints in the agreement restrict competition among the participants.

Therefore, RJVs may have a positive or negative effect on social welfare; assessing the net welfare effect of a particular RJV requires an examination of the particular facts. Nevertheless, certain types of RJVs most likely to yield a net benefit to society are identifiable.

For example, an RJV directed at basic or precommercial research is likely to generate significant benefits without imposing substantial social costs. Such an RJV will likely increase industry expenditures on basic research by permitting the sharing of costs and risks and by internalizing

37. See, e.g., *Berkey Photo, Inc. v. Eastman Kodak Co.*, 603 F.2d 263, 299-304 (2d Cir. 1979), cert. denied, 444 U.S. 1093 (1980) (Kodak's joint development with flash manufacturer of new camera flash held to violate section 1 of the Sherman Act); cf. Janusz A. Ordover & Robert D. Willig, *An Economic Definition of Predation: Pricing and Product Innovation*, 91 YALE L.J. 8 (1981) (analyzing introduction by single firm of new, but incompatible product system, as a form of predation).

38. See Grossman & Shapiro, *supra* note 16, at 317; Stockdale, *supra* note 21, at 71-72; cf. Paul L. Joskow & Alvin K. Klevorick, *A Framework for Analyzing Predatory Pricing Policy*, 89 YALE L.J. 213, 227-33 (1979) (private incentives for, and social costs resulting from, predatory conduct will be significant only in the presence of substantial market power and barriers to entry).

39. The following factors, among others, have been identified as limiting the effectiveness of or possibilities for oligopolistic collusion: (1) a large number of competitors, (2) a large variance in the size of competitors, (3) relatively free entry conditions, (4) differentiated products, (5) differential cost structures among competitors, (6) relatively elastic demand, (7) growing demand, and (8) significant non-price competition. See SCHERER & ROSS, *supra* note 34, at 277-315; Peter Asch & Joseph J. Seneca, *Characteristics of Collusive Firms*, 23 J. INDUS. ECON. 223 (1975); George J. Stigler, *A Theory of Oligopoly*, 72 J. POL. ECON. 55 (1964), reprinted in GEORGE J. STIGLER, THE ORGANIZATION OF INDUSTRY 39-63 (1968). See generally Alexis Jacquemin & Margaret E. Slade, *Cartels, Collusion, and Horizontal Merger*, in 1 HANDBOOK OF INDUSTRIAL ORGANIZATION 415 (Richard Schmalensee & Robert D. Willig eds., 1989). In addition, it is recognized that successful collusion is especially difficult in industries subject to rapid technological change. *Id.* at 420.

the externalities associated with such research.⁴⁰ It may also increase the efficiency of the R&D by reducing duplication.⁴¹ Moreover, because basic research and precommercial R&D are distanced from the competitive concerns of the market, the RJV will not likely spur collusion. Finally, a research-directed JV will unlikely injure nonparticipants because of the significant research spillovers and because the generally long lag between generation of the idea and its commercial application gives nonparticipants time to catch up.⁴²

Likewise, if the relevant research and product markets are unconcentrated with relatively free entry, it appears improbable that an RJV made up of a nonmajority of firms in those markets or of noncompetitors can impose significant social costs, since the participants would still face stiff competition from nonparticipants in both upstream and downstream markets. Such a venture could yield substantial benefits in the form of increased R&D expenditures and increased efficiency in the performance of R&D, however.⁴³ Moreover, in evaluating the conditions of the research market, it is generally accepted that the relevant geographic market is global in scope.⁴⁴ Accordingly, even if the RJV consists of a majority of U.S. competitors, this may not result in anticompetitive effects if there are foreign competitors with sufficient research capabilities.

In summary, the NCRA justly encourages the above-described RJVs in general because the social benefits outweigh the social costs.

C. The Lesser Benefits and Greater Costs of PJVs

In contrast to RJVs, joint ventures in production and distribution offer fewer social benefits and pose greater social costs. Although the market failures associated with R&D suggest that it often makes sense to include as many participants as possible in an industry-wide research consortia, no similar arguments justify industry-wide production

40. Based on survey and interview data, Wolek found that U.S. research consortia "are significantly more committed to basic research than are competitive, industrial programs." He further found that, on average, consortia devoted 23.4% of their budget to basic research in 1974. FRANCIS W. WOLEK, THE ROLE OF CONSORTIA IN THE NATIONAL R&D EFFORT (National Science Found., NTIS No. PB-277-366, 1977). Similarly Haklisch, Fusfeld, and Levenson found that 89% of the RJVs that they surveyed performed fundamental research, and that fundamental research represented 32% of the RJVs overall activities. HAKLISCH ET AL., *supra* note 32, at 18.

41. Grossman & Shapiro, *supra* note 16, at 333.

42. *Id.*; Katz, *supra* note 16, at 537.

43. See Grossman & Shapiro, *supra* note 16, at 326; Katz, *supra* note 16, at 540.

44. See, e.g., Antitrust Div., U.S. Dep't of Justice, Antitrust Guidelines for International Operations-1988, 4 Trade Reg. Rep. (CCH) ¶ 13,109.10, at 20589-3 (1989) [hereinafter International Guidelines]; William F. Baxter, *Antitrust Law and Technological Innovation*, 1 ISSUES SCI. & TECH. 80, 85 (1985); Ordover & Baumol, *supra* note 16, at 30.

consortia. To illustrate these differences, this Section will focus on production and distribution joint ventures that do not involve cooperation in research.⁴⁵

Because the production and distribution of goods and services do not suffer from the same market failures affecting R&D, the major justifications for research cooperation—internalizing the externalities associated with R&D, improving the *ex post* dissemination of research results, and eliminating wasteful duplication of research efforts—do not apply to PJVs. Rather the potential advantages of domestic PJVs⁴⁶ are considerably more narrow.

First, a PJV may permit the realization of economies of scale or scope, where the minimum efficient scale of a plant is beyond the capacity of individual companies or is large relative to total demand.⁴⁷ Empirical studies generally agree, however, that the minimum efficient scale of plant is small relative to market size in the vast majority of industries, and this ratio of scale to market size has been declining over time in many industries.⁴⁸ Furthermore, even in the most scale-intensive industries, numerous competing production facilities can coexist.⁴⁹ This suggests that, although economies of scale may justify production joint ventures between two or three smaller firms,⁵⁰ they do not justify

45. This limitation is chosen not only for expositional simplicity, but also because most of the bills currently being considered by Congress would not require PJVs to perform cooperative research to qualify for protection.

46. Joint ventures involving United States and foreign firms may be based on additional motivations, most importantly, the desire to gain access to foreign markets. See MICHAEL E. PORTER, THE COMPETITIVE ADVANTAGE OF NATIONS 66 (1990); David C. Mowery, *Collaborative Ventures Between U.S. and Foreign Manufacturing Firms: An Overview*, in INTERNATIONAL COLLABORATIVE VENTURES IN U.S. MANUFACTURING 12-15 (David C. Mowery ed., 1988); MOWERY & ROSENBERG, *supra* note 28, at 248-50.

47. See, e.g., J. PETER KILLING, STRATEGIES FOR JOINT VENTURE SUCCESS 7-8 (1983); PORTER, *supra* note 46, at 66; Robert Pitofsky, *Joint Ventures under the Antitrust Laws: Some Reflections on the Significance of Penn-Olin*, 82 HARV. L. REV. 1007, 1015 (1969); Carl Shapiro & Robert D. Willig, *On the Antitrust Treatment of Production Joint Ventures*, 4 J. ECON. PERSP. 113, 114 (1990).

48. See, e.g., C.F. PRATTEN, ECONOMIES OF SCALE IN MANUFACTURING INDUSTRY (1971) (compilation of studies for 25 industries); Leonard W. Weiss, *Optimal Plant Size and the Extent of Suboptimal Capacity*, in ESSAYS ON INDUSTRIAL ORGANIZATION IN HONOR OF JOE S. BAIN (Robert T. Masson & P. David Qualls eds., 1975). See generally SCHERER & ROSS, *supra* note 34, at 111-20 (reviewing empirical studies of minimum efficient scale relative to market size).

49. For example, in Japan there are nine competing automobile producers, six competing manufacturers of mainframe computers, and 34 competing producers of semiconductor chips. PORTER, *supra* note 46, at 412; cf. 1989b House Hearing, *supra* note 4, at 379 (statement of George Gilder) (in industries in which the Japanese surpassed the United States, "they had at least four times as many competitors in the marketplace").

50. Even for joint ventures among small numbers of firms there is reason to question the importance of scale economies as a motivating factor. For example, Mariti and Smiley found that only 11 of 70 cooperative agreements studied indicated that achieving economies of scale in production was a major motivating factor, and of those 11, six were

production consortia involving many or most of the firms in an industry.⁵¹

Second, PJVs permit the sharing of costs and risks, especially in cases involving uncertain demand or a new technology.⁵² The risks involved in producing a new product are generally significantly less, however, than the risk that basic or fundamental applied research will yield a reasonable return.⁵³

Third, PJVs may generate synergies resulting from the sharing of complementary assets and skills of the participants.⁵⁴ Again, however, the synergies resulting from joint production should not exceed those from joint research. Nor does it appear, in general, that a joint venture requires large numbers of cooperating firms to achieve such synergies.⁵⁵

Finally, RJV participants may benefit by extending cooperation from research into production. As previously noted, firms cooperating in R&D may dissipate any returns from the R&D by competing among themselves in the downstream product market.⁵⁶ To avoid such dissipation, firms may agree to cooperate in producing and/or marketing the results of the R&D or to limit downstream competition through the use of collateral restraints, such as geographic or field-of-use restrictions. Although such strategies are facially anticompetitive, they may be necessary to secure cooperation among the participants. Moreover, they

in the automobile industry. P. Mariti & R.H. Smiley, *Cooperative Agreements and the Organization of Industry*, 31 J. INDUS. ECON. 437, 445 (1983); cf. Jeffrey Pfeffer & Phillip Nowak, *Patterns of Joint Venture Activity: Implications for Antitrust Policy*, 21 ANTITRUST BULL. 315, 328 (1976) (in a survey of 163 joint ventures, the median level of assets and sales of participants exceeded \$500 million, suggesting that firms of this size did not require joint ventures to achieve economies of scale).

51. See 1987 Senate Hearings, *supra* note 1, at 128 (statement of Richard C. Levin).

52. See 1990 Senate Hearing, *supra* note 4, at 23 (statement of Assistant Attorney General James F. Rill); Shapiro & Willig, *supra* note 47, at 114.

53. In the former case, the risks concern whether the product will prove commercially successful. In the latter case, however, additional uncertainties exist concerning whether the research will yield any information that could lead to a new product or process, in addition to generally longer lag times before these uncertainties are resolved. See EDWIN MANSFIELD ET AL., THE PRODUCTION AND APPLICATION OF NEW INDUSTRIAL TECHNOLOGY 22-32 (1977); MOWERY & ROSENBERG, *supra* note 28, at 214; Dasgupta, *supra* note 18, at 6; see also Arrow, *supra* note 17, at 616 (discussing uncertainty connected with basic research); Nelson, *supra* note 17, at 298-300.

54. 1990 Senate Hearing, *supra* note 4, at 24 (statement of Assistant Attorney General James F. Rill); Shapiro & Willig, *supra* note 47, at 114.

55. Empirical studies of PJVs suggest that most involve a small number of firms. See, e.g., SANFORD V. BERG ET AL., JOINT VENTURE STRATEGIES AND CORPORATE INNOVATION 35 (1982) (in survey of chemical joint ventures, 90% had only two parents); Albert N. Link & Gregory Tassey, *Editors' Introduction to COOPERATIVE RESEARCH AND DEVELOPMENT: THE INDUSTRY-UNIVERSITY-GOVERNMENT RELATIONSHIP* vii-viii (Albert N. Link & Gregory Tassey eds., 1989) (two-firm joint ventures are most common for applied R&D).

56. See *supra* text accompanying notes 29-32.

are unlikely to impose significant social costs if the participants are few in number and collectively lack market power.⁵⁷

While the potential benefits of PJVs appear less than those of RJs, the potential anticompetitive effects are far greater. Most importantly, the PJV may have anticompetitive effects in the relevant product market. Where the participants are horizontal competitors and the joint venture controls a major portion of the production assets in the market, the participants will have a clear incentive to maximize their joint profits by reducing output and increasing price.⁵⁸

The PJV also increases the likelihood of either tacit or explicit collusion among the participants in other downstream product markets⁵⁹ and in upstream research markets.⁶⁰ Discussions concerning the appropriate prices for the joint venture's products may lead to discussions and collusion concerning the prices charged for products the participants

57. See Baxter, *supra* note 44, at 89-91; Grossman & Shapiro, *supra* note 16, at 332; Ordover & Baumol, *supra* note 16, at 30.

58. See, e.g., Katz & Ordover, *supra* note 16, at 156; Katz, *supra* note 16, at 541; Shapiro & Willig, *supra* note 47, at 114-15; cf. Joseph F. Brodley, *Joint Ventures and Antitrust Policy*, 95 HARV. L. REV. 1521, 1552 (1982) ("Of all joint ventures, the horizontal is inherently the most anticompetitive . . . [T]he parents, through their representatives in the joint venture, will necessarily agree on prices and output in the very market in which they themselves operate.").

Even if the participants distribute the joint venture's product independently, they can accomplish the same socially costly goal by raising the price at which the joint venture transfers its product to the participants.

Moreover, even if the participants did not control the pricing of the joint venture's product and do not coordinate their actions, their common ownership interests result in the internalization of a competitive externality which can lead to an increase in price-cost margins. Robert J. Reynolds & Bruce R. Snapp, *The Competitive Effects of Partial Equity Interests and Joint Ventures*, 4 INT'L J. INDUS. ORGANIZATION 141, 142 (1986). See generally Timothy F. Bresnahan & Steven C. Salop, *Quantifying the Competitive Effects of Production Joint Ventures*, 4 INT'L J. INDUS. ORGANIZATION 155 (1986) (examining effect on competitive incentives of non-cooperating oligopolists participating in joint ventures under alternative control arrangements).

59. See, e.g., Daniel R. Fusfeld, *Joint Subsidiaries in the Iron and Steel Industry*, 48 AMER. ECON. REV. 578, 585 (1958) (hypothesizing that joint ventures could be a mechanism through which emerging industries could be dominated by existing large firms in related industries); Walter J. Mead, *The Competitive Significance of Joint Ventures*, 12 ANTITRUST BULL. 819, 820-21 (1967) (finding that joint ventures formed to bid on government-owned property resulted in restrained bidding on subsequent bids).

60. If participants in a PJV collectively account for a large percentage of the competitors in the relevant product market, and if the participants do not independently manufacture goods that compete with the joint venture's products, then this may result in a significant reduction in research effort, since the participants need not worry that they will be preempted by new products resulting from other participants' research. See PORTER, *supra* note 46, at 621; cf. MOWERY & ROSENBERG, *supra* note 28, at 99 (weak British antitrust policy between the wars led to price and market-sharing agreements and "undercut the incentives for the pursuit of competitive advantage through innovation").

manufacture independently.⁶¹ Further, the joint venture will reduce the likelihood that individual participants would attempt to cheat on any collusive agreement because the ongoing relationship creates disincentives.⁶² The likelihood of collusion, moreover, is generally recognized as significantly greater in the case of production and distribution joint ventures involving direct competitors than with RJs involving direct competitors, especially RJs directed to basic or precompetitive research.⁶³ Such collusion will also be more likely when the combined market power of the participants is greater and the barriers to entry in the affected markets are higher.

Finally, the PJV may injure competition by excluding non-participants from an essential input. This "essential facilities" problem will most likely occur where competitors possessing market power organize a vertical joint venture to supply a particular, relatively unavailable, input. It may also occur, however, in the case of a horizontal joint venture, where participants deny competitors access to new technology or to a more efficient marketing facility.⁶⁴

61. Brodley, *supra* note 58, at 1530-31; Jacquemin & Slade, *supra* note 39, at 438-39; Pitofsky, *supra* note 47, at 1030. Econometric analyses of a large sample of U.S. joint ventures in a number of industries further suggest that where the participants are horizontal competitors, a potential for market-power augmentation exists. Sanford V. Berg & Philip Friedman, *Impacts of Domestic Joint Ventures on Industrial Rates of Return: A Pooled Cross-Section Analysis, 1964-1975*, 63 REV. ECON. & STAT. 293, 295 (1981); Jerome L. Duncan, Jr., *Impacts of New Entry and Horizontal Joint Ventures on Industrial Rates of Return*, 64 REV. ECON. & STAT. 339 (1982).

62. Brodley, *supra* note 58, at 1530-31; Reynolds & Snapp, *supra* note 58, at 148-49; see also Richard N. Clarke, *Collusion and the Incentives for Information Sharing*, 14 BELL J. ECON. 383, 384 (1983) (pooling of information "makes cheating more difficult and collusive quantity restriction more effective by improving the accuracy of every firm's market estimates"); cf. Walter Adams & James W. Brock, *The "New Learning" and the Euthanasia of Antitrust*, 74 CAL. L. REV. 1515, 1527-37 (1986) (discussing use of transnational joint ventures to solidify cartels and enforce oligopolistic collusion).

63. See, e.g., KATHRYN R. HARRIGAN, *STRATEGIES FOR JOINT VENTURES* 380 (1985); Grossman & Shapiro, *supra* note 16, at 334; David C. Mowery, *Collaborative Research and High-Temperature Superconductivity*, in *COOPERATIVE RESEARCH AND DEVELOPMENT*, *supra* note 55, at 151; MOWERY & ROSENBERG, *supra* note 28, at 241; Ordover & Baumol, *supra* note 16, at 30; Section of Antitrust Law, A.B.A., *Recommendations and Report on Production Joint Venture Legislation* 6 (Sept. 1, 1989) (unpublished manuscript, on file with the author) [hereinafter ABA Production Joint Venture Report].

64. See Brodley, *supra* note 58, at 1532; Jacquemin & Slade, *supra* note 39, at 439; Lawrence A. Sullivan, *The Viability of the Current Law on Horizontal Restraints*, 75 CAL. L. REV. 835, 868 (1987); see also 1989b House Hearing, *supra* note 4, at 199, 359, 374 (statements of Dr. T.J. Rogers, President and Chief Executive Officer of Cypress Semiconductor Corporation; Mr. D.R. Coelho, Chairman of Vantage Analysis Systems, Inc.; and Dr. L.R. Tomasetta, President and Chief Executive Officer of Vitesse Semiconductor Corporation) (detailing disadvantages of entrepreneurial firms when research consortia begin performing competing research). See generally PHILLIP E. AREEDA & HERBERT HOVENKAMP, *ANTITRUST LAW: 1990 SUPPLEMENT* ¶¶ 736.1-2 (1990) (discussing case law and applications of essential facilities doctrine).

As in the case of RJs, it is difficult to accurately predict whether a particular PJV is socially beneficial or socially costly without examining the specific characteristics of the participants, the markets involved, and the joint venture agreement itself. Nevertheless, certain generalizations can be made.

For example, it is generally recognized that anticompetitive effects are more probable where the participants are horizontal competitors,⁶⁵ although such effects are not limited to such ventures. Anticompetitive effects are also more likely where the relevant market is concentrated and exhibits entry barriers and where the participants collectively account for a significant portion of the market.⁶⁶ This suggests that production consortia involving a majority of the firms in an industry pose a special antitrust risk.

Therefore, the nature and structure of the joint venture and possible collateral restraints in the joint venture agreement can affect the likelihood that it will impose a net social cost. For example, a distribution or marketing JV is more likely to have anticompetitive effects than a production JV, since it prevents the participants from competing in marketing their products.⁶⁷ In addition, collateral restraints in the joint venture agreement may limit competition among the participants. For example, the joint venture agreement may contain field of use or geographic restrictions in intellectual property licenses.⁶⁸ Alternatively, the agreement may simultaneously prohibit the participants from independently manufacturing products that compete with those produced by the venture while limiting the amount of the venture's product that is distributed to each participant.⁶⁹ Also, collateral restraints that restrict distribution of the venture's product to the participants may disadvantage nonparticipants.⁷⁰ Thus, while PJs offer smaller potential

65. See Brodley, *supra* note 58, at 1552; Pitofsky, *supra* note 47, at 1031; cf. Fusfeld, *supra* note 59; Mead, *supra* note 59 (discussing possible anticompetitive effects of horizontal joint ventures in the iron and steel industry and in the bidding for oil and gas leases).

66. See International Guidelines, *supra* note 44, at 20,600; Brodley, *supra* note 58, at 1541-42.

67. See, e.g., 1990 Senate Hearing, *supra* note 4, at 63 (letter from James F. Rill, Assistant Attorney General, Antitrust Div., to Sen. Metzenbaum); 1989b House Hearing, *supra* note 4, at 129 (statement of Edward Rock); Brodley, *supra* note 58, at 1555-56; ABA Production Joint Venture Report, *supra* note 63, at 34. But see *supra* note 58.

68. See Grossman & Shapiro, *supra* note 16, at 329; Stockdale, *supra* note 21, at 80.

69. Cf. Brodley, *supra* note 58, at 1560-61.

If the parent must procure the input from the joint venture, regulation of the joint venture's output effectively controls the output of the parents. Moreover, in establishing the production level of the joint venture, the parents necessarily reveal their own output plans and thus diminish the uncertainty necessary for effective competition

Id.

70. See *id.* at 1563-65; ABA Production Joint Venture Report, *supra* note 63, at 14.

social benefits than RJs, they pose significantly higher social costs. This especially holds for production consortia consisting of a significant number of horizontal competitors.

D. Organizational Difficulties Associated with All Joint Ventures

In evaluating public policies that may encourage joint ventures, one must also consider the organizational difficulties and transaction costs associated with this form of business organization. These organizational difficulties will influence not only the types of joint ventures formed, but also the likely balance of social benefits and costs that will result. In addition, these difficulties will likely limit both the number of PJs formed and their likely success.

In attempting to overcome these organizational difficulties, joint venture participants will frequently attempt to limit competition among themselves. While this should not impose social costs for ventures involving small numbers of firms that collectively lack market power, it can pose dangers for joint ventures involving a large number of firms, especially where the firms are cooperating in production. In such cases, the major purpose of the venture may be to eliminate competition rather than to achieve efficiencies.

It is widely recognized that the presence of multiple participants makes management of joint ventures extremely difficult,⁷¹ and these difficulties tend to increase as the number of participants increases.⁷² As a result, decisionmaking tends to become slower and more cumbersome than in other forms of organization.⁷³

71. Joint venture participants often disagree on such fundamental matters as the goals of the venture, likely developments in technology or the market, and the relative contributions of the parents. *See, e.g.*, KILLING, *supra* note 47, at 8; PORTER, *supra* note 46, at 66; MOWERY & ROSENBERG, *supra* note 28, at 247; Michael E. Porter & Mark B. Fuller, *Coalitions and Global Strategy in COMPETITION IN GLOBAL INDUSTRIES* 326 (Michael E. Porter ed., 1986).

72. In the case of research consortia involving large numbers of participating firms, the Department of Commerce has estimated that one year is "the minimum time required to reach agreement on the research agenda and other management issues." Link & Tassey, *supra* note 55, at xix; *see also* 1989a House Hearing, *supra* note 4, at 74 (statement of Claude E. Barfield, American Enterprise Institute) ("Organizational difficulties will be [sic] tend to vary inversely with the number of firms involve [sic]: the more firms involved, the more illusive [sic] a consensus on agenda, increased potential for conflict in business cultures, and increased likelihood that the purpose of [the] venture will be defeated."); George R. Heaton, Jr., *The Truth About Japan's Cooperative R&D*, 5 ISSUES SCI. & TECH. 32, 37 (1988) (among Japanese RJs, "[a]s the membership increases, the difficulties in agreeing on a technical agenda rise proportionately; the larger the group, the less ambitious and more basic the research aims tend to be").

73. Based on a study of 37 joint ventures, J. Peter Killing found that management problems occurred not only at the board level of the parent firms but also at the management level of the joint venture itself. This latter problem occurred because the

That the venture itself can pose a competitive threat to its parents further complicates the management of the venture. The venture may itself become a competitor to one or more of the parents or may increase the competitive strength of one parent relative to the others.⁷⁴ More importantly, although the joint venture may depend on technological transfer, the participants are frequently reluctant to share strategic technological information.⁷⁵ In addition, firms often attempt to free-ride by contributing their less able personnel or by withholding their most advanced technology.⁷⁶ In other cases, participants will vigorously attempt to prevent the disclosure of proprietary information to their partners.⁷⁷

These problems in turn influence the structure of joint ventures. For example, in order to minimize competitive threats to participants, research consortia have tended to focus on basic research, pre-competitive research or non-competitive research and have eschewed applied

management staff of the joint venture tended to be drawn from the various parent organizations, and the working relationship among managers from different parents tended to be strained and inefficient. KILLING, *supra* note 47, at 9-10. A subsequent study of over 400 joint ventures found that decisionmaking was more cumbersome in joint ventures compared with a wholly-owned subsidiary and that it was "more difficult to get something done quickly." HARRIGAN, *supra* note 63, at 373. It further found that joint ventures having a 50-50 ownership split were disfavored by some parent managers, because this further slowed and complicated decisionmaking. *Id.* at 368.

74. See PORTER, *supra* note 46, at 66; Porter & Fuller, *supra* note 71, at 326. In some cases, a product developed independently by one of the participants may compete with jointly developed products, leading to the demise of the venture. See MOWERY & ROSENBERG, *supra* note 28, at 247; see also Mariti & Smiley, *supra* note 50, at 446 (giving examples of joint venture participants that were injured by the joint venture itself or by their partners).

75. See, e.g., HARRIGAN, *supra* note 63, at 344-47; Stockdale, *supra* note 21, at 252-53.

76. See, e.g., MOWERY & ROSENBERG, *supra* note 28, at 225; Shapiro & Willig, *supra* note 47, at 114; Stockdale, *supra* note 21, at 252. Microelectronics and Computer Technology Corporation (MCC) presents a clear example of this problem. Initially, MCC was designed to be operated with a staff drawn from its member companies. The members, however, sent their less able researchers. After MCC rejected 90% of the researchers sent by the member firms, it staffed its laboratories primarily with outside personnel. See HARRIGAN, *supra* note 63, at 231; MOWERY & ROSENBERG, *supra* note 28, at 270; cf. ALBERT N. LINK & LAURA L. BAUER, COOPERATIVE RESEARCH IN U.S. MANUFACTURING: ASSESSING INITIATIVES AND CORPORATE STRATEGIES 95 (1989) (discussing complaints of chemical firm researchers participating in RJsVs, who believed that none of the participants were sending their best scientists). Moreover, these same problems appear to have plagued the much-touted Japanese research joint ventures. See *id.* at 225; PORTER, *supra* note 46, at 635.

77. Frequently, participants will insist on confidentiality agreements that prevent a joint venture from disclosing information concerning one participant to another. See HARRIGAN, *supra* note 63, at 344-45; Stockdale, *supra* note 21, at 251. In other cases, such as the joint development of the International Aero Engines V2500 jet engine, the partners will each be assigned separate development of particular components in order to minimize technology transfer, even if this causes inefficiency. See DAVID C. MOWERY, ALLIANCE POLITICS AND ECONOMICS: MULTINATIONAL JOINT VENTURES IN COMMERCIAL AIRCRAFT 94-95 (1987).

research and development activities.⁷⁸ Firms appear to view applied RJs as a second or third best alternative and participate only when they cannot accomplish a task on their own.⁷⁹ Even then, they will seek to limit the number of partners to those absolutely necessary to accomplish the objective—usually two or three.⁸⁰ The predominance of two- and three-firm joint ventures appears to reflect both an attempt to minimize coordination problems and to reduce the likelihood that a partner would lose any competitive advantage to other partners.

As previously indicated, where two or three firms participate in a joint venture, they frequently will seek to protect their strategic technological knowledge and capabilities from their partners.⁸¹ In addition, participants often will try to minimize internal competition that can dissipate joint profits. Thus, they may extend cooperation from R&D through production and even distribution and marketing.⁸² Participants also frequently choose as partners firms that are not direct competitors,

78. See HAKLISCH ET AL., *supra* note 32, at 2; Link & Tassey, *supra* note 55, at viii, xii. Despite this focus, many firms have refused to participate, either because they believe they will be able to gain access to the research results without participating or because they fear disclosing sensitive proprietary information. WOLEK, *supra* note 40, at 130-33, 151-59; Johnson, *supra* note 24, at 80-81. And those who do join appear willing to contribute only modest amounts to the cooperative effort. See Stockdale, *supra* note 21, at 252; cf. HAKLISCH ET AL., *supra* note 32, at 16 (cooperative research at research consortia accounted for only 1.2% of total national R&D expenditures by industry in 1982).

79. Based on a questionnaire survey and interviews of firms participating in joint ventures, Professor Berg and colleagues concluded:

The ranking of alternatives to JVs was similar across firms. A wholly controlled internal project was deemed most preferable, everything else being equal. Where feasible, a merger ran second as a way to enter new markets. Since joint ventures provide an equity position, they were preferred to licensing by some firms; others asserted that the coordination problems of a joint venture render that interfirm linkage undesirable.

BERG ET AL., *supra* note 55, at 45 (1982); cf. HARRIGAN, *supra* note 63, at 56-57 (based on a study of over 400 joint ventures, the author concluded that the "most likely candidates for joint ventures are firms that lack the capabilities, strengths, or resources needed to exploit business opportunities alone," and that joint ventures "will not occur unless firms need to diversify, acquire new skills and resources, consolidate their positions, or attain objectives that they cannot reach alone"); PORTER, *supra* note 46, at 67 ("Alliances [of which joint ventures are a type] appear to be most common among second-tier competitors or companies trying to catch up.").

80. See *supra* note 55; cf. LINK & BAUER, *supra* note 76, at 28 (survey of early filings under NCRA found that degree of appropriability and average number of joint venture participants are inversely related).

There are three major exceptions to this generalization: Bell Communications Research, Inc. (Bellcore), the Electric Power Research Institute, and the Gas Research Institute. These RJs conduct not only significant basic research but also considerable applied research. It is plausible that the participants agreed to cooperate in applied research in these cases, because they are all subject to regulation and hence do not view themselves as competitors. See HAKLISCH ET AL., *supra* note 32, at 200.

81. See *supra* note 77.

82. Katz & Ordover, *supra* note 16, at 156; Stockdale, *supra* note 21, at 80.

such as firms that operate in related industries, that focus on different market niches, or that are located in different geographic markets.⁸³ Finally, as noted above, participants, in order to limit competition, may include collateral restraints in the joint venture agreement, such as field of use or geographic restrictions.⁸⁴

Such attempts to limit dissipation of profits should not pose significant anticompetitive risks where the market is competitive and the participants collectively lack market power. Significant antitrust concerns can arise, however, where the participants individually possess market power or where, as in the case of production consortia, a significant proportion of firms in the industry cooperate. These concerns are heightened by the significant possibility that, in such cases, one of the main purposes of the joint venture may be to facilitate or enforce collusion.

Thus, the organizational difficulties associated with joint ventures not only reduce their potential for achieving significant economies, but also frequently induce the participants to limit competition among themselves. This, in turn, raises the possibility of significant anti-competitive effects when the joint venture participants individually or collectively possess market power.

In summary, while PJVs offer much more limited benefits than RJs, they also create significantly greater dangers to competition. Therefore, serious doubt exists as to whether PJVs should receive the same favorable treatment accorded RJs.

IV. CURRENT LAW AND ENFORCEMENT POLICY CONCERNING PRODUCTION JOINT VENTURES AND THEIR EFFECT ON SUCH VENTURES

Proponents of the proposed PJV legislation contend that uncertainty over the legality of PJVs and the possibility that courts will condemn a JV as per se illegal deters potential PJV formation.⁸⁵ This concern appears exaggerated.⁸⁶ Current antitrust law and enforcement policy clearly

83. *Id.* at 253; cf. William G. Ouchi & Michele K. Bolton, *The Logic of Joint Research and Development*, 30 CAL. MGMT. REV. 9, 27 (1988) ("Most of the inter-corporate R&D which has occurred consists of contractual joint development of a new, applied technology by two companies at different stages in the vertical stream of an industry.").

84. See *supra* text accompanying notes 56-57.

85. See, e.g., 1990 Senate Hearing, *supra* note 4, at 35 (statement of Robert A. Mosbacher, Secretary of Commerce); 1989b House Hearing, *supra* note 4, at 57, 185-87 (statements of Rep. Thomas Campbell & Gordon E. Moore); Jorde & Teece (1989a), *supra* note 4, at 38-41.

86. In contrast to PJVs today, evidence exists that, in the early 1980s, uncertainty concerning the legality of *research* joint ventures may have deterred their formation and justified the special treatment accorded them under the NCRA. There were several factors that contributed to this uncertainty on the part of potential RJV participants. First, in one of the few decided cases involving RJs, the Government had successfully challenged a

indicate that legitimate PJVs will be judged under the rule of reason and that procompetitive PJVs will not be condemned. In addition, substantive and procedural changes in the law have reduced the threat of private antitrust challenges. Finally, the large numbers of PJVs that have been formed in recent years belies the need for any special treatment.⁸⁷

A. Courts Judge PJVs Under the Rule of Reason

Recent Supreme Court decisions clearly indicate that bona fide joint ventures—*i.e.*, joint ventures that are not merely shams to cover anticompetitive collusion—will be evaluated under the rule of reason.⁸⁸

RJV, made up of the three major automobile manufacturers, which had sought to develop automobile pollution reduction technology. *See United States v. Automobile Mfrs. Ass'n*, 1969 Trade Cas. (CCH) ¶ 72,907 (C.D. Cal. 1969); *see also SULLIVAN, supra* note 29, at 301-03. Second, in several business review proceedings, the Antitrust Division either indicated an intention to challenge proposed RJVs should the participants pursue their plans, *see ANTITRUST DIV., U.S. DEP'T OF JUSTICE, ANTITRUST GUIDE CONCERNING RESEARCH JOINT VENTURES* app. B., at 5, 9 (1980) [hereinafter ANTITRUST RJV GUIDE], made burdensome requests for information and unreasonably delayed in giving consent, *see Stockdale, supra* note 21, at 201-02, 228, 254; ABA Production Joint Venture Report, *supra* note 67, at 22 n.4, or refused to provide a definite answer, ANTITRUST RJV GUIDE, *supra*, app. B, at 13-15. Third, the Antitrust Division's 1980 *Antitrust Guide Concerning Research Joint Ventures*, while ostensibly intended to reduce legal uncertainty and encourage cooperative research, may have had the opposite effect. For example, the *Antitrust Guide's* general opposition to industry-wide research consortia, *id.* at 11, its statement that certain collateral restraints are per se illegal, *id.* at 14-15, and its suggestion that denying competitors access to a RJV, either *ex ante* or *ex post*, might constitute a per se violation of section 1 of the Sherman Act, *id.* at 22, could easily have discouraged firms from participating in RJVs. *See Stockdale, supra* note 21, at 114-22; *see also 1990 Senate Hearing, supra* note 4, at 93 (statement of Joseph Brodley). Finally, the courts had yet to issue certain decisions clarifying the appropriate method for analyzing joint ventures. *See infra* Section IV.B.

Further support for this view is found in the fact that 145 RJVs were notified in the first five years after passage of the NCRA, while only 21 RJVs were formed in the three years prior to the NCRA. *See Brodley, supra* note 6, at 100.

87. *See infra* Section IV.D.

88. *See Northwest Wholesale Stationers, Inc. v. Pacific Stationery & Printing Co.*, 472 U.S. 284 (1985); *National Collegiate Athletic Ass'n v. Board of Regents of Univ. of Okla.*, 468 U.S. 85 (1984); *Broadcast Music, Inc. v. Columbia Broadcasting Sys.*, 441 U.S. 1 (1979).

Even before these decisions were rendered, the Supreme Court had generally applied the rule of reason in evaluating joint ventures. *See, e.g.*, *United States v. Penn-Olin Chem. Co.*, 378 U.S. 158 (1964); *see also E. THOMAS SULLIVAN & JEFFREY L. HARRIS, UNDERSTANDING ANTITRUST AND ITS ECONOMIC IMPLICATIONS* 102 (1988) (Supreme Court sanctioned "broad rule of reason . . . for joint ventures" in *Chicago Bd. of Trade v. United States*, 246 U.S. 231 (1918)); *Brodley, supra* note 58, at 1534-35 ("Although joint ventures may be challenged under each of the three major antitrust statutes . . . [t]he guiding legal principle is the Rule of Reason, except in cases of flagrant cartel practices . . .").

Nevertheless, in some earlier cases, the Supreme Court held ancillary restraints to joint marketing efforts to be per se illegal. *See United States v. Topco Assoc., Inc.*, 405 U.S. 596 (1972); *United States v. Sealy, Inc.*, 388 U.S. 350 (1967); *Timken Roller Bearing Co. v. United States*, 341 U.S. 593 (1951). Although no decision has explicitly overruled these cases, a number of scholars have suggested that subsequent Supreme Court decisions have implicitly overruled them. *See, e.g.*, *Rothery Storage & Van Co. v. Atlas Van Lines*,

In *Broadcast Music, Inc. v. Columbia Broadcasting System, Inc. (BMI)*,⁸⁹ the Supreme Court reviewed a court of appeals decision holding that the blanket licenses issued by defendants BMI and the American Society of Composers, Authors and Publishers (ASCAP) to the television networks constituted a form of price fixing that was illegal per se under the Sherman Act. Rejecting the Court of Appeals' "literal" approach to price fixing, the Court stated:

Literalness is overly simplistic and often overbroad. When two partners set the price of their goods or services they are literally "price fixing," but they are not per se in violation of the Sherman Act.⁹⁰

Emphasizing that agreements that may increase economic efficiency or competition should not be held per se illegal, the Court held that the blanket license should be evaluated under the rule of reason, because the license "is not a 'naked restrain[t] of trade with no purpose except stifling of competition,' . . . but rather accompanies the integration of sales, monitoring, and enforcement against unauthorized copyright use."⁹¹ The Court further noted that, absent a blanket license, copyright holders might find it too expensive to enter into individual sales contracts or individually to monitor and enforce their copyrights.⁹²

The Supreme Court's decision in *National Collegiate Athletic Association v. Board of Regents of University of Oklahoma* (NCAA)⁹³ reaffirmed the analytical approach adopted in *BMI*. In *NCAA*, certain colleges challenged the NCAA's policy which limited the total number, and number-per-college, of televised intercollegiate football games and which prohibited any member college from selling television rights independently. Although it found that the plan was a horizontal restraint involving both prices and output,⁹⁴ the Supreme Court nevertheless rejected the per se approach adopted by the court of appeals. Instead, it applied a rule of reason analysis, because the "case involve[d] an industry in which horizontal restraints on competition are essential if the product is to be available at all."⁹⁵ The Court also recognized that intercollegiate sports required a "myriad of rules" to function and that the NCAA

792 F.2d 210, 226-27 (D.C. Cir. 1986) (Bork, J.), cert. denied, 479 U.S. 1033 (1987); Martin B. Louis, *Restraints Ancillary to Joint Venture and Licensing Agreements: Do Sealy and Topco Logically Survive Sylvania and Broadcast Music?*, 66 VA. L. REV. 879, 880 (1980).

89. 441 U.S. 1 (1979).

90. *Id.* at 9.

91. 441 U.S. at 20 (quoting *White Motor Co. v. United States*, 372 U.S. 253, 263 (1963)).

92. *Id.* at 10.

93. 468 U.S. 85 (1984).

94. *Id.* at 97.

95. *Id.* at 101.

played a vital role in enforcing the rules and preserving the character of the game.⁹⁶

Finally, in *Northwest Wholesale Stationers, Inc. v. Pacific Stationery & Printing Co.*⁹⁷, the Supreme Court refused to apply a per se rule to a concerted refusal to deal by a wholesale purchasing cooperative. Although it stated that a per se rule was appropriate where a concerted refusal to deal involved a "joint effort . . . to disadvantage competitors" and was unlikely to "enhance overall efficiency and make markets more competitive,"⁹⁸ the Court, following *BMI* and *NCAA*, nevertheless held that a rule of reason approach was warranted in the case of wholesale purchasing cooperatives, because such cooperatives "must establish and enforce reasonable rules in order to function effectively."⁹⁹

As expected, subsequent lower court decisions have followed the Supreme Court's rule of reason approach to joint ventures. For example, in *Rothery Storage Van Co. v. Atlas Van Lines, Inc.*,¹⁰⁰ the D.C. Circuit, in a decision by Judge Bork, refused to apply a per se rule to an alleged "group boycott" of the plaintiff by the defendants, Atlas Van Lines and several affiliated carrier agents. Instead, the court, applying the rule of reason, found that the challenged restraints were ancillary to the joint venture, that they "preserved the efficiencies of the nationwide van line by eliminating the problem of the free ride, and accordingly, that Atlas' decision to terminate plaintiff's agency contract did not violate the Sherman Act."¹⁰¹ Similarly, in *Polk Brothers, Inc. v. Forest City Enterprises, Inc.*,¹⁰² the Court of Appeals for the Seventh Circuit reviewed a district court's decision that a noncompetition agreement was per se illegal. Finding that the covenant played an essential role in inducing the firms to cooperate, the Court held that the restraint was ancillary and that it therefore should be evaluated under the rule of reason.¹⁰³

These cases clearly indicate that joint ventures, and collateral restraints in joint venture agreements, will be evaluated under the rule of reason if they have the potential for creating new products, increasing efficiency or promoting competition in the market. Only where a joint

96. *Id.* at 101-02.

97. 472 U.S. 284 (1985).

98. *Id.* at 294.

99. *Id.* at 296.

100. 792 F.2d 210 (D.C. Cir. 1986), *cert. denied*, 479 U.S. 1033 (1987).

101. *Id.* at 229.

102. 776 F.2d 185 (7th Cir. 1985).

103. *Id.* at 188-91; see also *National Bancard Corp. v. VISA U.S.A.*, 779 F.2d 592 (11th Cir. 1986) (interchange fee charged by VISA not a naked restraint of competition and therefore not per se price fixing), *cert. denied*, 479 U.S. 923 (1986); *Berkey Photo, Inc. v. Eastman Kodak Co.*, 603 F.2d 263, 299-302 (2d Cir. 1979) (joint R&D between monopolist in one market and major firm in complementary market not a per se violation of section 1), *cert. denied*, 444 U.S. 1093 (1980).

venture is a sham "with no purpose except stifling competition,"¹⁰⁴ will it be subject to a per se rule.

B. Current Antitrust Enforcement Policy Is Hospitable Towards Legitimate PJVs

In their published guidelines and enforcement actions, the Antitrust Division of the Department of Justice and Federal Trade Commission have likewise adopted hospitable joint venture policies based on the rule of reason.

For Example, in its *Antitrust Enforcement Guidelines for International Operations*,¹⁰⁵ the Antitrust Division emphasizes that it will apply a rule of reason analysis in evaluating joint ventures that involve "some form of economic integration that goes beyond the mere coordination of the parties' decisions on price or output and that in general may generate procompetitive efficiencies."¹⁰⁶ The Guidelines state that the Division will first consider whether the joint venture is likely to have any anticompetitive effects in the market in which it operates¹⁰⁷ or in any "spillover markets."¹⁰⁸ The Division performs a similar rule-of-reason analysis for any vertical non-price restraints associated with the joint venture to determine if the restraints could facilitate collusion or exclude competitors.¹⁰⁹ If the Antitrust Division concludes that anticompetitive

104. Broadcast Music, Inc. v. Columbia Broadcasting Sys., 441 U.S. 1, 20 (1979) (quoting White Motor Co. v. United States, 372 U.S. 253, 263 (1963)).

105. International Guidelines, *supra* note 44, at 20,600.

106. *Id.* In a footnote, the Antitrust Division explains that, in determining whether to apply the rule of reason, it does not consider whether the "economic integration involved in the particular transaction actually would generate efficiencies. It is enough if the form of integration involved in general generates efficiencies." *Id.* at 20,594 n.47. However, if a purported joint venture involves no economic integration, but rather is "simply a device to restrict output or raise price," then it will not hesitate to challenge it. *Id.* at 20,600.

107. If, under its Merger Guidelines, the Division would not challenge the merger of the joint venture participants, then it will conclude that the joint venture and any associated restraints are unlikely to have any anticompetitive effects in the joint venture market. *Id.* Moreover, even if a merger of the participants would raise concern, the Division recognizes that a joint venture "may have a less restrictive effect on the independent decision-making of the joint venture participants with respect to output and price than would an outright merger," and accordingly may treat the joint venture more leniently. *Id.* at 20,601.

108. The Antitrust Division acknowledges that a "joint venture may . . . include operational or procedural safeguards that substantially eliminate any risk of anticompetitive spill-over effects," *id.*, and states that the presence of such safeguards may render an elaborate structural analysis of the spill-over market(s) unnecessary, *id.* at 20,602. In the absence of such safeguards, the Division will perform a market-power analysis using the same factors it uses in merger analysis. *Id.*

109. *Id.* The Division emphasizes that selectivity in choosing partners may be important to the success of a joint venture and that, accordingly, it will be concerned with the exclusion of rivals only if "(i) an excluded firm cannot compete in a related market or markets . . . without having access to the joint venture and (ii) there is no reasonable basis

effects are unlikely, it will not challenge the venture regardless of whether it generates any efficiencies. If, however, the Division concludes that significant anticompetitive effects are likely, it then considers whether "those anticompetitive effects are outweighed by the procompetitive efficiency benefits" generated by the joint venture.¹¹⁰

The enforcement actions of the Antitrust Division and Federal Trade Commission have been consistent with the 1988 Guidelines. In recent years, for example, the enforcement agencies have allowed firms with significant market shares to enter into PJVs where those ventures involved genuine economic integration.¹¹¹ More importantly, between 1984 and 1990, the Antitrust Division challenged only three PJVs, none of which involved joint R&D,¹¹² while the Federal Trade Commission challenged only four PJVs.¹¹³ In those cases, moreover, the challenged joint ventures involved marketing collaboration along with severe market concentration.¹¹⁴

Thus, existing policies of the antitrust enforcement agencies impose no unreasonable obstacle to the formation of PJVs involving genuine economic integration.

C. The Threat of Private Antitrust Suits Is Exaggerated

Proponents of the proposed legislation also argue that the threat of private suits may deter procompetitive PJVs. This concern appears exaggerated for two reasons.

First, the economic incentives for private plaintiffs to bring an antitrust challenge have decreased. The courts' use of the rule of reason in evaluating joint ventures and willingness to consider the potential benefits as well as possible costs of the venture have raised the expected costs of bringing an antitrust challenge against a joint venture and reduced the likelihood that a plaintiff will prevail in challenging a joint

related to the efficient operation of the joint venture for excluding other firms." *Id.* (footnote omitted).

110. *Id.* The Division notes, however, that it "will not recognize claimed efficiencies if it is clear that equivalent efficiencies can be achieved by means that involve no anticompetitive effect." *Id.*

111. See, e.g., General Motors Corp., 103 F.T.C. 374 (1984) (General Motors and Toyota, the world's first and third largest automobile manufacturers, allowed to enter into PJV, partially because it would permit the diffusion of existing production techniques and know-how from Toyota to G.M., despite producing an existing Toyota model rather than a new product).

112. 1990 Senate Hearing, *supra* note 4, at 39 (letter from Bruce C. Navarro, Acting Assistant Attorney General).

113. *Id.* at 229-30 (letter from Janet D. Steiger, Chairman, FTC).

114. See, e.g., United States v. Ivaco, 704 F. Supp. 1409 (W.D. Mich. 1989) (enjoined joint venture which would combine operations of two of three remaining producers of automatic tampers to create firm holding 70% of the relevant market). See generally H.R. REP. NO. 516, *supra* note 5, at 8; Brodley, *supra* note 6, at 101.

venture. This should reduce the number of plaintiffs willing to sue, especially where the suit is primarily intended to harass or extort a settlement from the defendants.¹¹⁵

Second, in recent years, the Supreme Court and lower federal courts have developed various procedural barriers which make it more difficult for plaintiffs to maintain private antitrust actions. For example, elaboration of the concepts of antitrust standing and antitrust injury has limited the number and types of parties permitted to bring antitrust challenges.¹¹⁶ Similarly, the Supreme Court's decision in *Illinois Brick Co. v. Illinois*¹¹⁷ made it much more difficult for indirect purchasers to bring a private antitrust action. Finally, the Georgetown antitrust project¹¹⁸ found that antitrust defendants frequently succeeded in bringing pretrial motions for summary judgment and motions to dismiss against private antitrust plaintiffs.¹¹⁹

At least in part as a result of these legal developments, the number of private antitrust actions filed in federal court has declined over the past 15 years. According to the Administrative Office of the United States

115. See, e.g., HOUSE COMM. ON THE JUDICIARY, 98TH CONG., 2D SESS., STUDY OF ANTITRUST TREBLE DAMAGE REMEDY 31 (Comm. Print 1984) (prepared by G. Garvey) (reduction in number of private suits filed in recent years may be due to new judicial commitment to rule of reason); Steven C. Salop & Lawrence J. White, *Economic Analysis of Private Antitrust Litigation*, 74 GEO. L.J. 1001, 1019 (1986) ("A plaintiff is more likely to sue when his perceived probability of success is greater, [and] when his litigation costs are lower . . .").

116. See, e.g., *Cargill, Inc. v. Monfort of Colo., Inc.*, 479 U.S. 104 (1986); *Brunswick Corp. v. Pueblo Bowl-O-Mat, Inc.*, 429 U.S. 477 (1977), *cert. denied*, 429 U.S. 1090 (1977).

117. 431 U.S. 720 (1977) (holding that, in general, indirect purchasers are barred from recovering for overcharges allegedly passed down to the plaintiff purchaser through a chain of distribution).

118. The Georgetown antitrust project collected data on all private antitrust actions filed between 1973 and 1983 in five federal districts. Of those, usable data was obtained on 2357 cases. See Steven C. Salop & Lawrence J. White, *Private Antitrust Litigation: An Introduction and Framework*, in *PRIVATE ANTITRUST LITIGATION: NEW EVIDENCE, NEW LEARNING* 3-4 (Lawrence J. White ed., 1988).

119. See Stephen Calkins, *Equilibrating Tendencies in the Antitrust System, with Special Attention to Summary Judgment and to Motions to Dismiss*, in *PRIVATE ANTITRUST LITIGATION*, *supra* note 118, at 185, 200, 207; Stephen Calkins, *Summary Judgment, Motions to Dismiss, and Other Examples of Equilibrating Tendencies in the Antitrust System*, 74 GEO. L.J. 1065, 1127 (1986) [hereinafter Calkins (1986)].

In addition, the success of such motions may well increase following the Supreme Court's decision in *Matsushita Electric Industrial Co. v. Zenith Radio Corp.*, 475 U.S. 574 (1986). See, e.g., *International Distrib. Ctrs., Inc. v. Walsh Trucking Co.*, 812 F.2d 786 (2d Cir.), *cert. denied*, 482 U.S. 915 (1987) (summary judgment for defendants in section 2 monopolization case); *In re Apollo Air Passenger Computer Reservation Sys.*, 720 F. Supp. 1068 (S.D.N.Y. 1989) (summary judgment for defendant in antitrust challenge to computer reservation system); *Florida Fuels v. Belcher Oil*, 717 F. Supp. 1528 (S.D. Fla. 1989) (summary judgment for defendants in section 2 essential facilities claim). See generally Calkins (1986), *supra*, at 1127; John T. Soma & Andrew P. McCallin, *Summary Judgment and Discovery Strategies in Antitrust and RICO Actions after Matsushita v. Zenith*, 36 ANTITRUST BULL. 325 (1991); ABA Production Joint Venture Report, *supra* note 63, at 34.

Courts, the number of private antitrust suits filed per year in the federal courts peaked at 1611 in 1977. This represented 1.2% of all civil cases filed that year. By 1980 the number of private actions filed had dropped to 1457, or 0.8%, of all civil cases filed. In 1990, only 452 private antitrust actions were filed, which represented only 0.2% of all civil actions filed in federal courts.¹²⁰ The Georgetown antitrust study further indicates that of the 2357 private antitrust suits studied, only 5.8% were challenges to mergers or joint ventures.¹²¹

The changes in the law and the sharp decrease in the number of private actions together strongly suggest that: (1) joint venture participants are unlikely to be sued by private antitrust plaintiffs, and (2) if these participants are sued, they will have considerably greater protection against frivolous or harassment-motivated claims than they had in the past through the use of pre-trial motions.

D. Current Antitrust Law Has Not Deterred the Formation of an Increasing Number of PJs

The increasing number of domestic and international joint ventures that have been formed in recent years further suggests that the antitrust laws pose no obstacle to legitimate joint ventures. Although data on recent joint venture activity in the United States is inadequate for a comprehensive conclusion, the empirical studies that have been conducted all agree that the number of PJs formed each year since the middle to late 1970s has been both significant and growing. For example, an informal survey of joint venture announcements in the *Wall Street Journal* by the Antitrust Division's Office of Economic Policy found 130 joint venture announcements during a two and one-half year period in the late 1980s.¹²² Another, more in-depth, survey of domestic joint ventures formed between 1960 and 1984 found that joint venture activity had "blossomed" since 1978, and that in some industries, the number of joint ventures formed in 1983 alone exceeded all previously announced joint ventures in that industry.¹²³ Moreover, the growth in joint venture

120. ANNUAL REPORT OF THE DIRECTOR OF THE ADMINISTRATIVE OFFICE OF THE UNITED STATES COURTS Table C-2 (1977), (1980), & (1990).

121. Salop & White, *supra* note 118, at 6.

122. 1989a House Hearing, *supra* note 4, at 45 (statement of James Rill, Assistant Attorney General, Antitrust Division).

123. HARRIGAN, *supra* note 63, at 7. See also Link & Tassey, *supra* note 55, at vii (joint ventures involving two or three firms increased from less than 200 per year in the 1970s to more than 400 per year by the mid-1980s); Mowery, *supra* note 46, at 3 (in recent years, the number of domestic and international collaborations involving U.S. firms has increased considerably).

activity has been especially rapid in certain high-technology industries, such as semiconductors.¹²⁴

During this same period, the number of international joint ventures involving U.S. firms also increased significantly. According to a study by Hladik of international joint ventures formed between 1974 and 1982 involving at least one American firm, the number of such ventures formed during the latter half of the period roughly doubled that of the first half.¹²⁵ This increased rate of growth appears to have continued beyond 1984, the termination date of the study.¹²⁶

Moreover, many of the joint ventures have involved large companies with substantial market shares and which are direct competitors. For example, joint ventures have been formed between General Motors and Toyota, General Motors and Chrysler, Merck and Johnson & Johnson, Dow and Eli Lilly, IBM and Microsoft, and IBM and Apple.¹²⁷

This evidence of widespread joint venture activity clearly suggests that the antitrust laws do not pose a significant obstacle to the formation of PJVs. Further support for this view lies in the lack of hard evidence that antitrust concerns deterred any planned joint ventures. For example, the Antitrust Section of the American Bar Association reported that it found only one instance in which domestic firms declined to pursue an integrative PJV for reasons that would be remedied by the proposed legislation.¹²⁸

V. ARGUMENTS FAVORING THE PROPOSED SPECIAL TREATMENT OF DOWNSTREAM CONSORTIA ARE FLAWED

Proponents of antitrust reform for PJVs have advanced several supporting arguments, most of which relate to special needs of high-

124. See, e.g., Katz & Ordover, *supra* note 16, at 170 (between January 1985 and July 1989, U.S. firms formed over 140 joint ventures in the semiconductor industry); Shapiro & Willig, *supra* note 47, at 117 (there has been a "decade-long trend in the distribution of joint venture formations" from energy, chemical, and metals industries "towards computer, electronic components, communications systems, pharmaceuticals, medical equipment, and financial services industries").

125. KAREN J. HLADIK, INTERNATIONAL JOINT VENTURES: AN ECONOMIC ANALYSIS OF U.S.-FOREIGN BUSINESS PARTNERSHIPS 39 (1985).

126. MOWERY & ROSENBERG, *supra* note 28, at 243 n.5.

127. See Brodley, *supra* note 6, at 101; Shapiro & Willig, *supra* note 47, at 117; Richard Brandt et al., *IBM and Microsoft: They're Still Talking, But . . .*, Bus. Wk., Oct. 1, 1990, at 164; Deidre A. Depke & Kathy Rebello, *IBM-Apple Could Be Fearsome*, Bus. Wk., Oct. 7, 1991, at 28.

128. ABA Production Joint Venture Report, *supra* note 63, at 20; cf. MOWERY & ROSENBERG, *supra* note 28, at 253 ("There is little evidence to support the argument that U.S. antitrust policy is a central factor in the decisions of American firms to collaborate with foreign [rather than with domestic] enterprises.").

technology industries. This Part reviews these arguments and shows that they either do not justify the broad proposed antitrust relief or that they are of questionable empirical importance or validity.

A. The High Costs and Risks of Commercializing and Producing Innovative Products Do Not Exceed the Capacity of Most Individual Firms

A frequently cited justification for encouraging PJVs is that the costs and risks of developing and manufacturing a new product have increased beyond the resources of many individual firms.¹²⁹ Citing such examples as dynamic random access memory chips (DRAMs), high-definition television (HDTV), and high-temperature superconductors,¹³⁰ proponents argue that not only have costs of R&D risen, but so too have the costs of plants for manufacturing any products of R&D.

Accepting the validity of these cost increases in certain high-technology industries, it does not follow that we should encourage industry-wide production consortia. First, to the extent that high basic or fundamental applied research costs deter firms from developing new products, such research could be conducted in a research consortia, such as Sematech or Microelectronics and Computer Technology Corporation, with the results then transmitted to the member companies. Extending cooperation into development and production is not a necessary requirement.

Second, with respect to the cost of plants and equipment necessary to produce new products, empirical studies suggest that, in the vast majority of industries, the minimum efficient scale of a plant is small relative to market demand.¹³¹ This suggests that production consortia are seldom necessary to achieve efficient-scale plants, and that PJVs involving two or three firms would solve the problem. This appears true even in the industries cited as requiring industry-wide production consortia. For example, thirteen Japanese companies manufacturing semiconductor memory chips in thirty separate plants¹³² rebuts the argument that a single industry-wide DRAM PJV is necessary. Similarly, with HDTV,

129. See, e.g., 1990 Senate Hearing, *supra* note 4, at 23 (statement of James F. Rill, Assistant Attorney General, Antitrust Division); 1989b House Hearing, *supra* note 4, at 55-56 (statement of Rep. Tom Campbell); H. R. REP. NO. 516, *supra* note 5, at 1; Jorde & Teece (1990), *supra* note 4, at 81.

130. See, e.g., 1989a House Hearing, *supra* note 4, at 12-13 (statement of Rep. Thomas Campbell); 1989b House Hearing, *supra* note 4, at 182-84 (statement of G. Moore, Chairman, Intel Corp.).

131. See *supra* text accompanying note 48.

132. 1990 Senate Hearing, *supra* note 4, at 78 (statement of Michael Porter).

three joint ventures involving U.S. firms currently are developing competing HDTV systems.¹³³

Thus, neither research costs nor plant economies justify further relaxation of the antitrust laws.

B. The Cyclical Nature of the Development Process and Shorter Product Lives Does Not Necessitate Cooperation in Both R&D and Production

Professors Jorde and Teece find support for cooperative production in what they call a "cyclical" view of the innovation process.¹³⁴ Rejecting the traditional view of innovation as a sequential process proceeding from basic research through applied research, product development, and finally to production, Jorde and Teece argue instead that product development involves "tight linkages and feedback mechanisms" between the various levels of activity and frequent "mid-course corrections to design, and redesign." This "cyclical" innovation process together with shorter product lives, Professors Jorde and Teece contend, necessitate close linkages between those performing the research and those actually developing the product.¹³⁵ Furthermore, since smaller firms may have to go outside to obtain necessary complementary assets, Professors Jorde and Teece conclude that, in order to achieve rapid commercialization of innovation, joint ventures must operate from the research level through at least the production level.¹³⁶

Although this argument may have some validity in the case of applied research joint ventures involving two or three firms, there is reason to doubt its validity as applied to industry-wide research consortia. First, in stressing the need for communication between manufacturing and research groups, Professors Jorde and Teece fail to distinguish among basic research, applied research, and developmental activities. While those developing a new product will have to, and do, talk to those who will manufacture the product in order to ensure that the product can be manufactured efficiently and inexpensively, there appears to be no similar need for those involved in basic or fundamental applied research to be in frequent communication with the plant floor. Thus, the

133. See Andrew Kupfer, *The U.S. Wins One in High-Tech TV*, FORTUNE, Apr. 8, 1993, at 60, 63; Lucy Reilly, *Making HDTV All-Digital Delays FCC Selection*, WASH. TECH., Apr. 9, 1992, at 6.

134. Jorde & Teece (1989a), *supra* note 4, at 14-15; see also MOWERY & ROSENBERG, *supra* note 28, at 8 (arguing that many primary sources of innovation are located downstream and operate independently of frontier scientific research); Jorde & Teece (1990), *supra* note 4, at 77 (referring to the "simultaneous model of innovation").

135. Jorde & Teece (1989a), *supra* note 4, at 15.

136. Jorde & Teece (1990), *supra* note 4, at 82-84.

need for communication with and "feedbacks" from those in production clearly varies with the kind of research or development activity.

Consortia, however, have generally shunned applied or "competitive" research or developmental activities where "feed backs" might be important,¹³⁷ because of the participants' fears that they may be forced to share strategic proprietary technology. To the extent that this competitive threat deters cooperation in "competitive" research, this casts doubt on the likelihood that firms in a consortium would agree to cooperate from basic research through production, at least in the absence of collateral restraints that would effectively limit competition among the participants. On the other hand, companies may be willing to cooperate in consortia if their primary purpose is to facilitate collusion.

Finally, even if the participants could be persuaded to cooperate in a research and production consortia, it is doubtful that this would speed the commercialization of any resulting innovations, because decision-making in joint ventures is slow and cumbersome, and becomes more so as the number of participating firms increases.¹³⁸ This suggests that any gains from combining the complementary skills and resources of the participants would be far outweighed by the more inefficient decision-making processes involved in consortia. Thus, participants interested in production consortia should be viewed with suspicion.

In summary, the cyclical nature argument lends little support to the argument for further relaxing the antitrust laws.

C. International Competition Will Not Eliminate the Danger of Collusion Among the Joint Venture Participants

Proponents of PJVs argue that such ventures will likely not have anticompetitive effects because of the presence of international competition.¹³⁹ However, a number of factors may limit the effectiveness of foreign competition in restraining domestic collusion.

In assessing *research* joint ventures, it is recognized that the relevant geographic market for research should usually be worldwide.¹⁴⁰ This results because information, the product of the research, is easily communicated. Thus, a competing foreign firm with commercially

137. See *supra* text accompanying note 78.

138. See *supra* text accompanying note 73.

139. See, e.g., Jorde & Teece (1989a), *supra* note 4, at 4; Jorde & Teece (1990), *supra* note 4, at 91; see also 1989a House Hearing, *supra* note 4, at 32 (statement of James F. Rill, Assistant Attorney General, Antitrust Division) ("[I]ncreasing globalization of markets, particularly those incorporating advanced technologies, has dramatically reduced the risk that cooperative production efforts among some competitors would result in higher prices to American consumers.").

140. See H.R. REP. NO. 1044, 98th Cong., 2d Sess. 10 (1984); International Guidelines, *supra* note 44, at 20,625; Ordover & Baumol, *supra* note 16, at 30.

valuable technology can either incorporate it in products which it exports to the United States or license the technology to U.S. firms that are not participants in the RJV.¹⁴¹ In either case, the foreign technology will place downward pressure on prices and thus act to prevent collusive behavior on the part of RJV participants. In addition, because of appropriability problems, research results may spill over to nonparticipants, further limiting the ability of participants to collude. Thus, the presence of foreign competition reduces the need for strict antitrust enforcement of RJs.

Foreign competition will not necessarily play such an effective policing role in the case of PJVs, however. First, exchange rate fluctuations could raise the price of imported goods, rendering them less competitive. If foreign firms attempt to remain competitive on price, they make themselves vulnerable to charges of dumping. Fear of sanctions restricts the ability of foreign firms to compete effectively and prevent monopoly profits by a U.S. consortia.

More importantly, foreign competitors may be subject to trade restraints, such as tariffs, quotas, or other quantitative restrictions. Such restraints can significantly limit the ability of foreign competitors to respond to and take advantage of anticompetitive behavior by domestic firms. Quotas and other quantitative restrictions, such as voluntary restraint agreements (VRAs) or orderly marketing agreements (OMAs) are especially pernicious because they prevent the foreign firms from increasing exports to meet demand should the joint venture participants attempt to restrict output and raise price.¹⁴² Further, the foreign importers have no incentive to oppose such restrictions, since the restrictions generate scarcity rents for them as well.¹⁴³ A final reason that these trade restraints present a competitive danger is that the participants in a PJV may seek such trade protection at any time after notification to the antitrust authorities.¹⁴⁴ Thus, even if there exists strong competition

141. Baxter, *supra* note 44, at 89.

142. See Diane P. Wood, *Commentary: Antitrust and International Competitiveness in the 1990s*, 58 ANTITRUST L.J. 591, 600-01 (1989); see also ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, COMPETITION AND TRADE POLICIES: THEIR INTERACTION 49 (1984) ("VERs . . . ultimately involve actions by exporting firms to limit their volume of exports, and in some instances, to raise prices. This, in turn, can promote collusive behavior among firms and weaken competitive forces in both markets.").

143. See, e.g., GARY C. HUFBAUER & HOWARD F. ROSEN, TRADE POLICY FOR TROUBLED INDUSTRIES 15 (1986); PAUL R. KRUGMAN & MAURICE OBSFELD, INTERNATIONAL ECONOMICS: THEORY AND POLICY 191 (1988).

144. The Antitrust Division recognizes that trade restrictions can limit the competitive significance of foreign competitors and considers the existence of such restrictions in performing market analysis. See, e.g., Merger Guidelines § 3.23, 49 Fed. Reg. 26,823 (Antitrust Div., Dep't of Justice 1984), reprinted in 4 Trade Reg. Rep. (CCH) ¶ 13,103 at 20,551, 20,561 (1984); International Guidelines, *supra* note 44, at 20,598. It can not, however, anticipate the effect of trade restraints that have yet to be imposed.

from foreign firms at the time of the PJV's formation, there is no assurance that such competition will remain effective.

D. Use of Joint Ventures in Japan and Europe Does Not Require that the U.S. Encourage Domestic Joint Ventures

Joint venture proponents also rely upon the alleged widespread use of joint ventures in Europe and, especially, in Japan to support their arguments.¹⁴⁵ They argue that, in order to remain competitive, the United States must follow the examples of Japan and Europe and encourage more PJVs. But these proponents fail to understand the nature of the joint ventures that have been created in those jurisdictions.

1. JAPAN

There is little, if any, cooperation in production or marketing among Japanese companies.¹⁴⁶ Rather, cooperation is generally limited to R&D.

Even in the area of cooperative research, cooperation is more limited than generally thought. For example, although the Japanese Fair Trade Commission found that fifty-five percent of 250 major Japanese firms surveyed participated in cooperative research, ninety percent of such cooperative agreements were private contractual arrangements between two companies, most of which were already affiliated.¹⁴⁷

In addition, a widespread misunderstanding exists concerning the industry-wide research consortia, sponsored by Japan's Ministry of Trade and Industry (MITI). In the 1970s, MITI began to use existing engineering research associations¹⁴⁸ to launch a few large cooperative projects in the

145. See, e.g., 1990 Senate Hearing, *supra* note 4, at 131-32 (statement of David J. Teece); 1989b House Hearings, *supra* note 4, at 278-79 (statement of J.D. Kuehler, President, IBM Corporation); Jorde & Teece (1989a), *supra* note 4, at 27-33, 55-61.

146. 1990 Senate Hearing, *supra* note 4, at 79 (statement of Michael Porter). However, Japanese companies are increasingly entering into PJVs with non-Japanese firms. Shapiro & Willig, *supra* note 47, at 122.

147. Heaton, *supra* note 72, at 34. Moreover, only six percent of these 250 companies belonged to technology research associations, and those that belonged performed only a small fraction of their research in the associations. *Id.* at 34-35.

148. The engineering research associations (ERAs), were created by MITI in the 1960s and copied after the British Research Associations. See David B. Audretsch, *Joint R&D and Industrial Policy in Japan*, in COOPERATIVE RESEARCH AND DEVELOPMENT, *supra* note 55, at 106-07 (discussing differences between Japanese and British research associations); see also JOHNSON, *supra* note 24 (discussing development and operation of British Research Associations). The ERAs' principal function was to coordinate various member research projects and facilitate the exchange and diffusion of technological information. Audretsch, *supra*, at 107; Heaton, *supra* note 72, at 33. Most of the ERAs were quite modest in scope, and, because they lacked physical premises, research was performed in the laboratories of the member companies. See MOWERY & ROSENBERG, *supra* note 28, at 223; Heaton, *supra* note 72, at 33.

electronics, computer and aircraft industries.¹⁴⁹ The primary purpose of these projects was to close the technological gap between Japanese firms and more advanced foreign firms by adapting, disseminating and using existing technological knowledge.¹⁵⁰ Given this goal, the projects focused on generic or pre-competitive research and avoided applied research and product development.¹⁵¹ Despite this focus, participants frequently attempted to free-ride by contributing second-rate personnel or equipment or by withholding proprietary technological know-how,¹⁵² and they continued to spend far greater sums on independent research efforts.¹⁵³ Evidence also exists that as Japanese firms have caught up technologically to advanced firms of other countries, they have become increasingly reluctant to participate in cooperative research projects.¹⁵⁴ Consequently, cooperative research will probably play a less significant role in the future.¹⁵⁵

In summary, the Japanese experience in cooperative research suggests that it can be an effective mechanism for diffusing technological knowledge among firms that lag in technological sophistication. However, it hardly provides a justification for industry-wide consortia in production.

149. MOWERY & ROSENBERG, *supra* note 28, at 223; Audretsch, *supra* note 148, at 110-11. Among the more famous of these projects were the Very Large Scale Integration (VLSI) project of 1976-1979, the project that attempted to develop a fourth-generation computer, *see, e.g.*, Audretsch, *supra* note 148, at 112; Merton J. Peck, *Joint R&D: The Case of Microelectronics and Computer Technology Corporation*, 15 RES. POL'Y 219, 222 (1986), the 3.75 Computer project of 1972-76, *see* Daniel I. Okimoto, *Regime Characteristics of Japanese Industrial Policy*, in JAPAN'S HIGH-TECHNOLOGY INDUSTRIES: LESSONS AND LIMITATIONS OF INDUSTRIAL POLICY 35, 54 (Hugh Patrick ed., 1987), and the Fifth Generation Computer project, *id.* at 52.

150. *Id.* at 54; MOWERY & ROSENBERG, *supra* note 28, at 223, 225; Heaton, *supra* note 72, at 33, 37. Professor Porter sees an additional purpose for these consortia. He views them as a "signaling device to indicate important areas for long-term research attention, and as a stimulus to proprietary company research." PORTER, *supra* note 46, at 398; *see also* Okimoto, *supra* note 149, at 52.

151. George C. Eads & Richard R. Nelson, *Japanese High Technology Policy: What Lessons for the United States?*, in JAPAN'S HIGH-TECHNOLOGY INDUSTRIES, *supra* note 149, at 254; Heaton, *supra* note 72, at 35-37; *see also* Okimoto, *supra* note 149, at 52 (Japanese consortia tend to focus on the "development of precommercial prototype models").

152. MOWERY & ROSENBERG, *supra* note 28, at 225; PORTER, *supra* note 46, at 398; *see also* Okimoto, *supra* note 149, at 54 ("for the first several years, mutual suspicion and concerns about the leakage of proprietary information impeded the free exchange of information" in the VLSI project).

153. *Id.* at 53.

154. *See* MOWERY & ROSENBERG, *supra* note 28, at 226.

155. *See* Heaton, *supra* note 72, at 37, 39.

2. EUROPE

Like Japan, the European Community (EC) has attempted to strengthen the international competitiveness of its high-technology industries by encouraging and subsidizing cooperative research. The majority of these subsidized cooperative research programs, however, have been limited to *precompetitive research*¹⁵⁶ and have not extended to joint production or marketing. Furthermore, with respect to private unsubsidized PJs, EC competition policy appears to present a greater obstacle to cooperation than do the U.S. antitrust laws.

The European Strategic Programme for Research and Development in Information Technology (ESPRIT), created in 1984, was the first major EC-subsidized collaborative research project. It focused on pre-competitive research in information technology and was intended to revitalize the EC's information technology and electronics industries.¹⁵⁷ Several other major, and minor, subsidized collaborative research projects followed, including RACE (Research in Advanced Communications for Europe), BRITE (Basic Research in Industrial Technologies for Europe), EURAM (European Research in Advanced Materials), and several biotechnology projects.¹⁵⁸ These various programs, which were brought together under the FRAMEWORK Programme, all focus on precompetitive research and share the same goals: raising the research capabilities of European firms in certain high-technology industries and encouraging transnational research cooperation among EC firms and between those firms and universities.¹⁵⁹

In 1985, another collaborative project, EUREKA (the European Research Cooperation Agency), was launched.¹⁶⁰ Like the FRAMEWORK Programme, EUREKA seeks to improve the competitiveness of members' high-technology firms and to foster intercompany cooperation across borders. Although EUREKA projects are not restricted to precompetitive

156. See NATIONAL SCIENCE BD., SCIENCE AND TECHNOLOGY INTEGRATION IN EUROPE AND INFLUENCES ON U.S.-EUROPEAN COOPERATION 1 (1990) [hereinafter NSB REPORT]; OFFICE OF TECHNOLOGY ASSESSMENT, U.S. CONGRESS, COMPETING ECONOMIES: AMERICA, EUROPE, AND THE PACIFIC RIM 29 (1991) [hereinafter OTA REPORT]; MARGARET SHARP & CLAIRE SHEARMAN, EUROPEAN TECHNOLOGICAL COLLABORATION 42-74 (1987); Roy Rothwell, *Public Innovation Policies: Some International Trends and Comparisons*, 12 PAPERS IN SCI. TECH. & PUB. POL'Y 13-14 (1986).

157. OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 47-53.

158. NSB REPORT, *supra* note 156, at 1; OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 56-62.

159. NSB REPORT, *supra* note 156, at 2; OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 63-65.

160. OTA REPORT, *supra* note 156, at 29. EUREKA is not an EC program, though all the EC member countries and the EC Commission itself are members. *Id.*

research, and can extend to more commercial research,¹⁶¹ they receive significantly less public funding.¹⁶²

Although the EC and its member states have contributed significant funds to these programs,¹⁶³ the programs have received some criticism.¹⁶⁴ More importantly, both subsidized programs are limited to cooperative research, and in the case of the FRAMEWORK Programme, limited to solely precompetitive research. Neither program permits cooperation to extend to coproduction or to joint distribution.

If EC firms wish to enter into private R&D ventures or to extend cooperation beyond R&D to production or marketing, they must deal with the EC competition laws. Complying with these laws often proves more burdensome and uncertain than complying with U.S. antitrust law.

In the EC, joint ventures must be primarily concerned with article 85 of the Treaty of Rome (the "Treaty"),¹⁶⁵ which basically prohibits agreements that restrict competition.¹⁶⁶ Under article 85(2), any agreement that

161. OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 71.

162. The U.S. International Trade Commission estimated that less than 10% of EUREKA's funding comes from public sources. U.S. INT'L TRADE COMM'N, 1992: THE EFFECTS OF GREATER ECONOMIC INTEGRATION WITHIN THE EUROPEAN COMMUNITY ON THE UNITED STATES: SECOND FOLLOWUP REPORT 16-10 (1990) [hereinafter ITC REPORT].

163. According to a 1990 estimate by the U.S. International Trade Commission, public funding of the FRAMEWORK program will amount to about ECU 5.7 billion (\$6.7 billion) from 1990-94. *Id.* at 16-6.

164. Numerous and varied criticisms have been leveled against these programs. First, some have criticized the way project participants are selected, suggesting both that small- and medium-sized firms do not have sufficient access, *see, e.g.*, Leigh Bruce, *EUREKA Has Found It*, INT'L MGMT., Dec. 1987, at 38, 41, and that projects are being forced to accept firms that merely seek a free ride. *See* OTA REPORT, *supra* note 156, at 29. A number of projects involving a large number of participants have also been criticized for experiencing management problems. DIRECTORATE GEN. XIII, TELECOMMUNICATIONS, INFO. INDUS. & INNOVATION, THE REVIEW OF ESPRIT, 1984-1988: THE REPORT OF THE ESPRIT REVIEW BOARD 9 (1989) [hereinafter ERB REPORT]. Others have criticized the procedures for selecting research projects. *See, e.g.*, SHARP & SHEARMAN, *supra* note 156, at 74. The research projects selected have also been criticized for being on the periphery of firms' true concerns, ERB REPORT, *supra*, at 8, and for producing little commercially useful technology. Guy de Jonquieres, *ESPRIT, JESSI Come Under Attack*, NEW TECH. WK., Nov. 5, 1990; *see also* ERB REPORT, *supra*, at 8. Finally, and most importantly, questions have been repeatedly raised whether the programs are improving the competitiveness of EC firms or making them more dependent on government subsidies. *See, e.g.*, OTA REPORT, *supra* note 156, at 29; Bruce, *supra*, at 41; Jonquieres, *supra*.

165. Article 86 of the Treaty may also prove applicable if the parties individually or collectively enjoy a dominant position before the formation of the venture. *See* 2 BARRY E. HAWK, UNITED STATES, COMMON MARKET AND INTERNATIONAL ANTITRUST: A COMPARATIVE GUIDE 308 & n.65 (2d ed. 1990); Frank L. Fine, *EEC Antitrust Aspects of Production Joint Ventures*, 26 INT'L LAW. 89, 90 (1992).

166. Article 85(1) provides:

1. The following shall be prohibited as incompatible with the common market: all agreements between undertakings; decisions by associations of undertakings; and concerted practices which may affect trade between Member States and which have as their objective or effect the prevention,

falls within article 85(1) is automatically deemed null and void. In addition, the Commission may impose substantial fines on parties to an agreement that violates article 85(1).¹⁶⁷

The possibility that the agreement will be nullified or fined creates a significant incentive for joint venture participants to notify the agreement to the Commission and to request a negative clearance¹⁶⁸ or a special exemption under article 85(3).¹⁶⁹ Because the Commission has found the majority of joint ventures to fall within article 85(1), at least where the participants are actual or potential competitors,¹⁷⁰ the joint venture

restriction, or distortion of competition within the common market, and in particular those which:

- (a) directly or indirectly fix purchase or selling prices or any other trading conditions;
- (b) limit or control production, markets, technical development, or investment;
- (c) share markets or sources of supply;
- (d) apply dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage;
- (e) make the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts.

TREATY ESTABLISHING THE EUROPEAN ECONOMIC COMMUNITY [EEC TREATY] art. 85(1).

167. Under Regulation 17, the Commission may fine the parties up to 1 million ECU or 10% of the parties' preceding year's turnover, whichever is greater. Council Regulation 17/62 of 21 February 1962, *Premier règlement d'application des articles 85 et 86 du traité*, art. 15(2), 1962 J.O. (13) 204. See generally 2 HAWK, *supra* note 165, at 20; Sara G. Zwart, *Innovate, Integrate, and Cooperate: Antitrust Changes and Challenges in the United States and the European Economic Community*, 1989 UTAH L. REV. 63, 80 & n.74.

168. See Council Regulation 17/62, *supra* note 167, art. 2. A negative clearance basically is a Commission determination that an agreement does not violate article 85(1). See Zwart, *supra* note 167, at 80 n.75.

169. Article 85(3) provides:

The provisions of paragraph 1 may, however, be declared inapplicable in the case of:

- any agreement or category of agreements between undertakings;
 - any decision or category of decisions by associations of undertakings;
 - any concerted practice or category of concerted practices;
- which contributes to improving the production or distribution of goods or to promoting technical or economic progress, while allowing consumers a fair share of the resulting benefit, and which does not:
- (a) impose on the undertakings concerned restrictions which are not indispensable to the attainment of these objectives;
 - (b) afford such undertakings the possibility of eliminating competition in respect of a substantial part of the products in question.

See Commission Notice Concerning Assessment of Cooperative Joint Ventures Pursuant to Article 85 of the EEC Treaty, 1993 O.J. (C 43) 2 (describing Commission's procedures for evaluating cooperative joint ventures under article 85) [hereinafter Notice Concerning Cooperative Joint Ventures].

170. See 2 HAWK, *supra* note 165, at 310, 314; Fine, *supra* note 165, at 93-94, 98-101.

participants must usually prove that the venture qualifies for a special exemption under article 85(3).

This notification procedure can impose significant burdens and create substantial uncertainty for joint venture participants, however. First, the Commission is not required to render a decision within any specified time limit and, in practice, frequently waits years after the notification is filed before granting or refusing an exemption request.¹⁷¹ Second, in conducting its investigation of a notification, the Commission will frequently demand significant amounts of highly sensitive information about the participants and the venture.¹⁷² Third, the Commission is required to grant an exemption for a specified period of time, and, in many cases, this period is shorter than the proposed duration of the venture.¹⁷³ Fourth, the Commission frequently conditions the grant of an exemption upon the parties' agreeing to modify or eliminate provisions deemed unnecessarily restrictive¹⁷⁴ or to comply with certain continuing reporting and operating obligations.¹⁷⁵ Finally, all notices applying for an exemption are published in the Official Journal of the European Communities with an invitation for third parties to make comments.¹⁷⁶

A quicker and easier approval process is available for joint ventures that qualify under various block exemptions or the recent merger control regulation. However, many PJVs will not qualify.

Under a 1984 block exemption for certain R&D joint ventures,¹⁷⁷ any RJV agreement that satisfies the conditions of the block exemption is

171. Fine, *supra* note 165, at 105; Zwart, *supra* note 167, at 85-86.

172. In this regard, the Commission holds broader discovery powers than do the U.S. antitrust authorities, including the power to make on-site inspections without prior notice to those involved. See 2 HAWK, *supra* note 165, at 26-28; Fine, *supra* note 165, at 105.

173. Fine, *supra* note 165, at 105-06.

174. HAWK, *supra* note 165, at 138-39.

175. The Commission may require the parties to make periodic reports on marketing and pricing policy, licensing information, or market share data. In some cases, the Commission also has required the parties to give the Commission advance notice of planned changes in the agreement. *Id.* at 325-26.

176. Fine, *supra* note 165, at 106.

177. Commission Regulation 418/85 of 19 December 1984 on the Application of Article 85(3) of the Treaty to Categories of Research and Development Agreements, 1985 O.J. (L 53) 5 [hereinafter R&D Regulation]. The block exemption was intended to encourage the formation of such ventures by reducing any antitrust uncertainty concerning their legality under article 85(1). Even before the issuance of the 1984 block exemption, however, the EC had adopted a favorable stance towards cooperative research. For example, Regulation 17, issued in 1962, provided that agreements that had as their sole objective joint research to improve techniques did not have to file a notification requesting an individual exemption. Council Regulation 17/62, *supra* note 167. Similarly, in its 1968 Notice on Cooperation, the Commission stated that agreements, whose sole purpose is the joint implementation of R&D projects, the placing of R&D contracts, or the sharing of R&D projects among participants, were not restrictions of competition within the meaning

deemed to fall outside of article 85(1). Unlike the NCRA, the block exemption permits joint exploitation of the results of the cooperative R&D through joint production or joint licensing. Nevertheless, it contains a number of restrictions that disqualify many joint ventures. For example, where the agreement provides only for joint R&D, the parties must be free to license or otherwise exploit the results of the joint R&D independently.¹⁷⁸ Second, where the agreement provides for joint exploitation, the exploitation must relate only to research results which are protected by intellectual property rights or constitute know-how which contributes substantially to technical or economic progress, and the results must be decisive for the manufacture of the contract products.¹⁷⁹ Third, where the agreement provides for joint production, the venture may only supply the products to the participants; it cannot also engage in joint distribution or marketing.¹⁸⁰ Finally, and most importantly, the Regulation imposes strict twenty percent market share limitations on R&D joint ventures that involve joint exploitation.¹⁸¹

The EC's 1989 Merger Control Regulation (MCR),¹⁸² on the other hand, requires that "concentrations"¹⁸³ that have a "Community dimension," including certain "concentrative" joint ventures, be notified

of Article 85(1), and hence did not need to be notified. Commission Notice, O.C. 75/3 (July 19, 1968). *See generally* Notice Concerning Cooperative Joint Ventures, *supra* note 169, at 9-10; 2 HAWK, *supra* note 165, at 341.

178. R&D Regulation, *supra* note 177, art. 2(c). This means, however, that the parties may compete away, through licensing or otherwise, any short-term rents resulting from the R&D.

179. *Id.* art. 2(d). Although this provision is clearly intended to exclude joint ventures that are primarily joint production and/or marketing ventures, it may create uncertainty for joint ventures for which joint R&D is only a component. HAWK, *supra* note 165, at 348.

180. R&D Regulation, *supra* note 177, art. 2(e).

181. Specifically, where the participants are competing manufacturers, the exemption will only apply if, at the time in which the agreement is entered, the combined market shares of the participants with respect to products "capable of being improved or replaced by the contract products" does not exceed 20% of the market for such products in the Common Market or a substantial part thereof. *Id.* art. 3(2). In addition, the 20% market share limitation must be satisfied for the duration of agreements involving joint exploitation. *Id.* art. 3(3). Finally, where the joint production involves components used by the participants in the manufacture of other products, the 20% limitation applies to the latter products for which the components represent a significant part. *Id.*

182. Corrigendum to Council Regulation 4064/89 of 21 December 1989 on the Control of Concentrations Between Undertakings, 1990 O.J. (L 257) 13 [hereinafter MCR]. The MCR became effective in September 1990.

183. Concentrations are deemed to exist where:

- (a) two or more previously independent undertakings merge, or
- (b) one or more persons controlling at least one undertaking, or one or more undertakings, acquire, whether by purchase of securities or assets, by contract or by any other means, direct or indirect control of the whole or parts of one or more other undertakings.

Id. art. 3(1).

to the Commission and cleared by it before they can be implemented. Although the MCR offers certain advantages to "concentrative" joint ventures,¹⁸⁴ it also creates some problems for these ventures. First, although the Commission has issued guidelines to assist in distinguishing "cooperative" joint ventures (still subject to article 85) from "concentrative" joint ventures (subject to the MCR),¹⁸⁵ there may remain considerable uncertainty as to whether a particular joint venture qualifies as a "concentrative" joint venture.¹⁸⁶ In addition, even where a joint venture qualifies as a "concentration" subject to the MCR, complying with the notification is likely to prove burdensome.¹⁸⁷

In summary, the EC, like Japan, has subsidized and encouraged cooperative research of a precompetitive nature as a means of helping European firms catch up with more advanced American and Japanese rivals. This subsidized collaboration, however, does not extend to joint production. As shown above, private PJs in the EC face more burdensome and time-consuming clearance procedures than their U.S. counterparts. Therefore, neither the Japanese nor European policies concerning antitrust and cooperative research provide a precedent for the proposed joint venture legislation.

VI. WEAKNESSES IN THE PROPOSED LEGISLATION

The most obvious weakness with the current bills is that proponents have failed to demonstrate any compelling need for the legislation. Although they claim that fear of antitrust liability deters procompetitive joint ventures, they have produced no clear and convincing evidence that this results. On the contrary, the available evidence suggests that the antitrust laws do not constitute a significant deterrent to the formation of joint ventures.¹⁸⁸ The current law quite clearly indicates that bona fide

184. The major advantages of the MCR are that: first, agreements notified under the MCR do not have to be separately notified under article 85, see, e.g., Fine, *supra* note 165, at 101; Barry E. Hawk, *The EEC Merger Regulation: The First Step Toward One-Stop Merger Control*, 59 ANTITRUST L.J. 195, 202 (1990) [hereinafter Hawk (1990)] and second, the Commission is required to render a decision within strict time limits, basically one month to four months. MCR, *supra* note 182, art. 10; see also Barry E. Hawk, *European Economic Community Merger Regulation*, 59 ANTITRUST L.J. 457, 459 (1991) [hereinafter Hawk (1991)]; Patrick Thieffry et al., *The Notification of Mergers Under the New EEC Merger Control Regulation*, 25 INT'L LAW. 615, 618-19 (1991).

185. Commission Notice Regarding the Concentrative and Cooperative Operations under Council Regulation (EEC) 4064/89 of 21 December 1989 on the Control of Concentrations between Undertakings, 1990 O.J. (C 203) 10.

186. See Fine, *supra* note 165, at 101-02; Hawk (1990), *supra* note 184, at 202-06; Hawk (1991), *supra* note 184, at 460-61; Thieffry et al., *supra* note 184, at 621-24.

187. Professor Hawk has described the MCR notification form as requiring "something like a Hart-Scott second request combined with a white paper." Hawk (1991), *supra* note 184, at 462; see also Thieffry et al., *supra* note 184, at 628-34.

188. See *supra* Section IV.D.

PJVs will be evaluated under the rule of reason and that efficiencies resulting from economic integration will be weighed against any restrictions on competition. In addition, there is no evidence of any alarming pattern of public or private antitrust litigation brought against joint ventures; rather, the available evidence suggests that the total number of public and private antitrust suits of all kinds has declined dramatically within the last ten years.

The best argument made by proponents of the legislation is that businessmen may be under a "misperception" that the legality of PJVs remains uncertain under antitrust laws. Not only is this an unusual justification—that the law should be changed because the business community misunderstands it¹⁸⁹—but there also appears little evidence to support it. As previously noted, the American Bar Association's Section of Antitrust Law was unable to confirm the existence of such a widespread misperception.¹⁹⁰ Moreover, the large and increasing number of U.S. firms that are participating in joint ventures, despite the difficulties in collective management, suggests that the antitrust laws hardly constitute a significant deterrent.¹⁹¹

A second major weakness of the legislation is that the pending bills do not appear targeted to achieve their ostensible goals. Legislative proponents claim that the bills, by encouraging cooperation in production, will help U.S. firms become more innovative and competitive. Unfortunately, neither the House nor Senate bills requires a PJV to engage in cooperative R&D as well as joint production. Nor do the bills require qualified joint ventures to produce new or innovative products.¹⁹² Finally, neither bill contains any requirement that the joint venture be involved in a high technology sector. Accordingly, there is little reason to expect the legislation will achieve its apparent purpose of encouraging the formation of innovative joint ventures in high-technology industries.

More importantly, the proposed legislation may well have significant adverse effects on the U.S. economy and on U.S. competitiveness. First, the proposed legislation may encourage greater collusion and cartelization by U.S. firms. As indicated above, coop-

189. See 1989b House Hearings, *supra* note 4, at 126-27 (statement of Edward Rock).

190. See *supra* note 128.

191. See *supra* Section IV.D.

192. The Senate bill apparently attempts to deal with this problem by requiring joint ventures that use existing facilities produce or process a "new process or technology." Unfortunately, the language is insufficient to solve the problem. First, there is no explanation of what constitutes a "new product or technology." Accordingly, it appears that simply introducing a slightly modified version of an existing product where no new technology is involved—such as a sterling silver garlic press—would satisfy the requirement. Second, where the joint venture builds a new facility, there is no extra requirement that the venture produce a new product or use a new technology.

eration at the production or marketing levels results in significantly more collusion than cooperation at the R&D level.¹⁹³ But the proposed legislation contains no provisions that effectively reduce this risk. For example, there are no limits on the allowed market shares, either individual or combined, of the participants. However, the anticompetitive risks of such joint ventures rise substantially as the combined market shares and market power of the participants increase. At the same time, it appears unlikely that economies of scale or other justifications for collaborative production warrant participation by firms holding a majority share of the relevant market. Accordingly, the bills should include some requirement that participants collectively holding market power demonstrate some need for such broad participation.

In addition, although the bills prohibit participants from engaging in joint marketing and from exchanging information concerning "costs, sales, profitability, prices . . . that is not reasonably required to carry out the purposes of the venture,"¹⁹⁴ they do not prevent participants from combining their competing production facilities. Thus, theoretically, an entire domestic industry could cooperate in production, jointly determine total output, and hence, indirectly agree on price.¹⁹⁵ Moreover, despite the statutory prohibition, it is not unlikely that production cooperation may lead to cooperation or collusion in pricing or other dimensions of competition.

Finally, even if production cooperation does not lead to explicit collusion, it may reduce rivalry among the participants. Again, this danger increases as the proportion of industry firms participating increases. The reduction in rivalry in turn could lead to higher prices and reduced innovation, making the participants less able to respond to foreign competition.¹⁹⁶

The danger of collusion and reduced rivalry suggests that increasing the number of PJVs will heighten the need for vigilant antitrust enforcement. Unfortunately, the bills weaken, if not eviscerate, antitrust enforcement for PJVs. First, although the bills require joint venture participants to file notifications with the antitrust authorities if they wish immunity from treble damages, the information required under the notification is not sufficiently detailed¹⁹⁷ to enable the antitrust authorities to perform an accurate evaluation of the competitive effects of the

193. See *supra* text accompanying notes 59-62.

194. See, e.g., S. REP. NO. 146, *supra* note 5, at 21, reprinted in 61 Antitrust & Trade Reg. Rep. at 353.

195. See *supra* note 58.

196. See, e.g., 1989b House Hearing, *supra* note 4, at 142 (written response of Edward Rock); PORTER, *supra* note 46, at 117, 169-70, 530.

197. See *supra* note 174.

venture.¹⁹⁸ In addition, since the proposed legislation would provide no new resources to the relevant regulators, there is a significant risk that they will be overwhelmed if the legislation results in a flood of filings.¹⁹⁹ Finally, the proposed legislation contains no provision for periodic monitoring by the antitrust enforcement agencies. Accordingly, the authorities may receive no notice should the joint venture subsequently engage in anticompetitive practices or seek trade protection to eliminate foreign competition.

The effect of the bill on private antitrust enforcement will be even greater. Although the issue of the appropriate measure of antitrust damages is beyond the scope of this article,²⁰⁰ it appears reasonable to infer that a reduction in the damage multiplier from three to one will reduce the incentive for private plaintiffs to bring suit.²⁰¹ Similarly, the possibility that a plaintiff will be ordered to pay the defendants' attorneys' fees may deter even meritorious suits by plaintiffs with less resources.²⁰² At the same time, by reducing the likelihood of private suits, these changes will increase the incentives for joint venture participants to engage in collusion or other anticompetitive behavior.²⁰³

The relevant question then becomes whether a reasonable justification exists for singling out PJVs for special treatment as opposed

198. The limited information required and the ministerial review provided under the 1984 NCRA was arguably sufficient for RJs, given their limited possible anticompetitive effects. This does not mean, however, that such information and review are sufficient where the competitive dangers are much more significant.

199. See 1989b House Hearing, *supra* note 4, at 131-32 (statement of Edward Rock); *id.* at 433 (statement of Arthur Kaplan).

200. The literature on antitrust damages stems largely from the writings of Gary Becker on the economic theory of deterrence. See Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169 (1968). Breit and Elzinga were among the first to apply Becker's approach to antitrust damages. See KENNETH G. ELZINGA & WILLIAM BREIT, *THE ANTITRUST PENALTIES: A STUDY IN LAW AND ECONOMICS* (1976). For more recent analyses, see, e.g., SECTION OF ANTITRUST LAW, A.B.A., MONOGRAPH NO. 13, *TREBLE-DAMAGES REMEDY* (1986); WARREN F. SCHWARTZ, *PRIVATE ENFORCEMENT OF THE ANTITRUST LAWS: AN ECONOMIC CRITIQUE* (1981); William Breit & Kenneth G. Elzinga, *Private Antitrust Enforcement: The New Learning*, 28 J.L. & ECON. 405 (1985); Frank H. Easterbrook, *Detrebling Antitrust Damages*, 28 J.L. & ECON. 445 (1985); William M. Landes, *Optimal Sanctions for Antitrust Violations*, 50 U. CHI. L. REV. 652 (1983); A. Mitchell Polinsky, *Detrebling versus Decoupling Antitrust Damages: Lessons from the Theory of Enforcement*, in *PRIVATE ANTITRUST LITIGATION*, *supra* note 118, at 7.

201. See, e.g., 1989b House Hearing, *supra* note 4, at 257 (statement of Joseph Alioto); *id.* at 457 (written response of Arthur Kaplan); Peter W. Rodino, *Let's Fix Only What's Broken: Some Thoughts on Proposed Reform of Private Antitrust Litigation*, in *PRIVATE ANTITRUST LITIGATION*, *supra* note 118, at 421.

202. See Shapiro & Willig, *supra* note 47, at 128.

203. See 1989b House Hearing, *supra* note 4, at 457 (written response of Arthur Kaplan). See generally Edward D. Cavanaugh, *Detrebling Antitrust Damages: An Idea Whose Time Has Come?*, 61 TUL. L. REV. 777, 786-87 (discussing incentive effects of treble damages on potential violators); Salop & White, *supra* note 115, at 1017-21.

to other antitrust concerns.²⁰⁴ In the case of RJs, the justification for the 1984 NCRA was that R&D activities suffered from certain unique market failures and that RJs could help correct those market failures without imposing any significant anticompetitive costs.²⁰⁵ No similar market failures afflict production activities, however. Moreover, the potential benefits offered by PJs appear considerably smaller, while the potential anticompetitive costs appear significantly larger. The reasoning underlying the special treatment of R&D joint ventures thus does not appear to extend to PJs.

A final weakness of the House bill involves the attempt to benefit U.S. companies and workers by requiring that plant facilities be located in the United States and by limiting the companies eligible to qualify for protection.²⁰⁶ These provisions appear both protectionist and misguided. Foreign firms clearly offer access to foreign markets and, to an increasing extent, possess technological information that would be valuable to U.S. companies.²⁰⁷ These protectionist provisions could pose a barrier to cooperation with technologically advanced foreign firms and thereby substantially undermine the legislative purpose.²⁰⁸ In addition, such

204. One such justification is suggested by Shapiro and Willig. They note that the economic theory of deterrence suggests that multiple damages are more appropriate for violations that are less likely to be detected. Given this, they argue that the reduction in damages under the Act may be a suitable quid pro quo for notification. Shapiro & Willig, *supra* note 47, at 126; *see also* Easterbrook, *supra* note 200, at 456-57.

This argument appears unpersuasive. The examples usually cited of antitrust violations that are likely to go undetected include such acts as price fixing or market division, where the actions are intentionally concealed. *Id.* at 456. In the case of the formation of a PJ, however, the fact of the agreement is not generally hidden, but rather publicized in the press or at least generally known by competitors in the industry. Considering the limited information required by the notification, it would appear to provide no information that is not publicly available or available to those acquainted with the industry.

205. *See supra* Section III.B. *But see* Easterbrook, *supra* note 200, at 456 (arguing that single damages for violations by RJs is too low).

206. The Senate Bill, as it emerged from the Senate Judiciary Committee, contained provisions similar to those in the House version. Prior to passage, however, the Senate amended these provisions to reduce their protectionist tone. Unfortunately, the substitute language is vague and ambiguous.

207. Empirical studies of joint ventures indicate that technological transfer and market access are the two main reasons a firm may enter into a joint venture with a foreign partner. *See, e.g.*, MOWERY & ROSENBERG, *supra* note 28, at 248; Mowery, *supra* note 46, at 13-15.

208. Mowery and Rosenberg come to a similar conclusion:

Restrictions or controls on international collaborative ventures involving U.S. firms do not appear to be an effective means to improve U.S. international competitiveness and in fact might impair competitiveness. The complexity of international collaborative ventures, the fact that the pattern and impact of these ventures vary considerably across industries, and the historical evidence that restrictions on technology transfer are either

restrictions may well result in other countries retaliating by restricting participation by U.S. companies in foreign joint ventures.²⁰⁹

VII. CONCLUSION

One of the great virtues of the U.S. antitrust laws is that they are sufficiently broad and general so that the courts have been able to adapt their language to changing market and technological conditions so as to ensure that U.S. markets remain competitive and efficient. Congress has generally recognized this virtue and has accordingly shown considerable reluctance to make special exceptions and exclusions.²¹⁰ Before creating a special exemption or immunity, Congress has generally required a "convincing prior showing of public interest or compelling economic need."²¹¹ Congress should demonstrate similar restraint here. It should also demand substantial and convincing empirical evidence that justifies the extension of the NCRA's special protection to PJVs before it enacts any of the proposed legislation. Such evidence has not yet been produced.

There is obviously considerable political appeal to passing legislation intended to spur U.S. competitiveness, especially where the legislation will not require substantial federal expenditures. In the case of the proposed PJV legislation, however, this political allure should be resisted. If passed, the legislation at best will have little effect on the economy; at worst, it will foster collusion, undermine U.S. competitiveness, and impose significant costs on U.S. consumers.

ineffective or perverse in their impacts . . . all argue against controls on collaborations involving non-defense technologies.

MOWERY & ROSENBERG, *supra* note 28, at 252; see also 1990 Senate Hearing, *supra* note 4, at 102 (statement of Joseph Brodley); 1989b House Hearing, *supra* note 4, at 140 (statement of George Heaton); *id.* at 261 (statement of Thomas Jorde).

209. See, e.g., 1990 Senate Hearing, *supra* note 4, at 102 (statement of Joseph Brodley); 1989 House Hearing, *supra* note 4, at 140 (written response of George Heaton); *id.* at 328 (statement of J.D. Huehler, President, IBM Corporation).

210. See 1989b House Hearing, *supra* note 4, at 95 (statement of George Heaton); Hamilton Fish, Jr., *Antitrust Relief and the House Judiciary Committee*, 35 ANTITRUST BULL. 219 (1990); Rodino, *supra* note 201, at 421-23.

211. Fish, *supra* note 210, at 222; see also NATIONAL COMM'N FOR THE REVIEW OF ANTITRUST LAWS & PROCEDURES, REPORT TO THE PRESIDENT AND THE ATTORNEY GENERAL 186 (1979) ("Each existing or proposed exemption should be justified in terms of empirically demonstrated characteristics of the specific industry that make competition unworkable. The defects in the market place necessary to justify an antitrust exemption must be substantial and clear.").

ARTICLE

SOFTWARE LITIGATION IN THE YEAR 2000: THE EFFECT OF OBJECT-ORIENTED DESIGN METHODOLOGIES ON TRADITIONAL SOFTWARE JURISPRUDENCE

DAVID M. BARKAN[†]

Table of Contents

I.	INTRODUCTION	315
II.	TRADITIONAL SOFTWARE DESIGN METHODS	317
III.	THE OBJECT-ORIENTED MODEL	320
	A. Overview	320
	B. The Basic Concepts of the Object-Oriented Model	321
	C. The Process of Designing Software Under the Object-Oriented Model	335
IV.	COPYRIGHT PROTECTION FOR OBJECT-ORIENTED SOFTWARE	341
	A. <i>Whelan</i> and Its Progeny	343
	B. The Filtering Approach	346
	C. Economic Balancing Approach	352
	D. Copyright Doctrine Properly Applied Provides Little Protection for Object-Oriented Software	354
V.	PATENT PROTECTION FOR OBJECT-ORIENTED SOFTWARE	358
	A. Patentable Subject Matter	358
	B. Non-Obviousness and the Relevant Prior Art	363
VI.	IMPLICATIONS FOR INNOVATION POLICY	365

I. INTRODUCTION

Ever since *Whelan Associates v. Jaslow Dental Laboratory*,¹ courts dealing with software have attempted to understand the process of

© 1993 David M. Barkan.

[†] J.D. 1992, School of Law (Boalt Hall), University of California, Berkeley; A.B. 1987, Harvard University. The author wishes to thank Peter Menell and Jeremy Barkan for their helpful comments and criticism.

1. 797 F.2d 1222 (3d Cir. 1986), cert. denied, 479 U.S. 1031 (1987).

software design. In doing so, they have focused almost exclusively on traditional methods of procedural programming and "top-down" design.² Moreover, most commentators on software protection present this traditional model of software design before articulating their proposed level of protection.³ While this model accurately reflects software design in the 1980s, the traditional approach is not well-suited to the exponential increase in size and complexity that will characterize software projects in the 1990s.⁴ As one critic of traditional design noted: "If builders built buildings the way programmers wrote programs, then the first woodpecker that came along would destroy civilization."⁵

In order to address the problem of complexity, programmers are likely to turn to object-oriented design and analysis because it allows programmers to adopt an entirely different approach toward problem solving and strongly encourages the development of libraries of reusable software "components." The next generation of software cases is likely to involve alleged infringement of complete programs designed according to the object-oriented model or alleged infringement of reusable software libraries.⁶ This article attempts to explain object-oriented design to judges

2. See *id.* at 1229-31 (describing the process as identifying the problem, outlining the solution, and then creating a flowchart, modules, and subroutines); Computer Assocs. Int'l v. Altai, Inc., 775 F. Supp. 544, 559 (E.D.N.Y. 1991) (citing expert testimony describing a computer program as "made up of sub-programs and sub-sub-programs, and so on"), *aff'd in relevant part*, 982 F.2d 693 (2d Cir. 1992); E.F. Johnson v. Uniden Corp. of Am., 623 F. Supp. 1485, 1501-02 n.17 (D. Minn. 1985); SAS Inst., Inc. v. S & H Computer Sys., 605 F. Supp. 816, 825 (M.D. Tenn. 1985) ("Beginning with a broad and general statement of the overall purpose of the program, the author must decide how to break the assigned task into smaller tasks, each of which must in turn be broken down into successively smaller and more detailed tasks."). Other cases have implicitly adopted this model in determining similarities between the plaintiff's and defendant's programs. See, e.g., Plains Cotton Co-op v. Goodpasture Computer Serv., 807 F.2d 1256, 1260 (5th Cir. 1987) (analyzing evidence of organizational copying), *cert. denied*, 484 U.S. 821 (1987). Pearl Systems v. Competition Electronics, 8 U.S.P.Q.2d (BNA) 1520, 1523-24 (S.D. Fla. 1988) (accepting expert testimony on similarities at the "subroutine" and "module" level); Q-Co Indus., Inc. v. Hoffman, 625 F. Supp. 608, 614-15 (S.D.N.Y. 1985) (comparing similarities in program "modules").

3. See 3 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT, § 13.03 [F] at 13-78.30 to .32 (1991); Peter S. Menell, *An Analysis of the Scope of Copyright Protection for Application Programs*, 41 STAN. L. REV. 1045, 1055-56 (1989); David Nimmer et al., *A Structured Approach to Analyzing the Substantial Similarity of Computer Software in Copyright Infringement Cases*, 20 ARIZ. ST. L.J. 625, 637-38 (1988); Gary L. Reback & David L. Hayes, *The Plains Truth: Program Structure, Input Formats and Other Functional Works*, COMPUTER LAW., Mar. 1987, at 1, 5.

4. For a general discussion of the inherent complexity of modern software see GRADY BOOCHE, OBJECT-ORIENTED DESIGN WITH APPLICATIONS 2-8 (1991) [hereinafter BOOCHE, OBJECT-ORIENTED DESIGN].

5. Booch, *Reuse of Software Components Could Reduce Costs*, GOV'T COMPUTER NEWS, Sept. 25, 1987, at 86.

6. The object-oriented model is most likely to arise first in cases involving graphical user interfaces. Apple Computer distributes an object library called MacApp which provides programmers with many tools necessary to implement the Macintosh user

and lawyers and to raise the problems that this new method of writing software will pose for courts applying traditional copyright and patent doctrines. The article presumes no technical background, but it does assume some familiarity with basic copyright and patent concepts. Part II of this article briefly reviews the traditional model of software development and notes some of the limitations that are encouraging the switch to object-oriented design. Part III presents the basic concepts in the object-oriented model and explains the process of designing software under this model. Part IV then evaluates the scope of copyright protection both for complete programs written using the object-oriented model and for reusable object libraries. While some existing case law and commentary could be used to support broad protection for such programs, this Section shows that pure copyright doctrine should not provide any protection for the aspects of the program that make it object-oriented. Part V analyzes patent protection for object-oriented software and concludes that sufficiently innovative programs and libraries could qualify for patent protection. Finally, Part VI considers the potential effects that copyright and patent protection would have on the growth of object-oriented technology.

II. TRADITIONAL SOFTWARE DESIGN METHODS

Traditional programming languages are based on the concept of a procedure, which allows programmers to write a small section of code which performs one small task.⁷ Well-written programs begin with a top-down approach to software design that has been described by Nimmer:

In practice, a programmer usually will start with a general description of the function that the program is to perform. Then, a specific outline of the approach to this problem is developed, usually by studying the needs of the end user. Next, the programmer begins to develop the outlines of the program itself, and the data structures and algorithms to be used. At this stage, flowcharts, pseudo-code, and other symbolic representations often are used to help the programmer organize the program's structure. The programmer will then break down the problem into modules or subroutines, each of which addresses a particular element of the overall programming

interface. Similarly, Symantec Corp. distributes a similar object-oriented library with its Pascal and C compilers. For a description of major software vendors working on object-oriented development, see Richard K. Aeh, *OOPS Are Picking Up Speed*, J. SYS. MGMT., Feb. 1991, at 19; Rick Whiting, *The Quest for a Better Way to Develop Software*, ELECTRONIC BUS., July 10, 1989, at 16. For a general discussion of the some of the problems that may slow widespread adoption of the object-oriented model, see Jeff Moad, *Cultural Barriers Slow Reusability*, DATAMATION, Nov. 15, 1989, at 87.

7. See generally ELLIOT B. KOFFMAN, PROBLEM SOLVING AND STRUCTURED PROGRAMMING IN PASCAL 52-59 (1985). The terms "procedure," "module," and "subroutine" are used interchangeably to denote a small number of programming instructions which perform a single task.

problem, and which itself may be broken down into further modules and subroutines. Finally, the programmer writes specific source code to perform the function of each module or subroutine, as well as to coordinate the interaction between modules or subroutines.⁸

This process has also been described as "functional decomposition," since the "primary question addressed by the systems analysis and design is WHAT does the system do [or] what is its *function*?"⁹ The complex function identified at this stage must be further decomposed into smaller functions, a process which is repeated until the problem can be "expressed as some combination of many small, solvable problems."¹⁰

Several important implications flow from this model that affect the way programmers are taught to approach problem solving. First, the traditional model forces programmers to focus on specific tasks that must be achieved, resulting in strong analysis of the procedures and functions that must be used, but little emphasis on data structures.¹¹ Data structures are usually conceived only after the procedures have been generally defined.¹² While diligent programmers may go back and rethink some of their procedures after analyzing their data structures, as a general rule, "[i]n a typical procedural programming language such as C or Pascal, programmers approach data and algorithms as separate entities."¹³ Second, data to be used by several procedures are usually

8. NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.31 to .32.

9. Brian Henderson-Sellers & Julian M. Edwards, *The Object-oriented Systems Life Cycle*, COMM. ACM, Sept. 1990, at 142, 145 (emphasis in original).

10. Menell, *supra* note 3, at 1055 (citation omitted).

11. See Tim Korson & John D. McGregor, *Understanding Object-Oriented: A Unifying Paradigm*, COMM. ACM, Sept. 1990, at 40, 46.

The term "data structure" denotes the symbolic representation of a particular area of memory where specific data will be stored. For example, if we wished to store data about this article, we could create a data structure in Pascal called a "record" that includes various fields for storing different pieces of information. The record could be defined as follows:

```
TheArticle = record
    Author: str;
    NumberOfPages: int;
    Issue: int;
    StartingPage: int;
end;
```

In this data structure, the name of the entire record is "TheArticle." The name of the author is stored in the field "Author," the length of the article is stored in the "NumberOfPages" field, the issue number is stored in the "Issue" field, and the starting page in the issue is stored in the "StartingPage" field.

12. An obvious exception would be a database program since the programmer is usually given highly specific information on what type of information must be stored in the database, thus leading to an initial focus on data structures.

13. Randy Leonard, *OOP: The Future for Macintosh Development*, MACTECH Q., Spring 1989, at 22.

defined in one place and can be accessed by any module or subroutine.¹⁴ While this approach may seem logical and efficient, it creates severe problems as soon as a complex program needs to be updated or revised since it "leads to the stack of dominoes effect familiar to anyone working in program maintenance whereby changes to one part of a software system often cause a problem in an apparently dissociated program area."¹⁵ As a result, each time a data structure needs to be changed or refined, one would have to identify all the procedures that rely on the old definition of the data structure and change them accordingly. Finally, since the top-down approach forces programmers to think in terms of a series of single functions, programmers are less likely to incorporate evolutionary changes in the data structures into the big picture of the overall system.¹⁶

In general, the traditional model provides few easy ways to reuse existing pieces of software, thus making software development less efficient than other engineering disciplines that are accustomed to reusing existing components.¹⁷ The close dependence between procedures and the specific definition of data structures makes it extremely difficult to reuse selected procedures in a different project. While programmers certainly copy solutions to certain problems from earlier projects and from public domain sources, rarely can programmers lift an existing procedure from another project and incorporate it into their program without serious modification.

The traditional model leads to several additional problems from a project management perspective. While the top-down model has worked well until recently, software projects in the 1990s will involve new levels of complexity and will require better systems for managing multiple programmers working on different parts of one system. As software will increasingly be required to model real world behavior in meaningful

14. Ray Duncan, *Redefining the Programming Paradigm*, PC MAG., Nov. 13, 1990, at 526 ("[Y]ou visualize the data to be worked on as sitting in one place, while various routines call upon each other to do things to the data."); Henderson-Sellers & Edwards, *supra* note 9, at 146; Chris Terry, *Objects Facilitate Modular, Reusable Code*, EDN, Nov. 9, 1989, at 85.

15. Henderson-Sellers & Edwards, *supra* note 9, at 146.

16. *Id.*

17. Peter Coffee, *Honing the Software Equivalent of the Transistor*, PC Wk., Sept. 25, 1989, at 38 (comparing object-oriented development to electrical engineering which emphasizes the reuse of standard building block components); Chris Terry, *Reusable Software Requires Building Blocks*, EDN, Jan. 3, 1991, at 59 ("Power-supply designers don't have to manufacture their own capacitors and resistors, let alone their connectors, nuts and bolts. They use parts that conform to universal (or at least widely accepted) standards. Yet software-systems designers have to write code for almost every function in their system except the services the operating system provides.").

ways, the human capacity for managing complexity will be severely tested.¹⁸

While traditional methods of functional decomposition provide one way of managing complexity, they are constrained in ways that make it difficult to manage projects that require large numbers of programmers. Since changes made by one programmer can effect other disparate parts of the program in a ripple fashion, traditional design eventually breaks down in extremely complex projects. Of course, the traditional model can be improved by forcing each programmer to adhere to detailed specifications that dictate how each programmer's portion interacts with the program as a whole.¹⁹ While this modification of the traditional method was sufficient for "programming-in-the-large," it may not be sufficient for the "programming-in-the-colossal" that will be required in the 1990s.²⁰

III. THE OBJECT-ORIENTED MODEL

A. Overview

While object-oriented principles were originally developed in the 1960s,²¹ the current object-oriented model is an attempt to address the modern problem of "programming-in-the-colossal." Generally speaking, the object-oriented model emphasizes use of small, discrete components which can be used without any knowledge of how they work internally, thus breaking the tight dependency between data structures and procedures that constrained the traditional model. Several positive benefits flow directly from this one assumption:

Object-oriented decomposition yields smaller systems through the reuse of common mechanisms, thus providing an important economy of expression. Object-oriented systems are also more

18. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 14 ("As we first begin to analyze a complex software system, we find many parts that must interact in a multitude of intricate ways, with little perceptible commonality among either the parts or their interactions: this is an example of disorganized complexity. As we work to bring organization to this complexity through the process of design, we must think about many things at once. For example, in an air traffic control system, we must deal with the state of many different aircraft at once, involving such properties as their location, speed, and heading. . . . Unfortunately, it is absolutely impossible for a single person to keep track of all of these details at once.").

19. *Id.* at 29-30 (discussing the development of modular programming techniques in late third-generation programming languages).

20. *Id.* at 32.

21. For a good history of the development of object-oriented languages, see Duncan, *supra* note 14 (noting that object-oriented programming languages were developed first in 1967 and then implemented in the 1970s at Xerox's Palo Alto Research Center, but were not practical for true commercial programming projects until the development of fast and efficient C++ compilers).

resilient to change and thus better able to evolve over time, because their design is based upon stable intermediate forms. Indeed, object-oriented decomposition greatly reduces the risk of building complex software systems, because they are designed to evolve incrementally from smaller systems in which we already have confidence.²²

Object-oriented programming is better suited for complex projects, because it more naturally models complex behavior in the real world. Through the concept of *inheritance*,²³ object-oriented programming focuses on hierarchical relationships between different components in a particular real-world system. For example, rather than considering a rectangle and a square to be two separate objects, the object-oriented model views the square as a particular kind of rectangle with special properties. This view is far more than a simple conceptual device. On a practical level, a programmer who wishes to simulate the behavior and properties of a square does not have to start from scratch; rather, the programmer starts with a model of the rectangle which may already exist in a library and then simply adds those properties that are unique to the square. Moreover, this model closely fits the way humans approach physical systems in the real world:

For example, with just a few minutes of orientation, an experienced pilot can step into a multi-engine jet aircraft he or she has never flown before, and safely fly the vehicle. Having recognized the properties common to all such aircraft, such as the functioning of the rudder, ailerons, and throttle, the pilot primarily needs to learn what properties are unique to that particular aircraft. If the pilot already knows how to fly a given aircraft, it is far easier to know how to fly a similar one.²⁴

As a result, the programmer can approach a new software project as an incremental learning process that closely models real behavior. A programmer seeking to simulate the behavior of an F-15 fighter jet would first start with programs that simulate the behavior of a standard commercial jet. If those earlier programs were written according to object-oriented principles, the programmer can simply start with the old program and then incrementally add those behaviors and processes that make the F-15 different from the commercial jet.

B. The Basic Concepts of the Object-Oriented Model

Before further exploring this approach to software design, it is necessary to define some basic concepts that are essential to the object-oriented model. Rather than present these concepts in a vacuum, it is useful to illustrate them in the context of an actual software project.

22. BOOCHE, OBJECT-ORIENTED DESIGN, *supra* note 4, at 16.

23. See *infra* Section III.B.2.

24. BOOCHE, OBJECT-ORIENTED DESIGN, *supra* note 4, at 12.

Assume that you are given the following software project which we will call "QuadWorld."²⁵ You are told that QuadWorld must allow the user to draw and manipulate various types of quadrilaterals,²⁶ such as squares, rectangles, parallelograms, and rhombi. Specifically, the user must be able to choose a particular shape, draw the shape, rotate the shape, move the shape in any direction and erase the shape. The user should also be able to have the program calculate the area of any selected shape. We will use these requirements to illustrate the basic principles of the object-oriented model.

1. OBJECTS, ENCAPSULATION, AND MESSAGES

An *object* is the basic programming unit in the model and essentially combines the traditional concepts of data structures and procedures into a single entity. For example, an object designed to represent a rectangle would include two pieces of data: length and width. It would also contain a complete set of procedures to draw, rotate, move, and otherwise manipulate the rectangle. In this way, the object captures both the state (data regarding length and width) of the rectangle as well as its behavior (the set of procedures for manipulating the rectangle).²⁷ Once the object is defined, any other part of QuadWorld can use that rectangle by simply sending it a *message*, which is a command telling the object to perform one of its defined behaviors. The definition of such a rectangle object might look like this:²⁸

25. This example is taken directly from KURT J. SCHMUCKER, OBJECT-ORIENTED PROGRAMMING FOR THE MACINTOSH 32-35 (1986). Where figures or drawings are adapted directly from this text, they will be appropriately cited.

26. A quadrilateral is any geometric figure with four sides.

27. For a more formal definition of an object, see Korson & McGregor, *supra* note 11, at 42:

Objects are the basic run-time entities in an object-oriented system. Objects take up space in memory and have an associated address like a record in Pascal or a structure in C.

The arrangement of bits in an object's allocated memory space determines that object's state at any given moment. Associated with every object is a set of procedures and functions that define the meaningful operations on that object. Thus, an object encapsulates both state and behavior.

28. Each object-oriented programming language uses slightly different terms to define the basic concepts of object-oriented programming. The definition given here is not meant to mimic any particular language but rather to provide an easily understandable illustration of the basic concepts.

Object Definition of a Rectangle:

Internal Data:

length

width

current position of the top left corner of the rectangle on the screen

Messages that the object is able to perform:

Create, Draw, Move, Stretch, Rotate, Calculate Area

Internal implementation of those messages:

Calculate Area:

Area = length X width

Create: Code for implementing the create message

Draw: Code for implementing the draw message

Move: Code for implementing the move message

(Code for remaining messages as above)

Several observations must be drawn from this definition. First, note that the data fields are called "*internal*" data. In the object-oriented model, data structures are completely private to the object and cannot be used directly by any other part of the program. Similarly, the internal implementation of each message, such as the actual source code that would draw the rectangle, is also private to the object. When another part of the program wishes to draw a rectangle, it simply sends the "Draw" message to the object. Other parts of the program do not care how the object implements that message or what data structures the object uses to perform the draw message. In essence, the other part of the program tells the object "draw yourself, and I don't care how you do it."

This process of "hiding" the internal data structures and the implementation of messages within the object illustrates the principle of *encapsulation*. Since "no part of a complex system should depend on the internal details of any other part,"²⁹ encapsulation is critical to successful "programming-in-the-colossal." Since the object's internal structure is private to the object, changes can be made to the internal structure without having any effect on the rest of the program. For example, the definition of the object could be changed as follows (changes are shown in *italics*):

29. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 45. Booch defines encapsulation as the "process of hiding all of the details of an object that do not contribute to its essential characteristics." *Id.* at 46.

Alternative Object Definition of a Rectangle:

Internal Data:

Position of the top, left corner

Position of the bottom, right corner

(Note that the length and width pieces of data have been deleted)

Messages that the object is able to perform:

Create, Draw, Move, Stretch, Rotate, Calculate Area

Internal implementation of those messages:

Calculate Area:

Area = (first coordinate of bottom, right corner – first coordinate of the top left corner) X (second coordinate of the top left corner – second coordinate of the bottom right corner)

Create: Code for implementing the create message

Draw: Code for implementing the draw message

Move: Code for implementing the move message

(Code for remaining messages as above)

The beauty of the object-oriented model is that these internal changes can be made with absolutely no effect on any other part of the program. Unlike traditional programming, where changes in one procedure ripple through many other parts of the program, the internal definitions of objects can be changed many times with only minimal effect on the overall program. Moreover, tasks can be divided among numerous programmers without requiring the extensive coordination that is necessary when writing code with traditional methods. Each programmer need only be told the desired characteristics of a particular object; how the programmer chooses to write the internal code for that object doesn't matter to the other programmers working on the project. In fact, an object with the desired characteristics may already exist in a company-owned library of objects. As a result, many tasks can be completed without requiring any programmer to write new code.

In fairness to the traditional model, good programming practice can produce some of this modularity within procedural techniques. However, such modularity was usually achieved in a sporadic, informal, and inconsistent fashion.³⁰ In contrast, the object-oriented model requires

30. While late third-generation programming languages allow for separately compiled modules, this concept was used primarily to allow several programmers to work on the same project rather than as an independent tool for abstraction:

Modules were rarely recognized as an important abstraction mechanism; in practice they were used simply to group logically related subprograms. Most languages of this generation, while supporting some sort of modular structure, had few rules that required semantic consistency among module interfaces . . . Unfortunately, because most of these languages had dismal

programmers to incorporate the concept of encapsulation into their basic approach to building software.³¹

2. CLASSES AND INHERITANCE

The concept of a *class* extends the object-oriented model by providing a way to describe a group of objects that have similar properties and behavior. For example, if more than one rectangle is needed at a time, we need not define each rectangle individually. Instead, we define a *class* of rectangles, which serves as a template for creating rectangle objects. Each rectangle object is called an *instance* of the rectangle class. A class definition includes the data that describe each object, called *instance variables*, the messages that the classes must accept,³² and the code that implements each message, known as *methods*.³³ A class can also be thought of as a "factory," which is like a

cookie cutter that stamps out new instances of its class, new objects, whenever necessary. All instances of a given class have the same structure although the actual data stored in any one object may be different from the data stored in another object of the same type, just as all cookies stamped out with the same cookie cutter have the same shape even if they are made of different types of dough and decorated differently.³⁴

In other words, if the user wants to draw two different rectangles of different sizes, our program will use the class definition for the rectangle to create two rectangle objects which possess the same types of behavior but have different values stored in their width and length variables.

The concept of a class does not become truly powerful, however, until combined with the concept of *inheritance*.³⁵ Inheritance is the primary organizing principle behind object-oriented design and the major reason why this model naturally follows the hierarchical structure of real-

support for data abstraction and strong typing, such [semantic] errors could be detected only during execution of the program.

Id. at 30.

31. See the discussion of the object-oriented design process *infra* Section III.C.

32. The complete set of messages accepted by a particular class is sometimes called the class's protocol. SCHMUCKER, *supra* note 25, at 17.

33. Again, the precise terminology varies somewhat depending on the particular object-oriented programming language. Instance variables, messages, and methods are the terms used by Object Pascal and Symantec Corporation's object extensions to the C programming language. See, e.g., SCHMUCKER, *supra* note 25, at 17; SYMANTEC CORP., THINK C OBJECT-ORIENTED PROGRAMMING MANUAL 20 (1991) [hereinafter THINK C OBJECT-ORIENTED PROGRAMMING MANUAL]. The programming language C++ instead uses the terms data member, message, and member function, respectively. See generally BJARNE STROUSTRUP, THE C++ PROGRAMMING LANGUAGE (1st ed. 1986).

34. SCHMUCKER, *supra* note 25, at 18.

35. Inheritance also distinguishes the class concept from the idea of user-defined types that is found in Pascal and C. *Id.* at 21.

world entities.³⁶ Given any particular class, we can define a *sub-class*³⁷ or *immediate descendant* which automatically inherits all of the behavior and properties of the original class, (now called the *super-class*³⁸ or *immediate ancestor*); we then take this sub-class and add any additional behavior (through new messages and methods) and any additional properties (through new instance variables) that make this sub-class different from its super-class. For example, once we have defined the "rectangle" class, we might realize that a square is simply a special type of rectangle. We could then simply define the "square" class to be a sub-class of the "rectangle" class. The "square" class would inherit all the instance variables, messages, and methods of the "rectangle" class. The "square" class would already "know" how to respond to the messages for draw, move, and rotate. At this point, we would also add the features that make a square different from a rectangle.

Object-oriented languages provide two different ways to differentiate the behavior of a sub-class from its super-class. First, we could simply add a new message and a corresponding method to the definition of the sub-class. For example, we might add a message to our definition of the "square" sub-class that automatically draws the largest circle that just fits inside the square and touches each side exactly once. The super-class rectangle could not have had this message because it's geometrically impossible for a circle within a rectangle to touch each side exactly once. But, the square is a special kind of rectangle,³⁹ and this message could be performed on a square.

In addition, object-oriented languages allow a sub-class to *override* methods inherited from the super-class. For example, we would want our square to respond to the message "calculate area" just as we wanted our rectangle to respond to that message. Normally, the sub-class square inherits both the message and the method which contains the actual programming code that tells the object how to respond to that message. In the case of the rectangle, the method for "calculate area" multiplied the rectangle's length by its width. However, we know that the square is a special kind of rectangle whose length is equal to its width. Thus, we

36. Inheritance is also a more powerful tool for achieving meaningful abstractions: Inheritance is the most promising concept we have to help us realize the goal of constructing software systems from reusable parts, rather than hand coding every system from scratch. Procedural abstraction has worked well for some select domains, such as mathematical libraries, but the unit of abstraction is too small, the procedural focus not general enough, and the parameter mechanism too rigid.

Korson, *supra* note 11, at 43.

37. C++ uses the term "derived class." STROUSTRUP, *supra* note 33, at 30.

38. C++ uses the term "base class." *Id.*

39. The "is a special kind of" terminology is a particularly useful way to think of the relationship between sub-classes and super-classes.

might want to override the method for "calculate area" with a new method which calculates the area by simply squaring the value of any side.⁴⁰ When the program sends a "calculate area" message to an object of the "square" class, the object will calculate the area by squaring one side rather than using the method of "length times width" that it initially inherited. The ability to override methods provides an important tool for customizing the behavior of sub-classes and for taking advantage of efficiencies that might be available in sub-classes that cannot be used in the more generic super-classes.

This inheritance concept can be easily applied to the structure of the QuadWorld application. At the highest level, we must define a *root class* which is a class that has no super-classes above it. In this case, we might define a root class, called "quadrilateral," which defines only the most generic properties and behaviors common to all four-sided geometric figures. After that, we could construct the simple inheritance chart that is shown in Figure 1.⁴¹

40. Of course, this example is fairly trivial and would not provide any efficiency improvement, but it nonetheless shows how sub-classes can customize particular inherited behavior. A more useful example would arise when we originally define the rectangle class. While the rectangle class would inherit the methods of the parallelogram class, the formula for calculating the area of a parallelogram is considerably more involved than the simple "length X width" formula that can be used for the rectangle class.

41. This figure is adapted directly from SCHMUCKER, *supra* note 25, at 22.

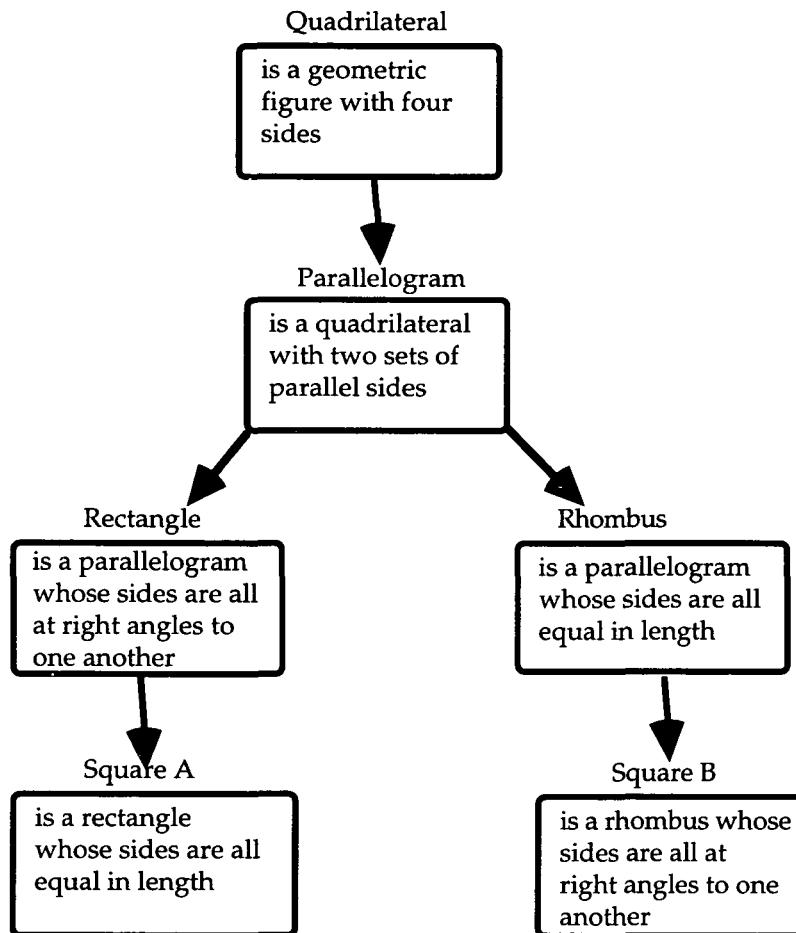


Figure 1: Simple Inheritance Chart for Quad World

A simple inheritance chart thus provides a logical method for organizing the basic relationships among various objects. Moreover, it forces us to recognize common features among different objects and then allows us to take pre-existing objects and modify them to fit our needs by building sub-classes from them. Moreover, "[I]nheritance not only supports reuse across systems, but it directly facilitates extensibility within a given system . . . [I]nheritance minimizes the amount of new code needed when adding additional features . . . and minimizes the amount of new code that must be changed when extending a system."⁴²

42. Korson & McGregor, *supra* note 11, at 43.

Real world behavior does not always fit this simple hierarchical structure, however. The object-oriented model can account for more complex relationships between classes with the concept of *multiple inheritance*.⁴³ Looking at Figure 1, we notice that a square has properties that make it a special kind of rectangle and properties that simultaneously make it a special kind of rhombus. Rather than having two classes of squares as in figure 1, we can improve our model by having a single class, called "square" that inherits from both the rectangle and the rhombus classes. The "square" class will inherit the ability to build a parallelogram with right angles from the rectangle class and will simultaneously inherit the ability to build a parallelogram with equal sides from the rhombus class.⁴⁴ As a result, we can create the "square" class without having to write any new methods, since all of the square's behavior is based on the combined properties of the "rectangle" and "rhombus" classes.⁴⁵ The improved model for QuadWorld is shown in Figure 2.⁴⁶

43. Since multiple inheritance is more difficult for the compiler to handle than simple inheritance, not all object-oriented development packages support multiple inheritance. For example, Symantec's Think C does not support it, while a full implementation of C++ would support it. See THINK C OBJECT-ORIENTED PROGRAMMING MANUAL, *supra* note 33, at 62.

44. SCHMUCKER, *supra* note 25, at 277.

45. If any messages are defined in both the rectangle class and the rhombus, then we must have some way to specify whether we want the square to inherit the rectangle's method for that message or the rhombus' methods. Multiple inheritance allows us to specify either the rectangle or the rhombus as the *primary immediate ancestor class*. In any inheritance conflict, the square will inherit the methods of the primary immediate ancestor class. *Id.* at 278.

46. Adapted from *id.* at 277.

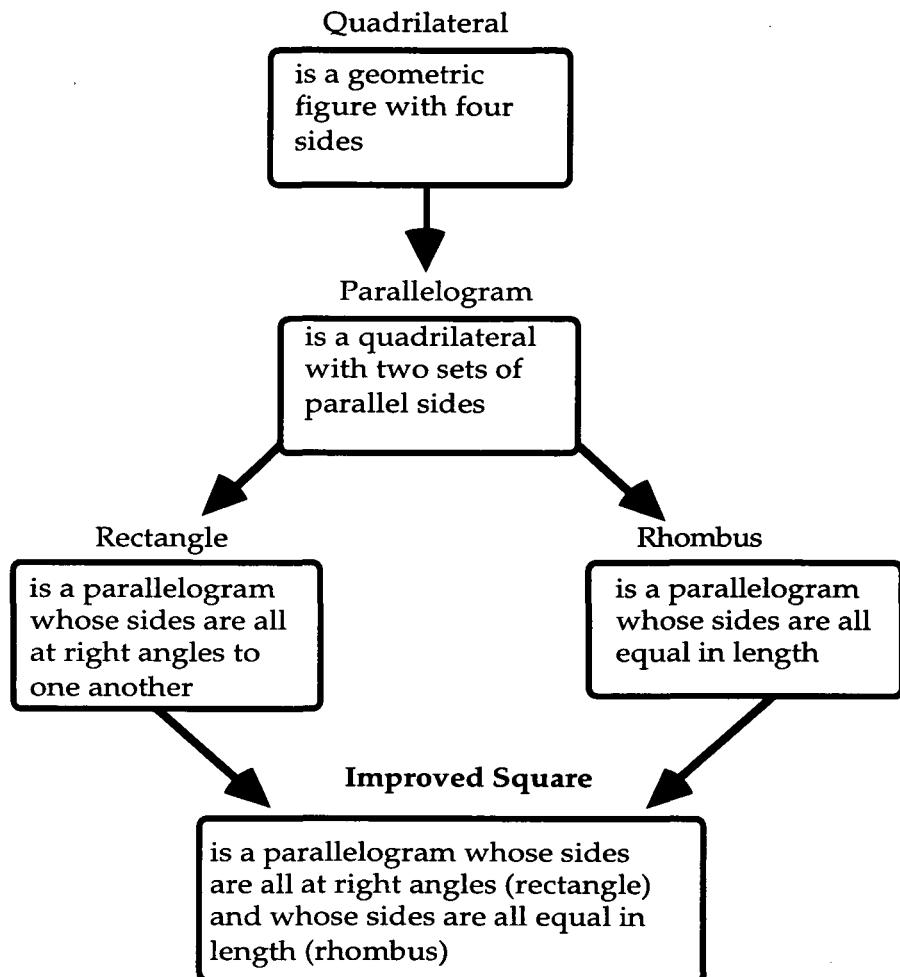


Figure 2: Improving Quad World with Multiple Inheritance

Multiple inheritance is a powerful concept that allows the programmer to model real-world entities which blend the characteristics of two or more super-classes. At the most basic level, multiple inheritance allows us to build classes as simple combinations of existing classes. For example, suppose we were working on an existing system that is designed to track the progress of efforts to save endangered wildlife, and we were instructed to add a class for "leopards."⁴⁷ Suppose further that the system already contained classes for "endangered" and "wild cat." Since leopards are both wild cats and endangered, we would

47. This example is taken directly from Korson & McGregor, *supra* note 11, at 58.

start by defining it as a sub-class of both of these existing classes. In addition, we can use multiple inheritance to model the behavior of entities that *blend* the characteristics of other objects. For example, a "houseboat" is not literally the combination of a house and a boat. Nonetheless, houseboats do possess some of the characteristics of a house and some of the characteristics of a boat. Again, a good starting point for building the "houseboat" class would be to define the class as a sub-class of both a "house" class and a "boat" class.⁴⁸ We could then add the characteristics that make a houseboat something more than the literal combination of a house and a boat by adding new messages and methods and by overriding the behavior of houses and boats that don't really apply to houseboats.⁴⁹

3. POLYMORPHISM AND DYNAMIC BINDING

In conventional programming languages, each variable has a *static type* which is defined when the program is written and remains unchanged while the program is running. In the object-oriented model, each object is defined as belonging to a particular class when the program is written; however, when the program is actually running, objects are not bound to their original class and instead may be treated as if they belong to any sub-class of the original class. Since any object can be treated as belonging to one class at one moment and as belonging to a different class at a later moment, object-oriented programming languages are said to allow for *dynamic typing*.

Polymorphism is simply a more general description of the concept of dynamic typing. The basic idea behind polymorphism is that if "Y inherits from X, [then] Y is an X, and therefore anywhere that an instance of X is expected, an instance of Y is allowed."⁵⁰ For example, consider the QuadWorld program again. Suppose the user had drawn a number of different quadrilaterals on the screen and we wished to display a textual list of all the different types of objects the user had already drawn. We might want to display a message to the user that reads, "you have drawn two squares, three parallelograms and one rectangle." One convenient way for the program to keep track of this information would be to create an object which keeps track of all the quadrilaterals currently displayed on the screen. Here's one way we could define that object:

48. SCHMUCKER, *supra* note 25, at 276.

49. Since houseboats probably have more in common with boats than houses, we would probably use the boat class as the primary immediate ancestor class. *Id.* at 278.

50. Korson & McGregor, *supra* note 11, at 45.

Class Definition of "Current Screen":

Internal Data:

- linked list⁵¹ of square objects currently displayed
- linked list of rectangle objects currently displayed
- linked list of parallelogram objects currently displayed
- linked list of rhombus objects currently displayed
- linked list of quadrilateral objects (not falling into any of the above categories) currently displayed

Messages that the object is able to perform:

- Add a square to the square list, delete a square from the square list
- Add a rectangle to the rectangle list, delete a rectangle
- Add a parallelogram to the parallelogram list, delete a parallelogram
- ... (similar messages for the other shapes)

Polymorphism allows us to find a better way to define this object. First, we note that every shape is a sub-class of the class quadrilateral. That means that wherever we have a data structure (such as a linked list) or a method that expects to use a quadrilateral, it will also accept any sub-class of the quadrilateral class. As a result, we can drastically reduce the complexity of the class "current screen" by redefining it as follows:

Class Definition of "Current Screen" Using Polymorphism:

Internal Data:

- linked list of quadrilateral objects currently displayed

Messages that the object is able to perform:

- Add a quadrilateral to the list, delete a quadrilateral from the list

When the program is actually running, we know that each element will be some type of quadrilateral, but we don't know which elements will contain which type of quadrilateral until the user has drawn some shapes on the screen. At any given moment, each element in the list will have a dynamic type which can be a square, rectangle, parallelogram, rhombus, or quadrilateral depending on which shapes are currently displayed on the screen. For example, if the user has drawn three objects on the screen, a square, a rectangle, and a quadrilateral, the "current screen" object would then contain the linked list shown in Table 1.

51. A linked list is simply a data structure that allows us to store pieces of data in a sequential chain. Each distinct piece of data in the list is called an "element" in the list. In the object-oriented world, each element can be an object.

Element #:	1	2	3
Original Definition:	Quadrilateral	Quadrilateral	Quadrilateral
Dynamic Type:	Square	Rectangle	Quadrilateral

Table 1: Illustration Of Dynamic Typing In A Linked List

Dynamic Binding is closely associated with the idea of polymorphism and dynamic typing. Dynamic binding builds on these concepts by allowing a particular object to respond differently to a particular message depending upon its dynamic type at a given moment while the program is running.⁵² For example, assume that our program accepts a command from the user to display the area of each shape currently on the screen. Dynamic binding makes this feature easy to

Element #:	1	2	3
Original Definition:	Quadrilateral	Quadrilateral	Quadrilateral
Dynamic Type:	Square	Rectangle	Quadrilateral
Class Method Invoked In Response To Message	Square	Rectangle	Quadrilateral
Formula Used In Response To The Message: "Calculate Area"	square of the length	length X width	generic formula applicable to any four-sided shape

Table 2: Illustration Of Dynamic Binding In A Linked List

implement and automatically incorporates the special methods we wrote to take advantage of the fact that there is a simpler formula to calculate the area of a rectangle than the area of a generic quadrilateral.

To implement this feature, the program responds to the user's request to calculate the area of each displayed shape by sending the "calculate area" message to each object in the linked list stored in the "current screen" object. When each object in the list receives the "calculate area" message, it will use the method associated with its dynamic type rather than the method associated with its originally defined class. Thus, the linked list will respond as shown in Table 2.

52. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 63. ("Static binding means that the types of all variables and expressions are fixed at the time of compilation; *dynamic binding* (also called *late binding*) means that the types of all variables and expressions are not known until runtime.").

Dynamic binding and polymorphism provide several immediate advantages. First, they encourage a high degree of generalization by permitting the programmer to write procedures that apply to any quadrilateral, but that respond with the optimal code depending on whether the particular quadrilateral is a square, rectangle, or parallelogram. In the object-oriented world, we simply send the "calculate area" message to some unknown quadrilateral and it doesn't matter that we have no way of knowing what type of quadrilateral will actually receive that message when the program is running. In contrast, traditional programming would require us to add code that essentially said "If the particular quadrilateral is a square then use the procedure for squares, but if the particular quadrilateral is a rectangle then use the rectangle procedure, but if the particular quadrilateral is a parallelogram, then use the parallelogram procedure, . . . otherwise use the generic quadrilateral procedure."⁵³

Moreover, dynamic binding and polymorphism also promote reusability by allowing other programmers to create new sub-classes of the quadrilateral class and know that they will automatically work in any procedure that expects to use the quadrilateral class. The programmers also know that if they have written new methods in the sub-class that override methods in the quadrilateral class the new methods will automatically be used when any procedure sends a message to their objects.

53. Of course traditional programming languages have a shorthand expression for this problem. In Pascal, the programmer would use a "case" statement that lists the names of different procedures for the different quadrilaterals, and in C the programmer would use a "switch" statement that listed the procedures. Nonetheless, these statements are no more than shorthand expressions for the long quotation in the text.

For those readers familiar with Pascal or C, consider a procedure that must re-draw a screen filled with various quadrilaterals. We could implement this procedure in Object Pascal by the following piece of code:

```
for i:= 1 to Number_of_Shapes do  
  current_figure.item(i).draw;
```

{current_figure is an array
of quadrilateral objects, and
draw is a message in each of
the quadrilateral subclasses
that tells the object to draw
itself}

Korson & McGregor, *supra* note 11, at 46. "At each pass through the loop, the code matching the dynamic type of current_figure.item(i) will be called. Note that if additional kinds of shapes are added to the system, this code segment remains unchanged. Contrast the resulting simplicity and extensibility as compared with a traditional case statement design." *Id.*

C. The Process of Designing Software Under the Object-Oriented Model

Given the concepts of objects, classes, inheritance, polymorphism, and dynamic binding, we can formulate an analytical approach for writing software that takes advantage of the object-oriented model. There have been many formal attempts to define an "object-oriented approach" to software design,⁵⁴ this Section outlines the basic features common to most of these models. In reading this Section, compare this model to the traditional model of software design as understood by the *Whelan* court and Nimmer.

1. IDENTIFY THE OBJECTS AND CLASSES THAT COMPRISSE THE "PROBLEM DOMAIN"

The first step in approaching object-oriented design is to learn as much as possible about the problem that the program is supposed to solve ("the problem domain"). As a first approximation, the programmer should approach the problem not as a computer scientist but rather by becoming an expert in a specific domain. For example, a programmer writing a navigational system for an airplane should initially learn from pilots, air controllers, and aeronautical engineers how navigation works:

Essentially, the developer must act as an abstractionist. By studying the problem's requirements and/or by engaging in discussions with domain experts, the developer must learn the vocabulary of the problem domain. The tangible things in the problem domain, the roles they play, and the events that may occur form the candidate classes and objects of our design, at its highest level of abstraction.⁵⁵

Once the developer learns the vocabulary and physical items used by pilots, air controllers, and aeronautical engineers, the programmer can begin to identify specific classes that will be needed to write navigational software. At this stage, the programmer is examining the problem domain from a fairly high level of abstraction.⁵⁶ Various commentators have identified formal categories that may help suggest candidate classes as illustrated in Table 3.

54. See, e.g., BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4; Henderson-Sellers & Edwards, *supra* note 9; Korson & McGregor, *supra* note 11; Ronald J. Norman, *Object-oriented Systems Analysis: A Methodology for the 1990s*, *J. Sys. MGMT.*, July 1991, at 32; Rebecca J. Wirfs-Brock & Ralph E. Johnson, *Surveying Current Research in Object-Oriented Design*, *COMM. ACM*, Sept. 1990, at 104.

55. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 191.

56. Norman, *supra* note 54, at 33 ("It is not necessary and certainly not required that all possible objects be identified during this step. Only the most intuitive and obvious ones may be identified here, while others or refinements of these may be identified during a later step.").

Physical Items	planes, wings, engines, fuel pump, radio beacon
Roles	pilot, copilot, navigator, passenger
Events	landing, take-off, turning, putting down landing gear
Interactions	clearance from air controller, radio contact, schedules, connections with other planes
Places	airport, destination, origin

Table 3: Types Of Classes That Are Likely To Be Used⁵⁷

This approach has several advantages over traditional software design. First, rather than asking "what tasks must the program perform," object-oriented design asks "how would those who will be relying on this program describe their problem, and what would *they* identify as the major actors (both human and inanimate) in the problem domain." This direct focus on the problem domain forces the programmer to address the specific needs of users in the problem domain before writing any code. In contrast, the traditional programming model promotes an early emphasis on the "tasks" that the software must perform and thereby removes the focus from the problem domain. At an early stage of the design process, the traditional programmer becomes bound to the specific instructions that will be used to write the program, often before potential users have identified all of their requirements. Second, the object-oriented approach helps to reveal commonalities that may exist across similar applications (vertical domain analysis) as well as commonalities that can be reused in different parts of the same application (horizontal domain analysis):

For example, when starting to design a new patient-monitoring system, it is reasonable to survey the architecture of existing systems to understand what key abstractions and mechanisms were previously employed and to evaluate which were useful and which were not. Similarly, an accounting system must provide many different kinds of reports. By treating these reports as a single domain, a domain analysis can lead the developer to an understanding of the key abstractions and mechanisms that serve all the different kinds of reports. The resulting classes and objects reflect a set of key abstractions and mechanisms generalized to the immediate report-generation problem; therefore, the resulting design is likely to be simpler than if each report had been analyzed and designed separately.⁵⁸

57. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 141 (summarizing categories proposed by Shlaer, Mellor, Ross, Coad, and Yourdon); see also Norman, *supra* note 54, at 40, Table 3 (identifying categories as "tangible items," "roles played by people or organizations," "incidents which happen at a specific point in time," "interaction [sic] that have a transaction-like quality," and "specification [sic] that have table-like qualities such as sales offices, state codes, standard industry codes, [and] tax rates").

58. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 142-43.

The object-oriented programmer's first written output is apt to be a rough list of classes and objects whose names imply their basic role in the problem domain and which will be used as the "common vocabulary of discourse among the developers."⁵⁹ Most important, these classes and objects should be subject to continual revision as the programmer follows the other three steps, thus leading to iterative and evolutionary changes in the original model of the problem domain, or what some commentators have called "Round-Trip Gestalt Design."⁶⁰ Steps 2, 3, and 4 in the model are all explicitly designed to foster such reevaluation of the problem domain.

In contrast, the traditional programmer's first written task using top-down design is to produce a rough flowchart of the program, which, by its very nature, is farther removed from the problem domain, closer to the stage of writing actual software code, and more likely to lock the programmer into tight dependencies among different parts of the program. The traditional programmer's tendency is to then parcel out pieces of the project to different programmers based on the original flowchart. While nothing stops the design team from refining the flowchart later, nothing in the traditional top-down model encourages iterative or evolutionary changes in the basic flowchart. In fact, the risk that changes in one part of the program will ripple through all other parts of the program actively discourages such changes.⁶¹

59. *Id.* at 192 (noting also that "[i]n most cases, this step takes a small amount of time relative to the other three steps. Often, a single chief designer will draft a list of candidate classes and objects and then review this list with peers as a kind of sanity check." *Id.* at 191-92).

60. *Id.* at 188 ("This style of design emphasizes the incremental and iterative development of a system through the refinement of different yet consistent logical and physical views of the system as a whole."); see also Henderson-Sellers & Edwards, *supra* note 9, at 148 ("Both top-down analysis and bottom-up class design, seen as the hardest part of the entire object-oriented software life cycle, must therefore be either concurrent or, at least iterative.") (footnote omitted).

61. BOOCHE, OBJECT-ORIENTED DESIGN, *supra* note 4, at 188. The entire "top-down" vs. "bottom-up" approach to design has been the subject of significant debate within the software community. It is important to recognize that object-oriented design is neither "top-down" nor "bottom-up":

Assume that we are faced with the problem of staffing an organization to design and implement a fairly complex piece of computer hardware. We might use horizontal staffing, in which we have a waterfall progression of products, with systems architects feeding logic designers feeding circuit designers. This is an example of top-down design, and requires designers who are "tall skinny men," as Druke calls them, because of the narrow yet deep skills that each must possess. Alternately, we might use vertical staffing, in which we have good all-around designers who take slices of the entire project, from architectural conception through circuit design. The skills that these designers must have leads Druke to call them "short fat men." Unfortunately, given its inherent complexity, software development often demands that we employ "tall fat people."

2. IDENTIFY THE STRUCTURE AND SEMANTICS OF THE OBJECTS AND CLASSES

At this stage, the programmer must specify the behavior and properties that each object will possess. One possible approach is to write "a script for each object, which defines its life cycle from creation to destruction, including its characteristic behaviors."⁶² For example, once we have identified a "radio beacon" object in our navigational software, we might write a script that reads "beacon object is created when a plane is close enough to receive the signal from that beacon, and then the beacon is expected to send a radio signal at a pre-set frequency and at pre-set intervals, and then the beacon is expected to continue this behavior until the plane passes the beacon and leaves the beacon's range." Similarly, we might define a "landing gear" object which is created as soon as the software is running and is expected to be able to keep track of whether the landing gear is up or down, and send an alarm message if the landing gear is in the wrong position. The process of writing these scripts should also force the programmer to reevaluate the original list of objects and classes identified in step 1. For example, when we define the landing gear as being able to send an alarm message, we then realize that we never identified the need for an "alarm bell" object that would receive the alarm message and be used to display an alarm message on the navigator's computer screen. In this way, step 2 is iterative because it forces the programmer to reevaluate the decisions made in step 1.

In addition, identifying the behavior of each object may reveal other sub-classes that could be introduced. For example, an analysis of a bookkeeping program in step 1 might reveal the need for an "invoice" object. However, once we specify the attributes of an invoice in this step, we might realize the need for additional objects, such as "header," "account summary," and "list of transactions," that represent the different sections that make up the invoice.⁶³ Conversely, this analysis

Id.; see also Henderson-Sellers & Edwards, *supra* note 9, at 146 (noting that "object-oriented (OO) design and analysis has many attributes of both top-down and, perhaps predominately, bottom-up design. Since one of the aims of an OO implementation is the development of generic classes for storage in libraries, an approach which considers both top-down analysis and bottom-up design simultaneously is likely to lead to the most robust software systems.").

62. BOOCHE, *OBJECT-ORIENTED DESIGN*, *supra* note 4, at 192 (noting that "[t]his step is much harder than the first and takes much longer. This is the phase in which there may be fierce debates, wailing and gnashing of teeth, and general name-calling during design reviews. Finding classes and objects is the easy part; deciding upon the protocol of each object is hard").

63. Norman, *supra* note 54, at 33 (suggesting that the programmer look for whole-to-part relationships and generalization-to-specialization relationships at this point). While this activity may blur some of the distinctions between step 2 and step 3, such blurring of

could also reveal the need to redefine some of the super-classes. For example:

[A] class of 'bird' with an attribute that "birds can fly" is successful until we consider the Southern Hemisphere and "penguins," "ostriches," "kiwis" etc. In this case, one solution is to introduce an additional level in the inheritance hierarchy by introducing two children classes of class bird as "flying bird" and "non-flying bird" and redefining the parent class to remove the attributes relating to flight. This process tries to develop a logical hierarchy of objects so there are no "missing" objects.⁶⁴

3. IDENTIFY THE RELATIONSHIPS AMONG OBJECTS AND CLASSES

In step 3, the programmer must identify the relationships among the previously identified objects and classes. First, the programmer must develop the inheritance relationships and define the structure of super-classes and sub-classes. In doing so, the programmer is likely to uncover additional "patterns among classes, which cause us to reorganize and simplify the system's class structure, and patterns among cooperative collections of objects, which lead us to generalize the mechanisms already embodied in the design."⁶⁵

One way to formulate a concrete representation of these relationships is by building a *semantic data model*.⁶⁶ For example, if we were designing a program that was intended to control the traffic lights at a busy intersection, we might construct the partial semantic data model shown in Figure 3 in which rectangles denote classes and circles denote the functional relationships between classes connected by arrows.

specific steps is consistent with the iterative and evolutionary style of object-oriented design.

64. Henderson-Sellers, *supra* note 9, at 150 (footnotes omitted).

65. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 193.

66. Korson, *supra* note 11, at 47.

Using this data model, we can then determine what messages each object must accept. One useful conceptual device is to consider each message as a "service" that the object is capable of providing to any other object. Then, using the scripts from step 2 to determine what behaviors each object will exhibit, the programmer can determine what kind of services each object will need to fulfill its role.⁶⁸ As would be expected from the theme of iterative development, this process is likely to reveal that certain objects require services that no current object yet provides. The programmer must then either add that service to an existing object or create a new object to handle that service. As an alternative, the relationship between two objects can be considered to be a contract in

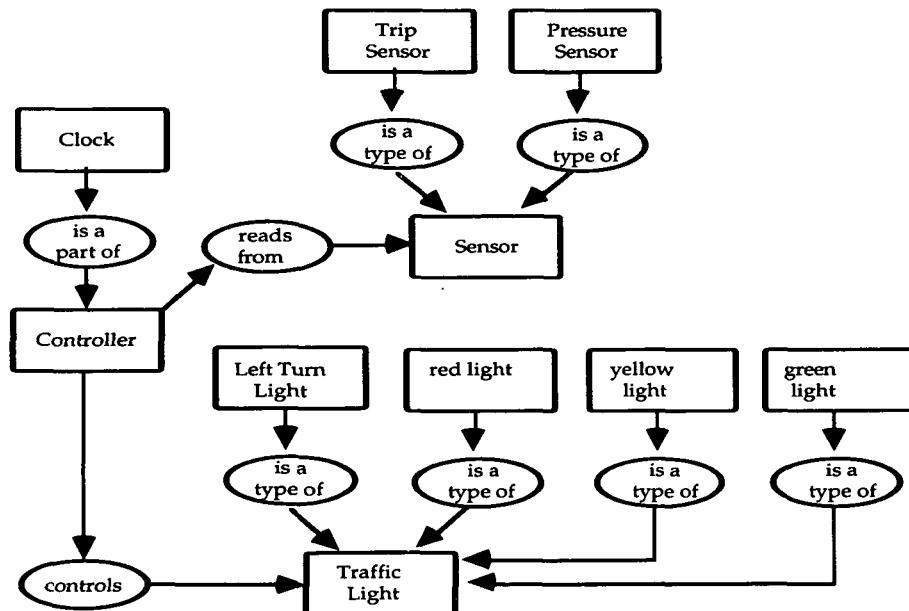


Figure 3: Semantic Data Model for Controlling Traffic Lights⁶⁷

which one object is a "client" that requests certain services from another object which is a "server" and fulfills those requests.⁶⁹ Again, the programmer must make sure that every object that needs a contract for a particular service has a corresponding server object to fulfill that contract.

67. This model is adapted from Korson & McGregor, *supra* note 11, at 48 (fig. 8).

68. Henderson-Sellers & Edwards, *supra* note 9, at 150.

69. Wirfs-Brock & Johnson, *supra* note 54, at 110-11.

4. IDENTIFY THE PUBLIC INTERFACES AND SERVICES PROVIDED BY EACH OBJECT AND CLASS

At this point, we know what role each object plays and what services the object provides to other objects. Using that information, we can define the general type of data structures needed by each object and the methods that the object will need to provide services to other objects. As the internal structure of a particular object is developed, the programmer may discover that this object can be built by using pre-existing libraries of more primitive objects.⁷⁰ At the end of this stage, the programmer can begin writing the actual source code for each method, perhaps treating each method as a miniature program that can be approached using traditional procedural techniques.

IV. COPYRIGHT PROTECTION FOR OBJECT-ORIENTED SOFTWARE

Before examining the scope of copyright protection, it is important to recognize when the fact that the object-oriented model was used to design a piece of software matters and when it does not. If a programmer writes software under the object-oriented model, the programmer will use an object-oriented programming language to write the high level source code for the program. This source code is then translated by a compiler program into object code which is a series of 1's and 0's. These 1's and 0's represent low level commands which the microprocessor can understand and execute. At the same time, the compiler can also produce an assembly language version of the source code. Assembly language is a human-readable listing of object code in which each low-level microprocessor instruction is represented by a single word, such as "jump," "store," or "link." However, once the program has been translated into object code or assembly language, the fact that the original source code was written in an object-oriented programming language is virtually impossible to detect. The microprocessor itself has no concept of object-oriented principles;⁷¹ therefore, the compiler produces a program that in object code form is indistinguishable from a program written according to traditional design methods.

As a result, in a case where the plaintiff alleges that the defendant copied the object code or assembler versions of the program, which can

70. Henderson-Sellers & Edwards, *supra* note 9, at 150.

71. While some computer manufacturers are touting "object-oriented operating systems," this statement does not mean that the microprocessor itself understands object-oriented principles. Instead, this feature means that the operating system is designed so that a program written in high level source code can interact with the operating system by using a pre-defined library of objects that perform the functions of the operating system. For example, the operating system running on a machine with a graphical user interface may supply libraries of objects for windows, icons, and menus.

arise only when verbatim copying of the program has occurred,⁷² the court can proceed without worrying about principles of object-oriented design. But, object-oriented principles are absolutely critical when the plaintiff alleges that the defendant had access to the original source code and copied it. During the trial, both plaintiff and defendant will have to produce the source code for their programs, and the court will have to determine whether the programs are "substantially similar."⁷³ In weighing the evidence of similarity, the court will need to understand how principles of object-oriented design affect that comparison.

The starting point for any discussion on copyright protection for software is 17 U.S.C. § 102(b), which excludes protection for "any idea, procedure, process, system, [or] method of operation."⁷⁴ While the theoretical limits imposed by § 102(b) seem clear, courts and commentators have struggled with the practical application of the section to computer software. This Section will review the two dominant approaches to § 102(b) and determine how they apply to object-oriented software. In addition, this Section analyzes an alternative approach which advocates an ad hoc balancing of the economic effects of protecting the plaintiff's work, but concludes that courts will not adopt this approach because it has no support in copyright doctrine and it is problematic as a matter of innovation policy.⁷⁵ Finally, this Section concludes that since the "behavioral" aspects of software are particularly dominant when the object-oriented model is used to write software, pure copyright doctrine provides almost no protection for the *object-oriented* aspects of software.

72. While it is theoretically possible to work backwards from the assembler version of the source code, this process is virtually impossible for a program of any complexity. Particularly as software projects are increasingly characterized by "programming-in-the-colossal," the assertion that a defendant could have reproduced the plaintiff's source code by disassembly is ludicrous. Thus, a defendant will find copying the object code useful only if the defendant can sell verbatim copies of the plaintiff's program, perhaps in a foreign market where explicit piracy is tolerated. For a discussion of the technical difficulties involved in disassembly, see G. Gervaise Davis III, *Reverse Engineering and the Computer Industry: A Battle Between Legal and Economic Principles* (1991) (unpublished presentation on file with the *High Technology Law Journal*); Ronald S. Laurie, *Protection of Trade Secrets in Object Form Software: The Case for Reverse Engineering*, COMPUTER LAW., July 1984, at 1. But cf. Allen R. Grogan, *Decompilation and Disassembly: Undoing Software Protection*, COMPUTER LAW., Feb. 1984, at 1 (arguing that disassemblers and decompilers may allow object code to be converted so that much of the logic of the program is revealed).

73. See NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.26 ("In many software cases, access is either conceded or easily proved, so that a finding of infringement turns entirely on whether the works are substantially similar.").

74. 17 U.S.C. § 102(b) (1988).

75. The term "innovation policy" is used to denote the best mix of legal incentives that would maximize the total value of new software inventions.

A. *Whelan* and Its Progeny

Whelan Associates v. Jaslow Dental Laboratory represents the broadest approach to protecting computer software. In *Whelan*, the court examined the idea/expression dichotomy⁷⁶ stated in § 102 and concluded that "the purpose or function of a utilitarian work would be the work's idea, and everything that is not necessary to that purpose or function would be part of the expression of the idea."⁷⁷ Since this rule forces the court to focus on *one* idea behind the program, courts applying this test will necessarily define an idea which represents an extremely high level of abstraction. For example, the *Whelan* court characterized the idea behind the plaintiff's program as the efficient operation of a dental laboratory.⁷⁸ At this high level of abstraction, there are of course many ways to write a program which performs that general function and all program elements at lower levels of abstraction would constitute copyrightable expression. As a result, the *Whelan* test protects the "structure, sequence, and organization" of source code as a general rule.

76. The limitations expressed in § 102(b) create what is known as the "idea/expression dichotomy" in copyright law:

The crucial consideration in the analysis that follows is that copyright law protects only an author's original expression, not ideas or elements taken from preexisting works. Infringement is shown by a substantial similarity of *protectable expression*, not just an overall similarity between the works. Thus, before evaluating substantial similarity, it is necessary to eliminate from consideration those elements of a program that are not protected by copyright.

NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.28 to .29.

77. *Whelan Assocs. v. Jaslow Dental Lab.*, 797 F.2d 1222, 1236 (3d Cir. 1986), *cert. denied*, 479 U.S. 1031 (1987).

78. *Id.* at 1238 n.34. Other cases applying the *Whelan* test have provided wide-ranging software protection either by broadly defining the "idea" behind the plaintiff's program or by finding that "other ways" exist to express a particular idea. See, e.g., *Johnson Controls v. Phoenix Control Sys.*, 886 F.2d 1173 (9th Cir. 1989) (general finding that "the structure of the JC-5000S [plaintiff's entire program] is expression, rather than an idea in itself," apparently because "each individual application is customized to the needs of the purchaser. This practice of adaptation is one indication that there may be room for individualized expression in the accomplishment of common functions."); *Lotus Dev. Corp. v. Paperback Software Int'l*, 740 F. Supp. 37, 65-66 (D. Mass. 1990) (repeatedly asking whether there were other ways to express the idea of "an electronic spreadsheet"); *Pearl Sys. v. Competition Elec.*, 8 U.S.P.Q.2d (BNA) 1520, 1524 (S.D. Fla. 1988) (defining the idea behind two subroutines as providing "a method for the user to set a par time" and as allowing "the user to review the shots he or she has fired and to learn of the time that elapsed between each shot"); *Digital Communications v. Softklone Distrib.*, 659 F. Supp. 449, 459 (N.D. Ga. 1987) ("The use of a screen to reflect the status of the program is an 'idea'; the use of a command driven program is an 'idea'; and the typing of two symbols to activate a specific command is an 'idea.' "); *Broderbund Software, Inc. v. Unison World*, 648 F. Supp. 1127, 1133 (N.D. Cal. 1986) (defining the idea of the plaintiff's program as "the creation of greeting cards, banner, posters and signs that contain infinitely variable combinations of text, graphics, and borders").

Courts applying *Whelan* to object-oriented software are likely to protect the basic inheritance relationships among objects. For example, the idea behind the QuadWorld program could be expressed as a program to allow for the efficient drawing of quadrilaterals on a computer screen. Since there are undoubtedly many ways to write such a program, the particular choice of classes, sub-classes, and messages is copyrightable expression. In fact, the court could argue that the idea behind QuadWorld could be achieved using traditional programming techniques, and since this would produce entirely different looking source code from the object-oriented version, that variation alone proves the necessary range of expression to justify copyright protection.⁷⁹

Moreover, the court could use our analysis of the design process to argue that high level inheritance relationships and class structures must be protected by copyright. In *Whelan*, the court justified the protection of structure, sequence, and organization in part on the basis that "among the more significant costs in computer programming are those attributable to developing the structure and logic of the program. The rule proposed here, which allows copyright protection beyond the literal computer code, would provide the proper incentive for programmers by protecting their most valuable efforts."⁸⁰ While the validity of this argument is highly doubtful in light of the Supreme Court's recent decision in *Feist*,⁸¹ lower courts may still be tempted to protect those parts of the plaintiff's software that are the products of significant time and effort. Under our analysis of the object-oriented design process, steps 2 and 3, identifying the structure and semantics of the objects and classes and identifying the relationships among these objects and classes, represent the most difficult parts of the design process.⁸² The decisions made in those steps are critical to the quality of the final program.⁸³ Thus, a court could use this

79. This argument is arguably analogous to one proposed in *Lotus*. The *Lotus* court held that the differences between the user interface for Microsoft Excel on the Macintosh and the user interface for the Lotus program were evidence that there are multiple ways to express the idea of an electronic spreadsheet. *Lotus*, 740 F. Supp. at 65-66. The court was oblivious to the fact that Excel's user interface was entirely attributable to the Macintosh operating system (all programs running on the Macintosh have that same interface) and had nothing to do with how Excel chose to express the idea behind an electronic spreadsheet.

80. *Whelan*, 797 F.2d at 1237.

81. *Feist Publications v. Rural Tel. Serv., Inc.*, 111 S. Ct. 1282, 1290 (1991) (repudiating "sweat of the brow" theories for copyright protection because "the primary objective of copyright is not to reward the labor of authors, but '[t]o promote the Progress of Science and useful Arts'").

82. See *supra* parts III.C.2, III.C.3.

83. While the iterative nature of object-oriented development encourages refinement of the decisions made in steps 2 and 3, the court will see only the final program and thus will not be able to determine which decisions were initially made during the first pass through the design process and which were added later by refining the decisions made in steps 2 and 3. As a result, the references to steps 2 and 3 in this discussion include all decisions

argument to protect the general inheritance relationships between classes, the detailed scripts for each object, and the collections of services each object is expected to provide.

The *Whelan* court's analysis of § 102(b) has been heavily criticized by commentators,⁸⁴ and for good reason. The primary criticism of *Whelan* has focused on *Whelan's* use of a *single* idea existing in each computer program:

The crucial flaw in [*Whelan's*] reasoning is that it assumes only one "idea," in copyright law terms, underlies any computer program, and that once a separable idea can be identified, everything else must be expression. All computer programs are intended to cause the computer to perform some function. The broad purpose that the program serves, be it managing a dental laboratory, automating a factory, or dispensing cash at a bank teller machine, is *an* idea. Other elements of the program's structure and design, however, may also constitute ideas for copyright purposes.⁸⁵

Similarly, in *Computer Associates v. Altai*, a district court adopted this criticism and then used the traditional model of programming to further reveal *Whelan's* flaws:

In the case at bar, Dr. Davis [court-appointed expert] pointed out further technical flaws in the *Whelan* analysis which render its reasoning inadequate. As he so convincingly demonstrated, a computer program is made up of sub-programs and sub-sub-programs, and so on. Each of those programs and sub-programs has at least one idea. Some of them could be separately copyrightable; but many of them are so standard or routine in the computer field as

that fall within the general subject matter of those steps whether or not they were actually made in those steps or at a later time.

84. See, e.g., NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.33 to .34; Richard A. Beutel, *Software Engineering Practices and the Idea/Expression Dichotomy: Can Structured Design Methodologies Define the Scope of Software Copyright*, 32 JURIMETRICS J. 1, 17-20 (1991); Nimmer et al., *supra* note 3, at 629-30, 639; Reback & Hayes, *supra* note 3, at 3-4. Several cases have also rejected *Whelan* or have recognized its existence but then implicitly failed to apply it. See *Computer Assocs. Int'l v. Altai, Inc.*, 982 F.2d 693, 706 (2d Cir. 1992) ("We think that *Whelan's* approach to separating idea from expression in computer programs relies too heavily on metaphysical distinctions and does not place enough emphasis on practical considerations."); *Sega Enter. v. Accolade, Inc.*, 977 F.2d 1510, 1524 (9th Cir. 1992) ("The *Whelan* rule, however, has been widely—and soundly—criticized as simplistic and overbroad."); *Plains Cotton Co-op Ass'n v. Goodpasture Computer Serv., Inc.*, 807 F.2d 1256, 1262 (5th Cir. 1987) (declining "to embrace *Whelan*"), cert. denied, 484 U.S. 821 (1987); *Computer Assocs. Int'l v. Altai, Inc.*, 775 F. Supp. 544, 558-59 (E.D.N.Y. 1991) (describing *Whelan* as setting "forth what now seems to be a simplistic test for similarity between computer programs"), aff'd in relevant part, 982 F.2d 693 (2d Cir. 1992); *Manufacturers Technologies, Inc. v. CAMS, Inc.*, 706 F. Supp. 984, 992 (D. Conn. 1989) (not explicitly rejecting *Whelan* but arguing that the *Broderbund* court's application of *Whelan* to screen displays, "overextended the scope of copyright protection applicable to those screen displays"); *Healthcare Affiliated Servs., Inc. v. Lippman*, 701 F. Supp. 1142 (W.D. Pa. 1988).

85. NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.33 to .34.

to be almost automatic statements or instructions written into a program.⁸⁶

Despite this criticism, *Whelan* has never been overruled and is still the starting point for most discussions of copyright protection for computer software.

B. The Filtering Approach⁸⁷

In contrast to *Whelan's* "one idea" approach, Nimmer starts with the "patterns of abstractions" test⁸⁸ and concludes that the court must apply a series of standard copyright doctrines to filter out unprotectable ideas at each level of abstraction. This test is easy to defend because each filter is closely tied to a specific copyright doctrine and thus forces the court to account for every theory that can limit the number of program elements entitled to protection. Nimmer proposes that the court apply four basic filters: abstract ideas, merger, *scenes a faire*, and public domain. While Nimmer's test was closely tied to the traditional model of software development, it can still be applied to object-oriented software by altering the relative importance of each filter.

1. ABSTRACT IDEAS

Nimmer's first filter revisits the basic problem of separating protectable ideas from non-protectable expression. In the context of traditional software, this filter provides a strong limit on copyright protection because the top-down approach to software development "provides natural divisions, which may correspond to the various levels of abstractions that the court seeks to identify and analyze."⁸⁹ In Nimmer's view, the court can divide the software into programs, sub-

86. *Computer Assocs.*, 775 F. Supp. at 559.

87. This approach was first developed in Nimmer et al., *supra* note 3, at 635-55, and is summarized in NIMMER & NIMMER, *supra* note 3, at § 13.03[F]. The Second Circuit has recently endorsed this approach to substantial similarity. *Computer Assocs.*, 982 F.2d at 706.

88. The test was first developed by Judge Learned Hand:

Upon any work, and especially upon a play, a great number of patterns of increasing generality will fit equally well, as more and more of the incident is left out. The last may be no more than the most general statement of what the play is about, and at times might consist only of its title; but there is a point in this series of abstractions where they are no longer protected, since otherwise the playwright could prevent the use of his "ideas," to which, apart from their expression, his property is never extended.

Nichols v. Universal Pictures Corp., 45 F.2d 119, 121 (2d Cir. 1930), *cert. denied*, 282 U.S. 902 (1931).

89. Nimmer et al., *supra* note 3, at 638 ("[T]he systematic method used to develop computer programs makes the abstractions test facially more applicable to computer software than other types of works. Traditional literary works are not created in such a consistently organized and orderly fashion.").

programs, and sub-sub-programs and then determine at which level the code passes from being an unprotectable idea to being protectable expression.

Two problems bar meaningful application of this filter to object-oriented software. First, Nimmer himself admitted that even in the context of structured top-down programming, the test is not easy to apply.⁹⁰ As one commentator complained, "simply to characterize the filter as eliminating 'abstract ideas' says very little about what is, and is not, an 'idea.' One man's 'abstract idea' may be another's protectable expression."⁹¹ Second, the iterative nature of object-oriented development prevents the court from finding easy lines to draw in determining what is a "level of abstraction." The process of "round-trip gestalt design" will tend to blur meaningful line drawing on the basis of the design process itself.

As an alternative, we could define the levels of abstraction by considering class lists, inheritance relationships, and the semantic data model to each be separate levels of abstraction. However, these lines create extremely broad categories which may encourage the court to find the same single idea behind each level of abstraction. For example, a court examining the traffic light problem might conclude that the idea behind the list of classes is the "efficient management of a traffic intersection." But, if the court then examines the inheritance structure and semantic data model, it seems that the idea at those levels of abstraction is also the efficient management of a traffic intersection. In the context of object-oriented software, this alternative leads courts back to the heavily criticized "one idea" approach of *Whelan*.

2. MERGER

The merger filter operates to exclude elements of the program that can only be expressed in one way.⁹² In the context of computer software, "merger issues may arise in somewhat unusual ways. Although theoretically many ways may exist to implement a particular idea, efficiency concerns can make one or two choices so compelling as to virtually eliminate any form of expression."⁹³ In this category, Nimmer lists such low-level routines as searching and sorting algorithms, which should not be protected, because "the fact that two programs both use the most efficient sorting or searching method available supports an inference of independent creation as readily as it supports one of copying, and thus

90. NIMMER & NIMMER, *supra* note 3, at 13-78.33.

91. Beutel, *supra* note 84, at 23.

92. NIMMER & NIMMER, *supra* note 3, at 13-78.35.

93. *Id.*

is not reliable evidence that copying occurred."⁹⁴ These considerations have finally received explicit judicial recognition by the Ninth Circuit which has expressed the impact of merger even more broadly than Nimmer:

To the extent that there are many possible ways of accomplishing a given task or fulfilling a particular market demand, the programmer's choice of program structure and design may be highly creative and idiosyncratic. However, computer programs are, in essence, utilitarian articles—articles that accomplish tasks. As such, they contain many logical structural, and visual display elements that are dictated by the function to be performed, by considerations of efficiency, or by external factors such as compatibility requirements and industry demands.⁹⁵

For most merger issues, object-oriented software can be analyzed in the same manner as traditional software. Sorting and searching routines would be used primarily by the *internal* implementation of a specific object's methods. Since this internal implementation may itself have been written using traditional structural programming techniques, courts should be able to apply this test without alteration. Similarly, other merger concerns, such as ensuring compatibility with particular hardware and software, should not raise issues unique to object-oriented software.⁹⁶ In general, courts should find the merger filter to be a powerful tool for limiting infringement claims relating to the internal implementations of specific objects.⁹⁷

94. *Id.* at 13-78.36. It must be emphasized that copyright law does not prevent a defendant from producing a substantially similar program, as long as the defendant did not actually copy the plaintiff's work. Copying is an absolute prerequisite for infringement, and the analysis of substantial similarity is used only to raise the inference of copying because direct evidence of copying rarely exists. See, e.g., *Computer Assocs.*, 982 F.2d at 708 ("Since, as we have already noted, there may be only a limited number of efficient implementations for any given task, it is quite possible that multiple programmers, working independently, will design the identical method employed in the allegedly infringing work. Of course, if this is the case, there is no copyright infringement.").

95. *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510, 1524 (9th Cir. 1992); see also *Computer Assocs.*, 982 F.2d at 708 ("[W]hen one considers the fact that programmers generally strive to create programs 'that meet the user's needs in the most efficient manner,' the applicability of the merger doctrine to computer programs becomes compelling. . . [T]he more efficient a set of modules are, the more closely they approximate the idea or process embodied in that particular aspect of the program's structure." (quoting Menell, *supra* note 3, at 1052; citation omitted)).

96. See *Sega*, 977 F.2d at 1526 (allowing intermediate copying in order to ensure compatibility with videogame hardware); *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832 (Fed. Cir. 1992) (same).

97. Merger analysis should not be used when evaluating semantic data models or the general structure of particular object classes and sub-classes. For most sophisticated and complex programs, it is highly unlikely that only one efficient object-oriented structure exists. The analysis of semantic data models is better addressed by the *scènes à faire* doctrine discussed in the next sub-section.

3. SCENES A FAIRE

Scenes a faire represents the most powerful filter for object-oriented software. Nimmer used the term to justify excluding program elements dictated by "external considerations," such as hardware standards, software standards, computer manufacturers' design standards, target industry practices, and computer industry programming practices.⁹⁸ While these considerations can certainly be applied to object-oriented software, traditional case law dealing with *scenes a faire* will actually be more important in eliminating elements of object-oriented software from copyright protection.

Under the *scenes a faire* doctrine, copyright protection is denied for "those elements that follow naturally from the work's theme rather than from the work's creativity."⁹⁹ In the literary context, *scenes a faire* has precluded protection for stock literary devices or stock character types that are inherent in the general theme of the work.¹⁰⁰ For example, in *Shaw v. Lindheim*,¹⁰¹ the court examined the mood, setting, and pace of the plaintiff's and defendant's television scripts and concluded that "[b]oth works are fast-paced, have ominous and cynical moods that are lightened by the [hero's] victory, and are set in large cities. These similarities are common to any action adventure series, however, and do not weigh heavily in our decision."¹⁰²

Particularly in the case of software designed to model real-world behavior, this approach to *scenes a faire* justifies excluding from protection software elements that are dictated by the real-world behavior being modeled. This understanding of the doctrine has already been accepted by several courts evaluating traditional software. For example, in *Data East USA, Inc. v. Epyx, Inc.*,¹⁰³ the Ninth Circuit analyzed two computer karate games and concluded that infringement could not be based on program elements that "encompass the idea of karate."¹⁰⁴ In doing so, the court approved of the district court's finding that:

[T]he visual depiction of karate matches is subject to the constraints inherent in the sport of karate itself. The number of combatants, the stance employed by the combatants, established and recognized moves and motions regularly employed in the sport of karate, the regulation of the match by at least one referee or judge, and the manner of scoring by points and half points are among the constraints inherent in the sport of karate. Because of these

98. NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.36 to .43.

99. Nimmer et al., *supra* note 3, at 642.

100. See *Hoehling v. Universal City Studios, Inc.*, 618 F.2d 972 (2d Cir. 1980), cert. denied, 449 U.S. 841 (1980).

101. 919 F.2d 1353 (9th Cir. 1990).

102. *Id.* at 1363.

103. 862 F.2d 204 (9th Cir. 1988).

104. *Id.* at 209.

constraints, karate is not susceptible of a wholly fanciful presentation.¹⁰⁵

Similarly, in *Plains Cotton Co-op Ass'n v. Goodpasture Computer Serv., Inc.*,¹⁰⁶ the Fifth Circuit refused to find infringement because the "appellees presented evidence that many of the similarities between the GEMS and Telcot programs are dictated by the externalities of the cotton market."¹⁰⁷ As a result, the plaintiff could not claim protection for program elements that were designed to imitate a "cotton recap sheet," because that was a stock element in the real-world cotton market and necessary to any program trying to model that market.¹⁰⁸ Finally, in *Q-Co Industries, Inc. v. Hoffman*,¹⁰⁹ the court examined two tele-prompting programs and found no protectable expression because "the same modules would be an inherent part of any prompting program. Their order and organization can be more closely analogized to the concept of wheels for the car rather than the intricacies of a particular suspension system."¹¹⁰

These cases provide strong authority for excluding many of the object-oriented elements in a program that models real-world behavior. For example, in QuadWorld, the entire class and inheritance structure flows directly from the natural relationships between squares, rectangles, parallelograms, and quadrilaterals which, in turn, are dictated by formal mathematical definitions in the real world. Similarly, in the traffic light control program, nothing in the semantic data model would be protectable because these relationships are dictated by the functional behavior of trip sensors, controllers, and traffic lights. Finally, the relevance of *scenes à faire* to object-oriented software is further underscored by our approach to design in step 2, in which we wrote "scripts" for each object, making it fairly easy for a court to compare each object to a "stock character" in the real-world system being modeled.

In most situations, the list of services that each object must provide will be largely dictated by these scripts and hence will be unprotectable. In certain cases, it might be possible to identify certain low-level objects¹¹¹

105. *Id.*

106. 807 F.2d 1256 (5th Cir. 1987), *cert. denied*, 484 U.S. 821 (1987).

107. *Id.* at 1262.

108. *Id.* at 1262 n.4.

109. 625 F. Supp. 608 (S.D.N.Y. 1985)

110. *Id.* at 616 (citation omitted).

111. By low-level objects, I mean certain objects which are simply building blocks in constructing more complex objects which model real-world properties. At this low level, the building block may be sufficiently removed from real-world behavior to render *scenes à faire* inapplicable. However, even these objects may often be taken from libraries of reusable objects and should be excluded from protection because they do not satisfy copyright's originality requirement. See *infra* Section IV.B.4 discussing the public domain filter.

that do not directly model real-world behavior and could therefore escape the *scènes à faire* filter. In general, however, the only elements that will survive this filter are low-level implementations of specific methods; at that level, those portions of code resemble traditional programs and embody few object-oriented principles.

4. PUBLIC DOMAIN

The "public domain" filter will also be extremely important in analyzing object-oriented software. Since object-oriented design focuses on reusable software components, many complex object-oriented programs will take advantage of existing objects that have been written for other programs. In some cases, these objects may be taken from public domain libraries, such as those provided on electronic bulletin boards. As Nimmer notes, "It is axiomatic that material in the public domain is not protected by copyright even when incorporated into a copyrighted work."¹¹² As a result, the court must eliminate any objects taken from public domain when determining which elements of the program are protectable.

However, the bulk of reusable objects may not come from entirely "public" sources. These reusable objects may come from vendors selling libraries of pre-defined objects on a license basis, particularly in the case of graphical user interfaces and database systems. These vendors clearly intend that their libraries will be incorporated into commercial products.¹¹³ Nonetheless, these objects should not be included in the scope of the copyright protection for the final commercial product, as they would not be original to the programmer claiming authorship of the final product, and hence could not pass copyright's threshold test for originality.¹¹⁴ As a result, the court must treat the use of licensed objects

112. NIMMER & NIMMER, *supra* note 3, at 13-78.43 (citing *Sheldon v. Metro-Goldwyn Pictures Corp.*, 81 F.2d 49, 54 (2d Cir. 1936), *aff'd*, 309 U.S. 390 (1940)); *see also Computer Assocs. Int'l v. Altai, Inc.*, 982 F.2d 693, 710 (2d Cir. 1992) (public domain "material is free for the taking and cannot be appropriated by a single author even though it is included in a copyrighted work").

113. In addition to the previously discussed object libraries for implementing the Macintosh user interface, see *supra* note 6, vendors are hawking a wide variety of object libraries for use in commercial applications. A quick perusal of advertisements and articles in any programming trade magazine will confirm the growth of this industry. *See, e.g., COMM. ACM*, Oct. 1991.

114. Copyright protection is allowed only for *original* works of authorship. 17 U.S.C. § 102(a) (1988). At a minimum, original authorship means that the programmer did not directly take the expression from any other source, whether public or not. *Feist Publications v. Rural Tel. Serv., Inc.*, 111 S. Ct. 1282, 1287 (1991) ("The sine qua non of copyright is originality. To qualify for copyright protection, a work must be original to the author. Original, as the term is used in copyright, means only that the work was independently created by the author (as opposed to copied from other works), and that it possesses at least some minimal degree of creativity." (citation omitted)).

on the same basis as truly "public domain" objects. In both cases, copyright protection would not be available for any object which the programmer seeking protection did not write.

C. Economic Balancing Approach

While no court has yet adopted the economic balancing approach, several recent commentators on software protection have suggested answering the idea/expression problem by balancing the copyright plaintiff's creative contribution against the loss to society from granting the plaintiff a monopoly over particular software code. In one version of this approach, the court would determine the existence of protectable expression by following a two-step test:

The first step is for the court to define as specifically as possible the thing that the defendant has taken from the plaintiff . . . the second step is to decide whether that thing is original to the plaintiff . . . That is, to get that thing the defendant took, did the plaintiff invest costly creative effort that presumptively relied on the promise of copyright? If so, judgment properly goes to the plaintiff, because, in conclusory terms, the defendant has taken the plaintiff's expression. Or did the plaintiff get that thing by copying it effortlessly from existing and available sources, or by otherwise responding entirely to incentives other than copyright? If so, judgment properly goes to the defendant because, again stating it in conclusory terms, the defendant took only the plaintiff's idea.¹¹⁵

In another version of the economic balancing approach, the court would divide the plaintiff's program at different levels of abstraction and then determine the dividing line between idea and expression by "balancing the need to provide an incentive to authors against the cost to society of losing the free use of the author's work at that level of expression."¹¹⁶

While the economic balancing approach seems intriguing as a matter of innovation policy, courts are not likely to endorse it primarily because it is not supported by copyright doctrine. Both versions require the court to parse the plaintiff's work in a manner similar to the "patterns of abstractions" test first articulated in *Nichols v. Universal Pictures Corp.*¹¹⁷ However, both versions ultimately depart from copyright doctrine by requiring the court to balance the economic return necessary to induce the author to produce a particular type of work against the cost to society of granting that author a monopoly over particular expression at a particular level of abstraction. This equation confuses the distinctions between copyright and patent law. In patent law, the author's creative

115. Wiley, *Copyright at the School of Patent*, 58 U. CHI. L. REV. 119, 158-59 (1991).

116. Reback & Hayes, *supra* note 3, at 5.

117. 45 F.2d 119, 121 (2d Cir. 1930), cert. denied, 282 U.S. 902 (1931).

contribution is assessed by the requirements of utility, novelty, and non-obviousness.¹¹⁸ The cost to society is controlled by requiring the inventor to define the invention with specific claim language sufficiently narrow to avoid the prior art and by requiring that those claims be supported by the specification, thus ensuring that most patents will have a fairly narrow scope. Moreover, the costs of protection are offset by the societal benefits resulting from full disclosure of the underlying technology in the patent specification.

In contrast, copyright asks little of the author except that the work not be copied from any other source and that the work reflect at least minimal creativity.¹¹⁹ While copyright law limits the monopoly costs to society by allowing independent creation to be an absolute defense to infringement, it provides no doctrinal tools for defining the scope of the monopoly against potential defendants who have had access to the work. As a result, the economic approach is difficult to support with copyright doctrine. In fact, the author of the first version acknowledged this dilemma and explicitly developed his test by applying the "good sense" of patent doctrine in order to "rationalize" copyright doctrine.¹²⁰

Even if the economic approach could somehow be justified under traditional copyright doctrine, it is not clear that the economic approach would be particularly desirable as a matter of policy. Under the first version, the court would face the elusive task of *objectively* determining whether the plaintiff would have authored the code appropriated by the defendant in the absence of copyright incentives. Beyond the obvious evidentiary problems in this analysis, the process of analyzing incentives *ex post* leads unavoidably to circular reasoning. Whether the plaintiff was motivated by the promise of copyright depends to a large degree on the generally perceived rule of law regarding the scope of software copyrights. But, at the same time, the purpose of the two-part test itself is to determine the proper scope of the idea/expression dichotomy and hence announce a new rule of law. This dilemma is further exacerbated by the small number of software copyright cases that result in published decisions.

The second version of the economic approach faces similar problems. First, the *ad hoc* nature of the inquiry makes it difficult for

118. See 35 U.S.C. §§ 101-103 (1988).

119. While the *Feist* case may have raised the standard of originality required by copyright, it did not raise that standard anywhere close to requiring an analysis of creative contributions.

120. Wiley, *supra* note 115, at 120 ("Using an economic perspective on innovation policy, this Part defends the notion that we should regard core portions of patent doctrine as intellectual successes worthy of imitation. Most fundamentally, patent law establishes a set of sensible and efficient *incentives* to creation. Copyright should learn this basic lesson, for a focus on sound incentives would give copyright doctrine the coherence it now lacks.").

software companies to make rational business decisions based on which aspects of their own software and their competitor's software are protectable in copyright.¹²¹ Moreover, the formulation of the test suggests a false empiricism. Even for software products aimed at mature business markets, it will be extremely difficult to determine the "cost" to society of granting the plaintiff a monopoly. More fundamentally, even if this cost could be accurately calculated, it must be balanced against the purely speculative "creative contribution" of the author, which inevitably invites judgments which are nothing more than a determination that the plaintiff's work is novel, and non-obvious, and therefore worthy of protection. This type of analysis is better left to patent law, where more precise standards exist for determining non-obviousness and where the court has the benefit of an initial expert analysis performed by the patent examiner.¹²²

D. Copyright Doctrine Properly Applied Provides Little Protection for Object-Oriented Software

While Nimmer's filtering test is closely linked to traditional copyright doctrine, it may present an unnecessary exercise in the case of object-oriented software. As a practical matter, Nimmer's filters will exclude from protection nearly every element that makes a particular program object-oriented in design. More important, these elements are precisely the elements which reveal software's "behavioral" rather than "textual" nature and which render object-oriented programs generally unsuitable for copyright protection.

While most courts recognize that computer programs are utilitarian articles, most infringement cases require the court to analyze only the textual representation of the program's structure, sequence, and organization as embodied in source code. As a result, most courts focus on the textual embodiment of software and quickly lose sight of the behavioral nature of software. This distinction was first recognized in *Computer Associates International v. Altai, Inc.*,¹²³ in which the district court adopted the findings of a court-appointed expert who explained that:

a computer program must be viewed both as text and as behavior.

The text perspective focuses upon the object code and source code A computer program, however, is more than a collection of zeros and ones. When properly loaded into a computer and

121. Beutel, *supra* note 84, at 27 (noting that "[j]ust as the application of the antitrust 'rule of reason' has taken years of dissection and analysis to take form, so too would the eventual parameters of software copyright under the policy-balancing approach set forth in the Reback/Hayes Abstractions Test").

122. Critics of software patents often question the competence of software examiners in analyzing software issues. This problem is discussed further *infra* Part V.

123. 775 F. Supp. 544 (E.D.N.Y. 1991), *aff'd in relevant part*, 782 F.2d 693 (2d Cir. 1992).

provided with appropriate input from, for example, the keyboard, the program behaves. In a word processing program, for example, text can be deleted, blocks of text can be moved, formatting of documents can be changed; all sorts of operations can be instituted; and these can only be described as behavior.¹²⁴

While the court used this analysis primarily to criticize *Whelan* for failing to distinguish between the static, textual view of the program and the dynamic, behavioral view,¹²⁵ the court also recognized that the behavioral aspect of software creates a much more fundamental problem when viewed against the statutory limits imposed by § 102(b):¹²⁶

Going beyond Dr. Davis' analysis, the court notes a possible statutory difficulty that arises when we recognize, as we must, that a computer program "behaves." . . . Since the behavior aspect of a computer program falls within the statutory terms "process", "system", and "method of operation", it may be excluded by statute from copyright protection. . . . Fortunately, this court need not wrestle with that possible development in the law of intellectual property, because CA's rights in this case are fully protected by viewing the ADAPTER program as text.¹²⁷

Although the *Computer Associates* court did not have to resolve this question, courts dealing with object-oriented software must address the behavioral nature of software. Courts will be confronted with this problem from three different angles. First, if courts approach object-oriented programs as they would approach programs written under the traditional model, they will find that the structure, sequence, and organization of the source code tell us little about the inheritance relationships and class structures in the program, which may be the part of the program that the plaintiff most wants to protect. Moreover, the closer the plaintiff adhered to the object-oriented model in the program's creation, the more pronounced this phenomenon will be. In fact, since programs that make good use of polymorphism and dynamic binding must include source code that is highly generalized, the source code behind the best written programs will tell us the least about the objects in the program.

Faced with this problem, the court will then have to focus on the specific class definitions used to specify inheritance relationships, messages accepted, and method implementations. However, while these definitions are expressed in the English words used by a particular programming language, they are simply a shorthand description of a

124. *Id.* at 559.

125. *Id.* at 560.

126. 17 U.S.C. § 102(b) provides: "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work."

127. *Computer Assocs.*, 775 F. Supp. at 560.

highly specific system. When we define the class, "rectangle" as a sub-class of the class "parallelogram," there is nothing expressive, in the copyright sense, about that definition. The term "sub-class" is a shorthand instruction that tells the compiler "whenever you see a rectangle, have it behave just like a parallelogram, except when you receive a message which has been overridden in the definition of the rectangle class, then use a different behavior."

Third, the court will have to approach the program by examining the high-level relationships among different classes and objects because the textual descriptions of a particular class cannot be protected,. In fact, the court may well be tempted to examine substantial similarity by asking the parties to create a semantic data model of each program, on the theory that if the semantic data models are substantially similar, the programs must be substantially similar. However, this approach effectively creates copyright protection for semantic data models themselves, a result which cannot be justified under fundamental copyright principles. True, a programmer who draws a semantic data model can claim a copyright in the pictorial representation that the programmer used to express the model; that programmer can prevent others from copying the *picture*. However, the copyright in the picture cannot be used to indirectly grant protection over the model itself, a result which follows directly from *Baker v. Selden*:¹²⁸

The copyright of a work on mathematical science cannot give to the author an exclusive right to the methods of operation which he propounds, or to the diagrams which he employs to explain them, so as to prevent an engineer from using them whenever occasion requires.¹²⁹

At a more fundamental level, the semantic data model cannot be used to determine infringement because it is simply a list of the constituent elements of a particular system. Copyright protection cannot be used to provide a monopoly over these elements, a point which was

128. 101 U.S. 99 (1879) (holding that copyright can reside in a particular explanation of a system, but not in the system itself). *Baker* is generally regarded as the inspiration for § 102(b). See Amicus Curiae Brief of Copyright Law Professors at 5, *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 799 F. Supp. 203 (D. Mass. 1992) (No. 90-1162-K) [hereinafter Copyright Professors' Amicus Brief] ("It is to cases such as *Baker v. Selden* and its progeny that courts should look in interpreting section 102(b) and its exclusion of systems and methods from the scope of copyright protection available to works of authorship." (citation omitted)).

129. *Baker*, 101 U.S. at 103; see also Copyright Professors' Amicus Brief, *supra* note 128, at 6 n.3 ("[T]he [Baker] Court pointed out that in most instances, useful arts were embodied in wood, metal, or stone, and what had given plausibility to Selden's claim was that his useful art was embodied in a writing. Nevertheless, the Court stated 'the principle is the same in all. The description of the art in a book, though entitled to the benefit of copyright, lays no foundation for an exclusive claim to the art itself.' " (quoting *Baker*, 101 U.S. at 105)).

recently restated in an amicus curiae brief submitted by eleven well-respected copyright professors in *Lotus v. Borland*:¹³⁰

It is in the nature of a method or system to have constituent elements, some of which may be quite detailed in character. In the "Shorthand cases," courts will decline to extend copyright protection not only to the set of abstract rules that a shorthand system developer might have devised for condensing words or phrases, but also to the vocabulary resulting from the implementation of these rules. Both are constituent elements of the system which copyright law will not protect.¹³¹

This analysis can be directly applied to a semantic data model. For example, in the traffic light program, the semantic data model tells us "a clock is a part of a controller, and a controller reads from a sensor which can be either a pressure sensor or a trip sensor." This semantic data model equally describes the real-world physical system and the system for modeling that behavior on a computer. Just as the traffic light and controller are constituent elements in the real-world traffic intersection, the representations of those entities as objects and classes are constituent elements of a *system* for modeling the behavior of a traffic intersection on a computer.

Finally, the semantic data model is exactly what its title implies, an attempt to explain a detailed system in words and pictures. The plaintiff presenting a semantic data model as the basis for proving substantial similarity is not arguing that the defendant used the same words and pictures to depict the system, but rather that the defendant used the same *system* of classes and inheritance relationships in writing the allegedly infringing program. As soon as the plaintiff presents a semantic data model as the basis for infringement, the court must recognize that the plaintiff is seeking protection for the constituent elements of a particular object-oriented system, a right which has no basis in copyright law.

The preceding analysis shows that copyright law does not protect the high-level relationships among objects. The fact that these relationships may represent the bulk of the programmer's effort and innovation during design is irrelevant in determining the scope of protection under copyright doctrine.¹³² If protection for such behavioral elements in object-oriented software is available, it can only be achieved through the patent system.¹³³

130. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 799 F. Supp. 203 (D. Mass. 1992) (granting partial summary judgment).

131. Copyright Professors' Amicus Brief, *supra* note 128, at 7.

132. See *supra* note 81.

133. See *Baker*, 101 U.S. at 105 ("The description of the art in a book, though entitled to the benefit of copyright, lays no foundation for an exclusive claim to the art itself. The object of the one is explanation; the object of the other use. The former may be secured by copyright. The latter can only be secured, if it can be secured at all, by letters-patent.");

V. PATENT PROTECTION FOR OBJECT-ORIENTED SOFTWARE

In some ways, the patentability of object-oriented software is easier to analyze than the patentability of traditional software. In cases involving traditional software, the primary question has been whether the software recites "a mathematical algorithm."¹³⁴ If it does, then the software is not patentable; otherwise the software is patentable subject matter.¹³⁵ Indeed, this analysis would still apply to a patent claim that was drawn to the low-level internal implementation of a specific method in an object-oriented program, since that portion of the program operates on the same principles as traditional software. In that case, the court would also have the benefit of examining the claim in light of a substantial body of critical commentary that has been written on the patentability of traditional software.¹³⁶ The more interesting question is what higher-level elements of the object-oriented model could qualify as patentable subject matter.

A. Patentable Subject Matter

The most promising candidate for protection is a patent claim drawn to a semantic data model. In fact, a purely textual description of a semantic data model would read very much like a standard apparatus claim. In the case of the traffic light example, we could construct a patent

Computer Assocs. Int'l v. Altai, Inc., 775 F. Supp. 544, 560 (E.D.N.Y. 1991) (noting in the context of the problems raised by the behavioral aspects of software, "indeed, it has been suggested that computer software is better protected by patent law than by copyright law"), *aff'd in relevant part*, 982 F.2d 693 (2d Cir. 1992).

134. Gottschalk v. Benson, 409 U.S. 63 (1972).

135. The Federal Circuit and its predecessor courts have devised a two-part test, the "Freeman-Walter" test, to determine whether a particular software claim is drawn to patentable subject matter. In the first step, the court must determine whether the claim directly or indirectly recites a mathematical algorithm. If it does not, then the claim is drawn to patentable subject matter. However, even if the claim does recite a mathematical algorithm, it may still be patentable if the claim "implement[s] the algorithm in a specific manner to define structural relationships between the elements of the claim in the case of apparatus claims, or limit or refine physical process steps in the case of process or method claims." *In re Walter*, 618 F.2d 758, 767 (C.C.P.A. 1980). See generally *In re Iwahashi*, 888 F.2d 1370 (Fed. Cir. 1989); *In re Pardo*, 684 F.2d 912 (C.C.P.A. 1982); *In re Abele*, 684 F.2d 902 (C.C.P.A. 1982); *In re Freeman*, 573 F.2d 1237 (C.C.P.A. 1978); *PTO Report On Patentable Subject Matter: Mathematical Algorithms and Computer Programs*, 38 Pat. Trademark & Copyright J. (BNA) 563 (1989) [hereinafter PTO Report].

136. See, e.g., Donald S. Chisum, *The Patentability of Algorithms*, 47 U. PITL. L. REV. 959 (1986); Pamela Samuelson, *Benson Revisited: The Case Against Patent Protection for Algorithms and Other Computer Program-Related Inventions*, 39 EMORY L.J. 1025 (1990); Randall M. Whitmeyer, Comment, *A Plea for Due Processes: Defining the Proper Scope of Patent Protection for Computer Software*, 85 NW. U. L. REV. 1103 (1991) [hereinafter Comment, *A Plea for Due Processes*].

claim for a real-life intersection control system that read something like this:

A traffic control apparatus consisting of:

a trip sensing means and a pressure sensing means, and a controller device which is operably connected to receive signals from said sensing means, and operably connected to send signals to a sequential display of different colored lights.

In the case of a real-world traffic control system, this claim would certainly recite patentable subject matter. In the case of object-oriented software, this claim is a close description of our semantic data model. Of course, in the case of software, the "trip sensing" means refers not to a physical object but to a location in the computer's memory that is designed to model the behavior of the real-world "trip sensing" means. While the case law on this issue is somewhat confused, a strong case can be made for holding that the above claim should be patentable subject matter whether it refers to the computer model of a traffic intersection or the physical apparatus used in the real world.

1. EXISTENCE OF MATHEMATICAL ALGORITHM

The initial inquiry for computer program related inventions focuses on the existence or absence of a mathematical algorithm.¹³⁷ If the claimed invention is drawn at the level of the semantic data model, no mathematical formulas will appear in the claim. Because the object-oriented model emphasizes encapsulation of data and procedures, the implementation of simple mathematical formulas should be hidden in the internal methods of each class and is generally invisible in the semantic data model.¹³⁸ At this point, the fact that the semantic data model may still embody a "non-mathematical" algorithm, in the broad sense, does not disqualify it from patent protection.¹³⁹

137. *Iwahashi*, 888 F.2d at 1374 ("[T]he proscription against patenting has been limited to *mathematical* algorithms and abstract *mathematical* formulae which, like the laws of nature, are not patentable subject matter.") (emphasis in original); *PTO Report, supra* note 130, at 570 ("The major (and perhaps only) exception in the area of computer processes is the mathematical algorithm . . . If a computer process claim does not contain a mathematical algorithm in the Benson sense, the second step of the Freeman-Walter-Abele test is not reached, and the claimed subject matter will usually be statutory.").

138. One might question whether the specification describing a semantic data model would be sufficiently enabling under 35 U.S.C. § 112. In most cases, the mathematical formulas necessary to construct a working program will be obvious to those skilled in the art. In those cases where the implementation is not obvious, the PTO could require the applicant to disclose those formulas in the specification, perhaps as part of the best mode requirement.

139. All apparatus claims could be considered to follow an algorithm in the broad sense of the term. See *Iwahashi*, 888 F.2d at 1375 ("[T]he fact that the apparatus operates according to an algorithm does not make it nonstatutory.").

The Federal Circuit has been quick to grant claims in which a single system of physical elements and a computer program are drafted as a single claim.¹⁴⁰ In such cases, “[t]he claim as a whole certainly defines [an] apparatus in the form of a combination of interrelated means and we cannot discern any logical reason why it should not be deemed statutory subject matter . . .”¹⁴¹ For many claims drawn to object-oriented programs, a strong argument can be made that the relationships between objects act much like the interaction between physical elements of a real-world apparatus in which different operational “means” send signals to one another and respond accordingly. In fact, as illustrated above by the sample claim for a traffic light control system, a single claim could equally describe the semantic data model or the real-world system itself.¹⁴² The close identity between the description of a real-world system and the object-oriented program which models that system reinforces the argument that a claim based on the object-oriented program presents the same statutory subject matter as a conventional claim for the physical system.

Finally, while software that models or controls real world objects presents the best candidate for patentable subject matter, protection may also be available for object libraries which have no real-world counterparts, such as object-oriented graphical user interfaces and database systems. In these cases, the software still represents the interactions of various “means” designed to control the internal workings of a general purpose computer. The Federal Circuit has already

140. See *id.* (auto-correlation unit for use in pattern recognition); *In re Abele*, 684 F.2d 902 (C.C.P.A. 1982) (software program for improved CAT-scan process); *In re Taner*, 681 F.2d 787 (C.C.P.A. 1982) (software which improved seismic exploration by translating spherical seismic waves into plane or cylindrical waves); *In re Freeman*, 573 F.2d 1237 (C.C.P.A. 1978) (software for controlling conventional phototypesetter).

141. *Iwahashi*, 888 F.2d at 1375.

142. The fact that such a claim could be drafted also implies that a single claim would grant the inventor a monopoly over both the real-world physical system and the object-oriented model of that system, a result which undoubtedly raises alarms in certain circles. However, three considerations mitigate the danger of this result. First, the single claim will have to withstand prior art from both the computer science field and field relating to the real-world physical system. See *infra* Section V.B. Few claims will be non-obvious when tested against such a wide range of prior art. Second, the claim grants a monopoly only to the extent that the invention is enabled by the specification under the standards in 35 U.S.C. § 112. In many cases, the inventor may be able to describe how to write the object-oriented program but will be unable to explain how to actually build some of the elements in the physical system. For example, while we can easily model a “trip sensor” in a program, it may be much more difficult to build one that works consistently when embedded in a roadway. Particularly since the vast majority of claims will have to be written in “means-plus-function” form, the specification will sharply limit the actual scope of the monopoly granted by the single claim. Finally, if the inventor has presented a single claim which is truly non-obvious and which enabled both the computer and physical versions of the systems, then the inventor has really *invented* both systems and should be entitled to protection over both.

recognized the patentability of pure software claims which direct the way the computer manages data internally.¹⁴³

2. MENTAL STEPS

The mental steps doctrine was historically used to deny patent protection for process claims involving simple measurements, calculations, and interpretations of data that could just as easily be performed by a human using paper and pencil.¹⁴⁴ However, the C.C.P.A. may have broadened the doctrine in 1982 when it denied patent protection to an expert system for neurological diagnosis on the basis that "their invention is concerned with replacing, in part, the thinking processes of a neurologist with a computer."¹⁴⁵ Moreover, the Federal Circuit has implicitly used the doctrine to invalidate a claim for an invention designed to determine whether any complex system is in a normal or abnormal state.¹⁴⁶

In its broadest form, the mental steps doctrine would deny patent protection to any expert system. Since object-oriented development emphasizes approaching the project from the perspective of an expert in the problem domain, semantic data models may mirror the mental process that an expert in that field would use to solve problems. However, many object-oriented programs will be able to survive the mental steps doctrine for two reasons. First, the software claims recently invalidated under the mental steps doctrine involved the calculation of a discrete result and in general seemed close to a simple process of mental calculations.¹⁴⁷ In contrast, many object-oriented programs will model the operation of systems with continuous behavior that produce no discrete "answer" to a problem. For example, the "QuadWorld" program solves no specific problem, but rather provides a system for drawing and manipulating a variety of shapes. It is difficult to conceive of QuadWorld

143. See *In re Pardo*, 684 F.2d 912, 913 (C.C.P.A. 1982) (invention which "converts a computer from a sequential processor . . . to a processor which is not dependent on the order in which it receives program steps"); *In re Bradley*, 600 F.2d 807 (C.C.P.A. 1979) ("firmware" designed to improve performance of multi-tasking), *aff'd sub nom.* *Diamond v. Bradley*, 450 U.S. 381 (1981).

144. Samuelson, *supra* note 136, at 1034-38.

145. *In re Meyer*, 688 F.2d 789, 795 (C.C.P.A. 1982).

146. *In re Grams*, 888 F.2d 835, 840 (Fed. Cir. 1989) (analogizing to *Meyer* and finding the existence of an algorithm in part because "the objective [in *Meyer*] of identifying malfunction is similar to the objective here of identifying abnormality").

147. In fact, the Federal Circuit never explicitly mentioned the mental steps doctrine but rather denied the claims because "[f]rom the specification and the claim, it is clear to us that applicants are, in essence, claiming the mathematical algorithm, which they cannot do . . ." *Grams*, 888 F.2d at 840; see also Comment, *A Plea for Due Processes*, *supra* note 136, at 1122.

as a series of discrete mental steps that could be performed to achieve the same result.

Second, the C.C.P.A. did not extend the doctrine to cases where the system could theoretically be performed as a series of mental steps but, as a practical matter, would be too complex to implement with pen and paper.¹⁴⁸ Similarly, many object-oriented programs will reflect complex relationships between elements of an extremely large system. At the level of "programming in the colossal", few object-oriented systems can be reduced to a series of human mental process steps. As a result, the mental steps doctrine should not provide an independent bar to patentability.

3. METHOD OF DOING BUSINESS

If the method of doing business limitation were applied seriously, it would exclude from protection any computer program that implemented familiar business systems, a category that could ensnare many object-oriented database systems. However, the doctrine is of questionable validity¹⁴⁹ and has only been weakly applied in computer cases. For example, in *Paine, Webber v. Merrill Lynch*,¹⁵⁰ a district court noted the existence of the doctrine, but dismissed it because

[t]he product of the claims of the '442 patent effectuates a highly useful business method and would be unpatentable if done by hand. The C.C.P.A., however, has made clear that if no *Benson* algorithm exists, the product of a computer program is irrelevant, and the focus of analysis should be on the operation of the program on the computer.¹⁵¹

On this basis, the court upheld a claim for a computer program that implemented Merrill Lynch's "Cash Management Account System" which allowed customers to combine brokerage, money market,

148. See *In re Toma*, 575 F.2d 872 (C.C.P.A. 1978) (allowing claims for a computer process of translating from any source language to any target language by examining a language dictionary, examining the syntax of the source and then producing a complete sentence in the target language).

149. Arthur J. Hansmann, *Method of Doing Business*, 50 J. PAT. OFF. SOC'Y 503, 504 (1968) ("Except for dicta, one can conclude that there is no basis in existing law for the rejection of claims as being directed to a 'method of doing business.'"); David J. Meyer, Note, *Paine, Webber, Jackson and Curtis, Inc. v. Merrill Lynch, Pierce, Fenner & Smith: Methods of Doing Business Held Patentable Because Implemented on a Computer*, 5 COMPUTER L.J. 101, 103-04 n.13 (1984) (reviewing the cases cited in *Merrill Lynch* and concluding that "examination of these cases reveals that the issue of patentable subject matter was never actually decided. Rather, the patent claims were held invalid for 'lack of invention.' . . . The issue of the patentability of a method of doing business was discussed only in dictum . . ."); Comment, *A Plea for Due Processes*, *supra* note 136, at 1119 ("[I]t is unclear whether this doctrine ever really existed . . .").

150. *Paine, Webber, Jackson & Curtis, Inc. v. Merrill, Lynch, Pierce, Fenner & Smith, Inc.*, 564 F. Supp. 1358 (D. Del. 1983).

151. *Id.* at 1369.

checking, and credit cards into one integrated account. Similarly, the C.C.P.A. has allowed claims for a program to control the optimal operation of plants, such as oil refineries, at multiple locations,¹⁵² and for a program that produced architectural specifications and project control instructions.¹⁵³ Finally, in the only case to invalidate a business methods program, the C.C.P.A. did not even mention the doctrine but declared the claim invalid on the basis that the claim recited and preempted a specific algorithm.¹⁵⁴

These cases indicate that the doctrine may have little effect on inventions related to computer programs in general. Moreover, object-oriented programs will be affected even less by the doctrine than traditional software. As discussed in the next Section, if the program merely implements a familiar business method, then prior art relevant to general business methods and practices will invalidate the claim on § 103 grounds.¹⁵⁵ Thus, courts evaluating such claims will never have to reach the business methods question, since it will be easier to resolve the issue on § 103 grounds and confine the analysis of patentable subject matter to determining whether the claim recites a mathematical algorithm.

B. Non-Obviousness and the Relevant Prior Art

Even though semantic data models qualify as patentable subject matter, few patents will actually be issued because few will pass the non-obviousness requirements of 35 U.S.C. § 103. While critics of software patents have claimed that the Patent Office lacks the expertise or the database files to accurately evaluate prior art for software patents,¹⁵⁶ this problem is considerably less severe in the case of object-oriented software. Since a standard for determining the relevant fields of prior art is "whether it deals with a problem similar to that being addressed by the

152. *In re Deutsch*, 553 F.2d 689 (C.C.P.A. 1977).

153. *In re Phillips*, 608 F.2d 879 (C.C.P.A. 1979).

154. *In re Maucorps*, 609 F.2d 481, 486 (C.C.P.A. 1979) (claim for a program that determined the optimal organization of a sales force).

155. See 35 U.S.C. § 103 (1988).

156. See, e.g., Brian Kahin, *The Software Patent Crisis*, TECH. REV., April 1990, at 53, 55.

The search [for software prior art] is extraordinarily difficult because the field's printed literature is thin and unorganized. Software documents its own design, in contrast to physical processes, which require written documentation. Also, software is usually distributed without source code under licenses that forbid reverse engineering. This may amount to suppressing or concealing the invention and therefore prevent the program from qualifying as prior art.... Many programmers suspect that patent examiners lack knowledge of the field, especially since the Patent Office does not accept computer science as a qualifying degree for patent practice

....

inventor,"¹⁵⁷ the examiner will have to search for references not only in the computer science area, but also in the literature relating to the real-world problem being addressed by the software. Because the programmer first approached the project in step 1 of our object-oriented design model by learning as much as possible from experts in the problem domain itself,¹⁵⁸ the examiner will also have to use the full range of literature in the problem domain as prior art. This search may prove fairly easy since the applicant's duty of candor will require the programmer to unilaterally disclose to the patent office all the sources used in developing the project.¹⁵⁹

The prior art problem will also be less significant because the examiner will be better equipped to determine non-obviousness.¹⁶⁰ For example, in the traffic light problem, the examiner will compare the semantic data model to literature that describes the operation of the physical entities that operate traffic lights in the real world. In general, this literature will reflect principles of electrical and mechanical engineering that are more familiar to most examiners than principles of computer science. The examiner can quickly determine whether the software is merely a straightforward model of the physical system and therefore obvious to the hypothetical person of ordinary skill in the art of object-oriented design. In many cases, the claim for the semantic data model will read almost exactly like a claim for a well-documented physical system and thus will quickly appear obvious to the examiner.

Under this analysis, the elements of object-oriented software that most embody object-oriented design are patentable subject matter. As a practical matter, however, only the small percentage of semantic data models that are truly non-obvious, in the face of an extremely broad range of relevant prior art, will be granted patents.

157. Union Carbide Corp. v. American Can Co., 724 F.2d 1567, 1572 (Fed. Cir. 1984) ("The determination that a reference is from a nonanalogous art is therefore two-fold. First, we decide if the reference is within the field of the inventor's endeavor. If it is not, we proceed to determine whether the reference is reasonably pertinent to the particular problem with which the inventor was involved." (quoting *In re Wood*, 599 F.2d 1032, 1036 (C.C.P.A. 1979)); see also *Bott v. Four Star Corp.*, 218 U.S.P.Q. (BNA) 358, 368 (E.D. Mich. 1983) ("The test for relevant or analogous prior art is 'similarity of element, problems, and purposes.' 'Analogous art is that field of art which a person of ordinary skill in the art would have been apt to refer in attempting to solve the problem solved by a proposed invention.'" (citations omitted)), *aff'd*, 732 F.2d 168 (Fed. Cir. 1984).

158. See *supra* Section III.C.1.

159. 37 C.F.R. § 1.56 (1992).

160. An often-heard complaint against traditional software patents is that examiners untrained in computer science "naturally have a lower standard in determining the hypothetical 'person having ordinary skill in the art,' and are thus more apt to grant patents for obvious processes." Kahin, *supra* note 156, at 55.

VI. IMPLICATIONS FOR INNOVATION POLICY

This article has presented the most important concepts of object-oriented programming and discussed one model of the object-oriented design process. A thorough understanding of object-oriented design shows that copyright protection cannot be justified for the elements of the software that make it object-oriented. In addition, one product of the object-oriented design process, the semantic data model, may be used to draft a patentable claim, if it is sufficiently innovative relative to an extremely broad range of prior art.

These conclusions are based on the application of existing legal doctrines, rather than on an analysis of which legal regime provides better protection as a matter of innovation policy. While the industry is still too new to perform any worthwhile empirical analysis, several intuitive observations about the advantages of each legal regime suggest that limited patent protection provides reasonable incentives for innovation.

Legal protection for software designed according to object-oriented principles is important only at the margins of innovation. Assuming that verbatim copying and distribution can be prevented,¹⁶¹ most companies that write commercial software are not motivated by the promise of broad legal protection for their products but by economic returns that result from being first in the market or being the first to introduce programs with new features into an existing market. While there is no method to test this hypothesis, the nature of object-oriented development itself suggests that companies will be motivated to innovate without extensive legal protection. Object-oriented programming will be adopted because it allows complex programs to be created with fewer errors, encourages the use of existing software components, leads to easier software maintenance, and eases the process of improving the software in subsequent versions. As a result, companies utilizing object-oriented techniques will produce better products and face lower development costs than companies using traditional techniques. This improvement in software "manufacturing" will provide most of the incentive necessary to stimulate innovations in object-oriented software. Indeed, these incentives are evident in the growth of the Object Management Group

161. Of course, outright piracy destroys all incentives for innovation. However, verbatim copying and distribution is easy to prohibit, at least within the United States, through traditional copyright protection for literal source and object code. As in the prior discussion of legal doctrines, this discussion is concerned with the more interesting problems presented by copyright and patent protection for the object-oriented elements of software.

(OMG), a "technology endorsement group" whose membership includes almost 200 leading software companies.¹⁶²

Against this background of strong innovation, legal protection can provide some additional incentives. Analyzing legal protection from the standpoint of innovation policy suggests that copyright protection would be inappropriate for object-oriented software¹⁶³ while patent protection may be justified for highly innovative programs.

Copyright protection has been popular for traditional software primarily because it is easy to obtain, it provides strong protection against literal and non-literal copying, and it still allows for a defense of independent creation. However, copyright protection presents serious problems when applied to the aspects of a particular program that make it object-oriented. When applying for a copyright, the author need not make any attempt to define the scope of the copyright being claimed. The author simply submits a copy of the source code or object code of the program with the registration form, and copyright protection instantly attaches. As the Second Circuit recently noted, "we think that copyright registration—with its indiscriminating availability—is not ideally suited to deal with the highly dynamic technology of computer science."¹⁶⁴

The net result of this process is that the scope of any particular copyright is not defined until litigation occurs and, even then, is only defined relative to the particular program accused of infringement. As a result, business competitors cannot legitimately plan future products since they cannot be sure of how to "design around" an existing copyright. This problem is most acute when protection is claimed for the non-literal elements of a program, such as the semantic data model. While the independent creation defense provides some protection for competing software developers, the high mobility of software engineers¹⁶⁵ combined with the questionable legal status of reverse engineering techniques¹⁶⁶ may render any protection highly illusory.

162. Object Management Group Purpose and Definition Statement (1992) (on file with author). OMG was originally formed in April 1989 by Data General, Hewlett-Packard, Sun, Canon, American Airlines, Unisys, Philips, Prime, Gold Hill, Soft-Switch, and 3-Com. Major software players such as AT&T, Digital, NCR, Borland, Microsoft, and IBM have subsequently joined.

163. Remember that copyright protection would still be available to prevent verbatim copying of source and object code, thus supplying the necessary prerequisite to innovation discussed earlier.

164. Computer Assocs. Int'l v. Altai, Inc., 982 F.2d 693, 712 (2d Cir. 1992).

165. Absent costly clean room development procedures, the accused infringer may find it extremely difficult to prove that every engineer involved in the project was completely ignorant of the plaintiff's copyrighted program.

166. Once reverse engineering has been used, the defendant can no longer claim independent creation. Even if the final program is non-infringing, the process of reverse engineering itself could constitute copyright infringement. At the present time, reverse engineering is a risky business strategy. However, the judicial attitude toward reverse

Moreover, the basic fit between copyright protection and continuing innovation must be questioned. Even though the Supreme Court's decision in *Feist* breathed new life into the originality requirement, copyright makes little distinction between the protection afforded to trivial innovations and the protection given to major innovations. If courts adopt broad *Whelan*-style protection for semantic data models, then fairly trivial applications of object-oriented principles are likely to be granted strong protection.¹⁶⁷ This protection will not be offset by the benefits of disclosure of the innovation since commercial programs are distributed in object code form only and copyright registration can be obtained without disclosing the source code to the public. Since any object-oriented innovations occur at the source code level, the public gains no new knowledge from the grant of copyright. At most, the public gains access to a commercial product that might not have been created without the promise of copyright. On balance, copyright protection seems likely to stifle competition and discourage continuing innovation.

In contrast, while the patent examination process makes patent protection more difficult and costly to obtain, this process also addresses the primary deficiencies in copyright protection. First, the inventor must specifically define the scope of the software invention through technical claim language. This language is likely to be narrowed during the examination process in order to overcome prior art rejections. As a result, only highly innovative programs will be granted protection, and the scope of that protection will be sharply limited by prior art. Business competitors can then rationally plan competing products by performing patent searches and then determining how to design around existing patents.¹⁶⁸ If designing around an existing patent is not feasible, the precise definition provided by the claim language will make it easier for the parties to estimate the value of the patent and negotiate licenses. Finally, since patent protection is given only as the *quid pro quo* for full disclosure of the innovation, the costs of protection are offset by dissemination of the new technological knowledge behind the invention as well as by dissemination of the commercial product itself.

engineering may be changing,. Two appellate courts have recently applied the "fair use" doctrine to allow reverse engineering in certain contexts. See *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510, 1520 (9th Cir. 1992), amended, 1993 U.S. App. Lexis 78 (9th Cir. 1993); *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832 (Fed. Cir. 1992).

167. See, e.g., *Computer Assocs.*, 982 F.2d at 712 (noting that "serious students of the industry have been highly critical of the sweeping scope of copyright protection engendered by the *Whelan* rule, 'in that it enables first comers to "lock up" basic programming techniques as implemented in programs to perform specific tasks'" (quoting Menell, *supra* note 3, at 1087; citations omitted)).

168. Admittedly, business competitors still face some risk since patent applications are kept secret during the examination process.

Patent protection does present several drawbacks. First, the costs of obtaining protection or defending a potential infringement suit may deter some smaller companies from innovation. Second, the lack of an independent creation defense sharply increases the societal costs of protection by stifling innovation that would have occurred in the absence of patent protection. Finally, the seventeen-year term for the patent monopoly is excessively long given the short product cycles for most software projects. Nonetheless, if patent examiners utilize a broad range of relevant prior art both to reject applications that are not highly innovative and to limit the scope of patents actually granted, then patent protection is a more effective incentive for innovation than copyright protection.

COMMENT

THE EXPERIMENTAL USE EXCEPTION TO INFRINGEMENT APPLIED TO FEDERALLY FUNDED INVENTIONS

SUZANNE T. MICHEL [†]

Table of Contents

I.	INTRODUCTION	369
II.	BACKGROUND OF EXPERIMENTAL USE EXCEPTION	371
	A. Creation and Early Development.....	371
	B. The Federal Circuit	374
III.	PROBLEMS WITH COMMON LAW DOCTRINE	376
	A. Uncertainty for Universities and Federal Laboratories.....	376
	B. Foreclosure of New Inventions When A Basic Technique is Patented	386
	C. Polymerase Chain Reaction Example	387
IV.	PAST PROPOSALS AND CRITIQUE.....	388
	A. Proposals For a Broad Exception.....	388
	B. Incentives of the Patent System.....	391
	C. A Critique of the Broad Exception	394
V.	NEW PROPOSALS.....	397
	A. Non-profit Researchers Allowed Broad Exception	397
	B. Government-Funded Inventions Subject to a Broad Exception.....	400
VI.	CONCLUSION	409

I. INTRODUCTION

With one minor exception¹ the patent statutes do not suggest any instance in which use of a patented invention is not infringement. According to 35 U.S.C. § 154, "[e]very patent shall contain . . . a grant to the patentee . . . of the right to exclude others from making, using or

© 1993 Suzanne T. Michel.

[†] J.D. candidate 1993, School of Law (Boalt Hall), University of California at Berkeley; Ph.D. 1989, Yale University; B.S. 1984, Northwestern University.

1. 35 U.S.C. § 271(e) provides that it is not an act of infringement to make, use, or sell a patented invention for purposes reasonably related to obtaining FDA approval of drugs.

selling the invention throughout the United States." Section 271(a) provides that "whoever without authority makes, uses or sells any patented invention . . . infringes the patent."

In spite of the seemingly unyielding dictate of the statutes, courts have recognized experimental use as an exception to infringement. Use of a patented invention "for the mere purpose of philosophical experimentation, or to ascertain the verity and exactness of the specification" is exempt from infringement.² While it is well settled that a patented invention may be made and used to test the verity and exactness of the specification, the scope of the "philosophical experimentation" prong of the exception is much less clear. The Federal Circuit has called this prong "truly narrow."³ To be deemed philosophical experimentation, the experiment must be "for amusement, to satisfy idle curiosity, or for strictly philosophical inquiry."⁴ The exception does not "allow a violation of the patent laws in the guise of 'scientific inquiry,' when that inquiry has definite, cognizable and not insubstantial commercial purposes."⁵ Part II of this Comment describes the history and current scope of the experimental use exception.

In view of this narrow interpretation of the "philosophical experiment" prong of the experimental use exception, several commentators have called for a legislative broadening of the exception to encompass all activity short of commercialization.⁶ A House bill, the Research, Experimentation and Competitiveness Act of 1990, also proposed broadening the exception.⁷

Those proposing the broad exception point to two key problems which they contend the broad exception would either clarify or solve. First, it is unclear whether university and other non-profit research done under contract with industry or with a purpose to patent the results is "strictly for philosophical inquiry." The uncertain limits of the doctrine might chill research or lead to litigation. Second, when a patent owner controls important information, that control might prevent a subsequent researcher from building on the information in a way that benefits society. The broad exception would allow subsequent research on patented inventions and would clarify the position of non-profit

2. Sawin v. Guild, 21 F. Cas. 554, 555 (C.C.D. Mass. 1813) (No. 12,391).

3. Roche Prods., Inc. v. Bolar Pharmaceutical Co., 733 F.2d 858, 863 (Fed. Cir.), cert. denied, 469 U.S. 856 (1984).

4. *Id.*

5. *Id.*

6. Rebecca S. Eisenberg, *Patents and the Progress of Science: Exclusive Rights and Experimental Use*, 56 U. CHI. L. REV. 1017 (1989); Ned A. Israelsen, *Making, Using, Selling Without Infringing: An examination of 35 U.S.C. Section 271(e) and the Experimental Use Exception to Patent Infringement*, 16 AM. INTELL. PROP. L. ASS'N Q.J. 457 (1989).

7. H.R. 5598, 101st Cong., 1st Sess. §§ 401-403 (1990).

researchers. Part III describes the conditions which caused these two problems.

The task at hand is to find the wisest limits for the exception while providing a workable solution to the problems of foreclosed research and the uncertain position of non-profit researchers. Any proposal must take into account the economics and incentives of the patent system. Part IV critiques the wisdom of the proposals for a generally applicable broad exception. Part IV also argues that a generally applicable broad experimental use exception weakens the incentives to invent, to develop and to disclose provided by the patent system to too great an extent when applied to patents resulting from private research efforts.

Instead, this Comment proposes in Part V that the experimental use exception (extending up to commercialization) be made applicable only in the special circumstances in which its harm to patent incentives is minimal compared to the resulting benefits. First, university and other non-profit researchers should be allowed the advantage of the broad exception. This first proposal clarifies the position of non-profit researchers with minimal harm to the patent holder. Second, any party should be allowed to use a patented, federally funded invention in research and development. This second proposal provides a number of benefits without the disincentives which result when a broad experimental use exception is applied to privately funded patents. For instance, federally funded inventions will not foreclose subsequent research, but federal grantees will not lose their incentive to invent and disclose because those incentives come from outside the patent system.

II. BACKGROUND OF EXPERIMENTAL USE EXCEPTION

Understanding how the critique and proposals presented by this Comment fit into the framework of the patent laws first requires understanding the judicially created experimental use exception.

A. Creation and Early Development

The experimental use doctrine as a defense to patent infringement originated in 1813 in *Whittemore v. Cutter*, an opinion written by Justice Story while sitting on the Massachusetts Circuit Court.⁸ The defendant in that case challenged a jury instruction that "the making of a machine fit for use, and with a design to use it for profit, was an infringement of the patent right."⁹ Justice Story approved the instruction on the grounds that "it could never have been the intention of the legislature to punish a man, who constructed such a machine merely for philosophical experiment, or

8. *Whittemore v. Cutter*, 29 F. Cas. 1120 (C.C.D. Mass. 1813) (No. 17,600).

9. *Id.* at 1121.

for the purpose of ascertaining the sufficiency of the machine to produce its described effects."¹⁰

Justice Story referred to this exception again in *Sawin v. Guild*.¹¹ In holding that the defendant's use of patented machines constituted patent infringement, he noted that the machines had been used for profit rather than "for the mere purpose of philosophical experimentation, or to ascertain the verity and exactness of the specification. . . . In other words, that the making must be with an intent to infringe the patent-right, and deprive the owner of the lawful rewards of his discovery."¹² Even though experimental use was not an issue in either case, meaning that the exception originated in dicta, by 1861 the law on this subject was deemed "well-settled."¹³

Very few early cases applied the experimental use doctrine created by Justice Story to excuse use of a patented invention that would otherwise constitute infringement.¹⁴ Even so, the second prong of Justice Story's test which allows activity for "ascertaining the verity and exactness of the specification" does appear to be "well settled." A party may wish to challenge a patent as invalid for not being enabling or useful and therefore must use the invention without a license to assemble proof of this invalidity. A party may also wish to test a patent before taking a license. Although there is little case law on the point, most commentators agree that this sort of activity is and should be protected by the exception.¹⁵

The scope of the "philosophical experiment" prong is much less clear. The cases that applied this prong simply concluded that the use in question was "experimental" without offering an elaboration of that term.¹⁶ The commercial character of a use or the commercial intent of a user usually forfeited the protection of the doctrine in other early cases.¹⁷ Overall, these early cases provide little guidance in setting the contours of the exception today.

Two more recent cases developed the "philosophical experiment" prong more fully, but neither found the doctrine to be applicable. In *Pitcairn v. United States*, the Court of Claims considered whether

10. *Id.*

11. 21 F. Cas. 554 (C.C.D. Mass. 1813) (No. 12,391).

12. *Id.* at 555 (citation omitted) (citing *Whittemore*).

13. *Poppenhusen v. Falke*, 19 F. Cas. 1048, 1049 (C.C.S.D.N.Y. 1861) (No. 11,279).

14. The history of the experimental use exception from its creation to its application by the Federal Circuit is described elsewhere. See Ronald D. Hantman, *Experimental Use as an Exception to Patent Infringement*, 67 J. PAT. & TRADEMARK OFF. SOC'Y 617 (1985). Accordingly, this Comment presents only a summary.

15. Eisenberg, *supra* note 6, at 1074.

16. See Israelsen, *supra* note 6, at 460 n.11.

17. See *id.* at 460 n.14.

helicopters produced under contract for the United States infringed patents that had been previously declared valid by that court.¹⁸ The court rejected the government's argument that the helicopters were purchased for testing and experimental purposes and therefore did not infringe.¹⁹ The court held that "[t]ests, demonstrations and experiments of such nature are intended uses of the infringing aircraft manufactured for the defendant and are in keeping with the legitimate business of the using agency."²⁰ The helicopters were not built solely for experimental purposes and thus were excluded from the exception.

In *Pfizer, Inc. v. International Rectifier Corp.*, a federal district court held International Rectifier (IR) in contempt of court for violating an injunction which ordered IR to cease manufacture, use and sales of doxycycline, a pharmaceutical compound patented by Pfizer.²¹ In spite of the injunction, IR had continued to manufacture doxycycline in order to conduct various tests such as bioequivalency and serum level tests.²² IR also shipped doxycycline to laboratories in and out of the United States accompanied by a notice that the compound constituted laboratory samples for experimental purposes only.²³

IR defended its activities on the grounds that they were solely experimental, and that the compound was never sold in the United States after the injunction. The court held these arguments to be "utterly without merit."²⁴ The court interpreted the history of the experimental use doctrine to suggest that "the underlying rule of permissible experimental use demands there must be no intended commercial use of the patented article, none whatsoever, if the exception is to be recognized at all."²⁵ Because IR's activities were for the purpose of competing with Pfizer after its patent expired, the court held IR in contempt. In addition, the court ordered IR to destroy all the doxycycline it possessed as well as all data it illicitly acquired regarding doxycycline.²⁶

Both *Pitcairn* and *Pfizer* make clear that when a use is consistent with the "legitimate business" of the infringer or has an ultimate commercial purpose, the use is not "philosophical experimentation" and falls outside of the exception.

18. *Pitcairn v. United States*, 547 F.2d 1106 (Ct. Cl. 1976), *cert. denied*, 434 U.S. 1051 (1978).

19. *Id.* at 1124-25.

20. *Id.* at 1125-26.

21. *Pfizer, Inc. v. International Rectifier Corp.*, 217 U.S.P.Q. (BNA) 157 (C.D. Cal. 1982).

22. These tests are required for FDA approval of generic drugs.

23. *Id.* at 158-59.

24. *Id.* at 160.

25. *Id.* at 161.

26. *Id.* at 163.

B. The Federal Circuit

In *Roche Products, Inc. v. Bolar Pharmaceutical Co.*, the only Federal Circuit²⁷ case discussing at length the scope of the experimental use doctrine, the court interpreted the doctrine narrowly.²⁸ Bolar had imported five kilograms of Roche's patented compound flurazepam hydrochloride which Roche sold as a sleeping pill, Dalmane. Bolar used the compound to conduct the bioequivalency studies required for FDA approval with an eye toward marketing a generic version of the drug when Roche's patent expired a year later. Roche argued that this use constituted infringement, but the district court held that the use of a patented drug for testing related to FDA drug approval during the last six months of the patent term was *de minimis*, experimental and noninfringing.²⁹

The Federal Circuit reversed, calling the experimental use exception "truly narrow."³⁰ The court's analysis first addressed the statute, noting that "[s]ection 271(a) prohibits, on its face, any and all uses of a patented invention," but admitted that the definition of "use" is a matter of judicial interpretation. The court cited *Pitcairn* for both the proposition that experimental use may be a defense to infringement and as setting forth the controlling law.³¹ The court quoted *Pitcairn*'s statement that "[t]ests, demonstrations, and experiments . . . [which] are in keeping with the legitimate business of the . . . [alleged infringer]" are infringements for which '[e]xperimental use is not a defense.'³²

Bolar did not come within the exception because its use was "not for amusement, to satisfy idle curiosity, or for strictly philosophical inquiry."³³ The court explained:

[U]nlicensed experiments conducted with a view to adaption of the patented invention to the experimenter's business is a violation of the rights of the patentee to exclude others from using his patented invention. . . . We cannot construe the experimental use rule so broadly as to allow a violation of the patent laws in the guise of "scientific inquiry," when that inquiry has definite, cognizable and not insubstantial commercial purposes.³⁴

27. The Federal Circuit, established in 1982, has jurisdiction over all appeals in cases "arising under" the federal patent laws. 28 U.S.C. § 1295 (1988).

28. *Roche Prods., Inc. v. Bolar Pharmaceutical Co.*, 733 F.2d 858 (Fed. Cir.), *cert. denied*, 469 U.S. 856 (1984).

29. *Roche Prods., Inc. v. Bolar Pharmaceutical Co.*, 572 F. Supp. 255 (E.D.N.Y. 1983), *rev'd*, 733 F.2d 858 (Fed. Cir.), *cert. denied*, 469 U.S. 856 (1984).

30. *Roche*, 733 F.2d at 863.

31. *Id.* at 861, 863.

32. *Id.* at 863 (quoting *Pitcairn v. United States*, 547 F.2d 1106, 1125-26 (Ct. Cl. 1976), *cert. denied*, 434 U.S. 1051 (1978)).

33. *Id.*

34. *Id.*

Nor did the court consider the use *de minimis* even though the quantity used was small, because the testing could have had a significant economic impact on Roche if Bolar released the generic drug on the market earlier than it would have absent the infringement.³⁵

1. THE OVERRULING OF ROCHE V. BOLAR

Shortly after the *Roche v. Bolar* decision, Congress passed the Drug Price Competition and Patent Term Restoration Act of 1984³⁶ which legislatively overruled that decision. That law exempts from infringement activity which is "reasonably related" to seeking FDA approval for a generic drug. The portion of the bill codified as 35 U.S.C. § 271(e)(1) states that "[i]t shall not be an act of infringement to make, use, or sell a patented invention . . . solely for purposes reasonably related to the development and submission of information under a federal law which regulates the manufacture, use or sale of drugs."

The scope of the exemption is fairly narrow. The legislative history indicates that only a limited amount of testing to establish the bioequivalency of a generic drug substitute is permitted.³⁷ Whether an activity is "reasonably related" to seeking FDA approval has been narrowly interpreted in the case law.³⁸

The legislation is interesting because it demonstrates a Congressional attitude which is willing to allow exceptions to infringement under some circumstances. The committee report states that the exemption did not substantially interfere with the rights of the patent holder because "[t]he patent holder retains the right to exclude others from commercial markets during the life of the patent."³⁹ In spite of this statement, Congress concurrently enacted a law which extended the patent grant for human drugs and other products which must undergo federal approval before marketing to compensate patentees for the time lost in which they can monopolize the market.⁴⁰ Patent owners essentially receive an extension of the patent term in exchange for their toleration of infringing use which enables a competitor to market a product as soon as the pertinent patent expires. This trade-off implies

35. *Id.* at 866.

36. 35 U.S.C. § 271(e) (1988).

37. H.R. REP. NO. 857 pt. 2, 98th Cong., 2d Sess. 8 (1984), reprinted in 1984 U.S.C.C.A.N. 2686, 2692.

38. *Scripps Clinic v. Genentech, Inc.*, 666 F. Supp. 1379, 1396 (N.D. Cal. 1987) (a multiple purpose use of a patented invention is not exempted where only one purpose is reasonably related to FDA testing). However, the Supreme Court's decision in *Eli Lilly v. Medtronics* affirms a Federal Circuit decision to extend the scope of 271(e) to include the testing of medical devices. 872 F.2d 402 (Fed. Cir. 1989), *aff'd*, 496 U.S. 661 (1990).

39. H.R. REP. NO. 857 pt. 2, *supra* note 37, at 8, reprinted in 1984 U.S.C.C.A.N. at 2692.

40. *Id.* at 14, reprinted in 1984 U.S.C.C.A.N. at 2691; 35 U.S.C. § 156 (1988).

that Congress may have conflicting views as to whether the harm to the patentee caused by the exempted experimental testing is as insubstantial as the legislative history suggests. The nature of an experimental use exception's interference with the patent right is discussed below in Section IV.C.

III. PROBLEMS WITH COMMON LAW DOCTRINE

Although the holding of *Roche* was overruled through legislation, that case is illustrative of the Federal Circuit's attitude toward the experimental use exception as a defense to infringement. The rationale of *Roche* remains the common law of experimental use in contexts other than the limited conditions of section 271(e). Given the narrow limits which that case places on the doctrine, any activity with a long-range profit motive or with any profit potential is unlikely to fall within the exception. Corporate research will nearly always be "in keeping with the legitimate business of the alleged infringer."

This narrow interpretation of the "philosophical experiment" prong of the experimental use exception engenders two related problems. First, it is unclear whether university research done under contract with industry or with a purpose to patent the results is "strictly for philosophical inquiry." Second, when a patent owner controls important information, that control might prevent a subsequent researcher from building on the information in a way that benefits society. The uncertain limits of the doctrine might chill research or lead to unnecessary litigation.

A. Uncertainty for Universities and Federal Laboratories

The extent to which use of a patented invention is permissible noninfringing experimentation when conducted by nonprofit researchers such as universities and federal labs remains unclear. Only one 1935 case, *Ruth v. Stearns-Roger Manufacturing Co.*, has addressed the issue of whether university use can be infringement. The district court in that case held that use of an infringing machine by the Colorado School of Mines was experimental and exempt from infringement because the machines were used in a laboratory and were cut up and changed from day to day. The school used the machines in furtherance of its educational purpose.⁴¹

Whether all research conducted in universities and federal laboratories today can be categorized as "philosophical experiments" is extremely problematic given the Federal Circuit's narrow interpretation

41. *Ruth v. Stearns-Roger Mfg. Co.*, 13 F. Supp. 697, 703 (D. Colo. 1935), *rev'd on other grounds*, 87 F.2d 35 (10th Cir. 1936).

of that term. To understand how university research, which would appear to epitomize "philosophical experimentation," could fall outside the exception, we must examine the trend toward patenting and licensing university research and the relationships universities have forged with industry. It is through the universities' own attempts to monopolize research results and collaborate with the commercial sector that they have potentially lost claim to the experimental use exception. Part I through Part III below describe the current landscape of industry-university, industry-federal laboratory relationships. Part IV explains why these relationships make application of the experimental use exception uncertain.

1. PATENTING AND LICENSING BY UNIVERSITIES

Prior to the 1980 and 1984 amendments to the patent laws, patents resulting from federally funded research belonged to the government, who often licensed them on a royalty-free, non-exclusive basis, although policies varied depending on the granting agency.⁴² The government had a poor record for advancing the development of its patents. For instance, in 1976, less than four percent of the twenty-eight thousand patents held by the federal government were commercially developed.⁴³

The perceived need for the 1980 and 1984 amendments was prompted in part by the concern that federally funded research was not being efficiently commercialized because a company wishing to use that research confronted "a bewildering array of 26 different sets of agency regulations governing their rights to use such research."⁴⁴ In response, the amendments created a single, uniform national policy. Non-profit research institutions and small businesses now retain the rights to patents resulting from federally funded research which they perform. The amendments also give universities the right to own inventions made in federally owned research facilities run by the university under contract with the government.

The amendments encourage government-funded researchers to patent resulting inventions by simplifying the bureaucratic obstacles to licensing and by allowing the patent holder to keep the royalties.⁴⁵ Private industrial firms can exclusively license these patents from the

42. James A. Dobkin, *Patent Policy in Government Research and Development Contracts*, 53 VA. L. REV. 564 (1967) (describing policies of the AEC, NASA, the FAA, the Department of Defense, and the Department of Health, Education and Welfare).

43. S. REP. NO. 480, 96th Cong., 1st Sess. 2 (1979).

44. H.R. REP. NO. 1307, 96th Cong., 2d Sess. 2 (1980), reprinted in 1980 U.S.C.C.A.N. 6460, 6461.

45. See *id.* at 5, reprinted in 1980 U.S.C.C.A.N. at 6464.

university or another government contractor for specific uses they intend to commercialize.⁴⁶

Congress designed the amendments to encourage private industry to commit the capital necessary to develop government-funded inventions to the point of commercial application. Supporters of the amendments argue that without the profit incentive provided by exclusive rights, commercial development lags and research results do not become socially useful. The Secretary of Commerce stated, "Direct access to the university and the university's right to transfer the results of its research on an exclusive basis is an important incentive for business to invest in the further development and commercialization of new technologies."⁴⁷

Thus, the patent system accomplishes the policy goal of transferring the products of university research to the public by allowing a university to license its inventions.⁴⁸ However, the license must be exclusive before companies will invest in development. Inventions arising from university research are often at an early stage of development and the licensee may need to do further development simply to identify a commercial product. Because biotechnology products in particular require expensive regulatory approval, it is difficult to find a licensee who is willing to make the required investment without receiving an exclusive license.⁴⁹

In the past the university scientific community viewed private ownership of discoveries as contrary to the university's mission and the public interest.⁵⁰ Especially in the biomedical fields, some researchers held a belief that new knowledge should be made as widely available as possible to serve humanity.⁵¹ This attitude has changed for several reasons, making universities increasingly likely to patent publicly and privately funded research.

First, the view that basic research should be freely available to everyone was predicated on the assumption that the work being done

46. 35 U.S.C. § 202(c)(7) (1988). The government retains a royalty-free worldwide license to practice the invention or have it practiced for the government. *Id.* § 202(c)(4). In addition, the government has march-in rights that terminate the rights of the contractor if the contractor does not effectively attempt to apply the invention. *Id.* 35 U.S.C. §§ 202(c)(8), 203.

47. S. REP. NO. 662, 98th Cong., 2d Sess. 4 (1984), reprinted in 1984 U.S.C.C.A.N. 5799, 5803.

48. Phyllis S. Lachs, *University Patent Policy*, 10 J.C. & U.L. 263, 276-77 (1983). Of course, this argument assumes that the private sector would not commercially develop the university invention absent an exclusive license.

49. See DAVID DICKSON, THE NEW POLITICS OF SCIENCE 91 (1984); Joyce Brinton, *Biotechnology Licensing: Issues from the University Perspective*, 16 AM. INTELL. PROP. L. ASS'N Q.J. 479, 484 (1988).

50. DICKSON, *supra* note 49, at 89-90; BERNARD BARBER, SCIENCE AND THE SOCIAL ORDER 130 (1952).

51. MARTIN KENNEY, BIOTECHNOLOGY: THE UNIVERSITY-INDUSTRIAL COMPLEX 32 (1986).

had no immediate commercial value. When this premise broke down in fields like molecular genetics, and laboratories produced results with commercial value, various entrepreneurial interests insisted that results be privatized.⁵² Consequently, patent protection for basic research discoveries with potential commercial value has become more commonplace. This is especially true in biotechnology-related fields where the dividing line between basic and applied research is not clear.⁵³ Academic and industrial scientists often work on the same or closely related problems.⁵⁴

Second, universities had little incentive to pursue patent rights before the 1980 amendments because the common practice of government agencies supporting the research was to require that the patent be assigned to the government and then freely licensed.⁵⁵

Because the amendments allow the universities to keep royalties, they are looking to licenses as a way to supplement government money for research. As government support of university research has decreased in terms of constant dollars, the cost of scientific research has rapidly escalated.⁵⁶ Erich Bloch, then director of NSF, testified before a Senate Committee that the federal government is unable to meet all research needs of the universities and, therefore, the universities have a continuing need for additional funding.⁵⁷

Allowing universities to patent and license faculty inventions has produced a number of success stories for different universities. The Cohen-Boyer gene-splicing patent which forms the basis of the biotechnology industry is expected to bring more than \$100 million in royalties to the University of California and Stanford.⁵⁸ The Massachusetts Institute of Technology registers more patents than any other university, over one hundred per year, and licenses up to 53% of

52. *Id.* at 107.

53. DICKSON, *supra* note 49, at 75-76.

54. *Id.* at 74-75; see David Blumenthal et al., *Industrial Support of University Research in Biotechnology*, 231 SCIENCE 242 (1986).

55. Dobkin, *supra* note 42, at 568-84, 591-607.

56. Lachs, *supra* note 48, at 268.

57. *National Science and Technology Issues: Hearing Before the Senate Committee on Commerce, Science and Transportation*, 101st Cong., 2nd Sess. 22 (1990) (statement of Erich Bloch, Director, National Science Foundation) [hereinafter *Technology Issues Hearing*]. For instance, the NIH budget has been rising rapidly, from \$3.2 billion in 1980 to \$7.5 billion in 1990. However, the soaring cost of doing research, the fact that more money is tied up in long-term grants, and the increasing number of scientists applying for grants have created a money drought, especially for younger scientists. NIH research grants account for about 75% of all biomedical research funds provided by the federal government and private nonprofit sources. Gina Kolata, *Beginning Scientists Face a Research Fund Drought*, N.Y. TIMES, June 5, 1990, at C1.

58. Marjorie Shaffer, *All About University Patents: When Research Labs Go After Business*, N.Y. TIMES, Feb. 23, 1992, § 3, at 10; see also DICKSON, *supra* note 49, at 90.

those. In 1991, M.I.T. grossed \$5.5 million from its licensing activities.⁵⁹ Some forty companies employing more than one thousand people have been started based on M.I.T.-licensed technology.⁶⁰

Third, the requirement of the 1980 amendments that universities share royalties with inventors gives researchers an incentive to be alert to patent rights.⁶¹ Universities generally include a patents rights clause in employment contracts with faculty so that the patent must be assigned to the university. Often the university awards between one third and one half of any resulting royalty to the inventor, with the remainder going to the university.⁶² Consequently, the inventor profits from any licensing.

2. UNIVERSITY-INDUSTRY RELATIONSHIPS

For universities, patents provide more than just royalty income. Patents are also a means of strengthening ties with industry and gaining private support for academic research.

Universities are contracting with industry to conduct specific research with the understanding that the industrial firm receives the right to license and commercially develop the results.

In the past, university-industry agreements were generally of a small scale and seldom controversial.⁶³ The situation began to change in the mid-1970s at a time when universities experienced economic pressures from rising operating costs coupled with federal funding that failed to keep pace with the expanding number of scientists. In this atmosphere, university faculty and administrators welcomed increased collaboration with and funding from industry.⁶⁴ Industrial support of academic research made up only 3.8% of the total university research budget in 1980 but has been generally increasing since then.⁶⁵ A 1984 study reveals that industry may be funding as much as one fourth of all biotechnology research in universities.⁶⁶ Fueling industry's increased

59. M.I.T. netted only \$500,000 from the \$5.5 million it grossed from royalty licenses in 1991 due to the costs associated with filing and licensing patents and the \$1 million it distributed to hundreds of individual scientists. Shaffer, *supra* note 58, at 10.

60. *Id.*

61. 35 U.S.C. § 202(c)(7)(C) (1988).

62. Lachs, *supra* note 48, at 281, 285-86 (recommending that universities include a patent rights clause in their employment contracts).

63. DOROTHY NELKIN, SCIENCE AS INTELLECTUAL PROPERTY 18 (1984).

64. *Id.* Universities are partially motivated to accept the corporate sponsorship in order to keep their best scientists, who may move to another university or to industry if denied the corporate funding. KENNEY, *supra* note 51, at 62.

65. KENNEY, *supra* note 51, at 35; NELKIN, *supra* note 63, at 18, 23. Industry contributed \$667 million to university research in 1986. Gretchen Morgenson, *In Pecunia Veritas?*, FORBES, Nov. 1988, at 204, 208.

66. Blumenthal et al., *supra* note 54, at 244.

CONTENTS

HIGH TECHNOLOGY LAW JOURNAL FALL 1992 VOLUME 7 NUMBER 2

ARTICLES

Science and Toxic Torts:

Is There a Rational Solution to the Problem of Causation?

- Susan R. Poulter 189

Antitrust and International Competitiveness:

Is Encouraging Production Joint Ventures Worth the Cost?

- Donald K. Stockdale, Jr. 269

Software Litigation in the Year 2000:

The Effect of Object-Oriented Design Methodologies on Traditional Software Jurisprudence

- David M. Barkan 315

COMMENT

The Experimental Use Exception to Infringement Applied to Federally Funded Inventions

- Suzanne T. Michel 369

ARTICLE

SCIENCE AND TOXIC TORTS: IS THERE A RATIONAL SOLUTION TO THE PROBLEM OF CAUSATION?

SUSAN R. POULTER[†]

Table of Contents

I.	INTRODUCTION	190
II.	HARD CASES MAKE BAD LAW	197
III.	ACTIVE REVIEW OF SCIENTIFIC EVIDENCE.....	205
	A. Active Review and the Rules of Evidence	205
	B. Active Review and Scientific Reasoning	207
IV.	ACTIVE REVIEW OF CAUSATION EVIDENCE IN TOXIC TORTS	213
	A. Validity, Reliability, and the Determination of Probative Value	213
	B. Validity and Reliability of Causation Evidence in Toxic Torts	216
V.	DIVERGENCE OF OPINION	241
	A. Deferential Review and the Accumulation of Errors	241
	B. Active Review Exemplified	250
VI.	ACTIVE REVIEW: THE ANTIDOTE FOR JUNK SCIENCE.....	252
	A. Courts' Ability to Review Scientific Evidence	252
	B. Overcompensating for the Deficiencies and Inequities of the Tort System	254
	C. The Costs of Overcompensation	264

© 1993 Susan R. Poulter.

[†] Associate Professor, University of Utah College of Law; J.D. 1983, University of Utah College of Law; Ph.D., 1969, University of California, Berkeley; B.S. 1965, University of California, Berkeley. The author wishes to thank the following colleagues and friends for their thoughtful review and many helpful comments on drafts of this article: Dean Lee Teitelbaum, Professors Leslie Francis and Wayne McCormack and Associate Professor Paul Cassell of the University of Utah College of Law, and Professor Gary Yost of the Department of Pharmacology and Toxicology of the University of Utah. Any errors are, of course, the author's own.

I. INTRODUCTION

Recent controversies over the safety of breast implants,¹ electrical power transmission lines,² and even cellular phones³ portend yet another period of protracted litigation in which the courts will confront issues of what constitutes admissible and sufficient evidence⁴ of causation in toxic torts.⁵ Questions have surfaced regarding the safety of each product, but there is no clearly established causal link between chronic exposure to any of them and disease or injury. News reports indicate that while anecdotal reports abound regarding breast implants, little if any systematic testing has been done to confirm suspicions of harmful effects.⁶ Concerns about cellular phones were prompted by an even sparser array of anecdotal reports and studies.⁷ Electromagnetic radiation from electrical power lines has been studied more extensively, but many scientists remain unconvinced of the purported link between such

1. See, e.g., Philip J. Hilts, *Experts Suggest U.S. Sharply Limit Breast Implants*, N.Y. TIMES, Feb. 21, 1992, at A1.

2. See Bill Richards, *Elusive Threat: Electric Utilities Brace for Cancer Lawsuits Though Risk Is Unclear*, WALL ST. J., Feb. 5, 1993, at A1.

3. See Natalie Angier, *Cellular Phone Scare Discounted*, N.Y. TIMES, Feb. 2, 1993, at C1.

4. The United States Supreme Court recently granted the plaintiffs' petition for certiorari in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992). In *Daubert*, the Ninth Circuit held that animal testing and chemical studies provided insufficient foundation for expert testimony that Bendectin causes limb reduction defects. 951 F.2d at 1131. The court also held that unpublished reanalyses of epidemiologic studies which had not been peer reviewed and which were generated solely for use in litigation were inadmissible on the issue of causation. *Id.*

5. This article uses the term "toxic tort" for cases, including products liability and environmental exposure cases, in which disease or injury is alleged to have resulted from exposure to harmful substances (i.e., chemicals). See 1 MICHAEL DORE, THE LAW OF TOXIC TORTS § 2.02 (1992). The toxic tort rubric also applies to cases involving radiation exposure. See, e.g., *Allen v. United States*, 588 F. Supp. 247 (D. Utah 1984), *rev'd*, 816 F.2d 1417 (10th Cir. 1987), cert. denied, 484 U.S. 1004 (1988). For discussion of the characteristics of toxic torts cases, see *infra* notes 43-49 and accompanying text.

6. This statement is intended to apply to the issue of whether breast implants or their constituents pose systemic risks. There are, of course, cases in which the implants have ruptured or produced localized effects, where the injuries and the causal role of breast implants is not subject to the same level of doubt.

As the breast implant controversy came to a head, *Chemical & Engineering News* reported:

After 30 years of silicone gel breast implant use, the biological, physiological, physical, and chemical reactions of silicones in the human body are likely, finally, to be systematically studied. A major goal of these studies will be determining how often the devices rupture, and what happens when they do.

Lois Ember, *Breast Implants: Silicone Effects in Body to Be Probed*, CHEMICAL & ENGINEERING NEWS, Mar. 2, 1992, at 4. Almost a year later, the *Wall Street Journal* reported that some researchers have identified diseases that they believe are unique to or more common in breast implant recipients. Joan Rigdon, *Breast Implants Raise More Safety Issues: Saline Implants Appear to Carry Hazard as Well*, WALL ST. J., Feb. 4, 1993, at B1.

7. See *infra* notes 10-13 and accompanying text.

exposure and disease.⁸ Nonetheless, all three exposures are the subject of recently filed, and in some cases adjudicated, lawsuits.⁹

The current scare over cellular phones is instructive. The primary "evidence" of a causal link between the phones and brain cancer is the fact that a number of cellular phone users have been diagnosed with brain cancer, several with the cancer located near the location of the phone's antenna in use. Using newspaper estimates of over three million users of hand held portable cellular phones in the United States¹⁰ and 11,000 expected deaths from brain cancer this year,¹¹ it is hardly surprising that several cases of brain cancer in cellular phone users have been reported. One reported laboratory study which reported that radio-frequency radiation increased the growth rate of tumor cells is consistent with the possibility that such radiation could increase the growth rate of preexisting cancers,¹² but it does not prove that there is any effect in humans from cellular phone use.¹³

8. See Richards, *supra* note 2.

9. *Plaintiffs in Georgia, Texas Sue Makers, Contending Devices Caused Various Ailments*, Current Report, Toxics L. Rep. (BNA) 937 (Jan. 8, 1992) (breast implants). On February 4, 1993, the *Wall Street Journal* reported a plaintiffs' lawyer's estimate that 2000 breast implant cases have been or soon will be filed in consolidated court proceedings in Birmingham, Alabama. Rigdon, *supra* note 6. At least one California case produced a verdict for the plaintiff. *Federal Court Upholds \$7.3 Million Award, Says Verdict Supported, Punitives Proper*, Toxics L. Rep. (BNA) 1480 (May 6, 1992). Regarding radiation from electrical power transmission lines, see *Suit Seeks to Hold Two Utilities Liable for Injuries to Family Living Near Substation*, Toxics L. Rep. (BNA) 927 (Jan. 8, 1992). See also Richards, *supra* note 2, at A1 (describing a "nationwide group of law firms eager to turn [electromagnetic field radiation] into a legal battleground").

Cellular phones are at issue in at least one lawsuit. See Angier, *supra* note 3.

10. See Stephen Nolhgren et al., *A Lethal Connection?*, ST. PETERSBURG TIMES, Jan. 10, 1993, at 1A (reporting estimates of 10 million owners of cellular phones, approximately one third of which are hand-held portables).

11. See Mary Lu Carnevale, *Scientists Doubt Phones Cause Brain Tumors*, WALL ST. J., Feb. 3, 1993, at B1. Richard Adamson, a researcher at the National Cancer Institute, was quoted as predicting 11,800 deaths from brain cancer in the U.S. this year. *Id.* Estimating the population of the U.S. at 250 million, the brain cancer death rate would then be approximately 47 per million, leading to an expected mortality of approximately 140 cases per year among the 3 million hand-held cellular phone users. Even if the age-adjusted cancer rates are lower for the age groups who use cellular phones, it is not unexpected that there would be a number of cases of brain cancer among cellular phone users each year. Further, incidence of brain cancer in the United States is undoubtedly somewhat higher than mortality from the disease.

12. See *supra* note 11.

13. Even the study's author, Stephen Cleary, a physiology and biophysics professor at the Medical College of Virginia, was quoted by the *Wall Street Journal* as stating that he does not believe that portable cellular phones cause cancer. Carnevale, *supra* note 11, at B1. The *Journal* cited scientists from the National Cancer Institute, the Food and Drug Administration, the Environmental Protection Agency, and the Federal Communications Commission as stating that they do not believe that phone use causes brain cancer, but they might pose a small risk of increasing the growth rate of existing cancers. *Id.*

Despite the obvious lack of evidence to prove that cellular phone use causes brain cancer given the current state of knowledge, the evidence available today on cellular phones does not differ substantially in quantity or quality from the evidence that courts have found admissible and sufficient in other recent toxic tort cases. Those problematic cases are likely to be supported only by a combination of anecdotal evidence that amounts to no more than coincidence, speculation in the guise of scientific explanation, and testing based on unvalidated methodology or studies that have limited predictive value for human disease. Sometimes, as in the Bendectin litigation, such evidence is urged upon and accepted by courts in the face of overwhelming scientific consensus, supported by evidence, that a substance is unlikely to be a cause of injury. In other cases, very tenuous evidence is deemed sufficient where more probative positive or negative evidence is unavailable. Such unprobative and insufficient evidence and testimony, termed "junk science" by some observers,¹⁴ has been the subject of increasing commentary and criticism.¹⁵

Erroneous plaintiffs' verdicts and the corresponding overcompensation and overdeterrence are not just academic concerns. The prospect of useful products being driven from the market or of economic resources being diverted from productive uses is real, as the cases of vaccines¹⁶ and

14. The term "junk science" has been popularized by Huber. See PETER HUBER, GALILEO'S REVENGE: JUNK SCIENCE IN THE COURTROOM (1991). At least one court has used the term in a toxic tort case as of this writing. Landrigan v. Celotex Corp., 605 A.2d 1079, 1086 (N.J. 1992).

15. See generally Bert Black, *A Unified Theory of Scientific Evidence*, 56 FORDHAM L. REV. 595 (1988); Jude P. Dougherty, *Accountability Without Causality: Tort Litigation Reaches Fairy Tale Levels*, 41 CATH. U. L. REV. 1 (1991); Peter Huber, *Junk Science in the Courtroom*, 26 VAL. U. L. REV. 723 (1992). For commentary on the courts' tendency to ignore probative evidence in favor of unproven mechanistic explanations and medical testimony, see Troyen A. Brennan, *Causal Chains and Statistical Links: The Role of Scientific Uncertainty in Hazardous Substance Litigation*, 73 CORNELL L. REV. 469 (1988).

16. The cost of litigation and the threat of liability have discouraged research and development of new vaccines, as well as production of existing vaccines, activities that are already of marginal interest to pharmaceutical companies because of high production costs and low return on investment. Louis Lasagna, *The Chilling Effect of Product Liability on New Drug Development*, in THE LIABILITY MAZE 335, 341-45 (Peter W. Huber & Robert E. Litan eds., 1991). In 1991, there was only one U.S. manufacturer of vaccines for measles, mumps, rubella, and polio, down from three to six for each. *Id.* at 344. The high price of vaccines for childhood diseases has recently become the focus of public health concerns about low immunization rates among children in the United States. See Richard L. Berke, *President Assails "Shocking" Prices of Drug Industry*, N.Y. TIMES, Feb. 13, 1993, § 1, at 1. Those prices are attributable in part to liability concerns. See Lasagna, *supra* note 16 at 344; James V. Aquavella, *Profits Don't Explain High Drug Costs*, N.Y. TIMES, Feb. 23, 1993, at A20 (letter to the editor) (attributing high costs to product liability insurance and limited life of patent protection).

Bendectin¹⁷ illustrate. Submission of a case to the jury may result in a plaintiff's verdict where even the most cursory examination of the evidence reveals its deficiencies.¹⁸ Verdicts may be very large,¹⁹ and an occasional plaintiff's verdict may even encourage other suits and increase the settlement value of other cases.²⁰ The social and economic significance of breast implants, electrical transmission lines and cellular phones varies considerably, but clearly the costs to society of an erroneous conclusion that any of them causes harm are significant, potentially even catastrophic.

To deal with the problems of junk science in court, several commentators have suggested that courts regularize the standard for admissibility of scientific evidence. One frequent suggestion is that courts reinstate or continue to apply the standard announced in *Frye v. United States*,²¹ which requires that novel scientific evidence have general acceptance within the relevant scientific discipline,²² an issue that the United States Supreme Court is expected to address this year in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*²³ As will be demonstrated in this article, however, many of the issues that arise are more properly viewed as questions about the sufficiency of relevant evidence to meet the more probable than not standard of proof. Thus, solutions that depend on tightening the criteria for admissibility will either require distortion of the

17. Bendectin was eventually withdrawn from the market despite defense verdicts in the overwhelming majority of cases. *Lasagna, supra* note 16, at 340; see also Joseph Sanders, *The Bendectin Litigation: A Case Study in the Life Cycle of Mass Torts*, 43 HASTINGS L.J. 301, 357 (1992).

18. See, e.g., *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984). Obviously, no plaintiff's verdict can result where a case is not submitted for a decision on the merits. It is understood among plaintiffs' lawyers that the objective is to get to trial. Thus, plaintiffs often propose to fully try a few "bellwether" cases, while defendants move for exclusion of evidence and summary judgment on causation issues. See, e.g., *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1547 (D. Colo. 1990), aff'd, 972 F.2d 304 (10th Cir. 1992).

19. In *Ealy v. Richardson-Merrell, Inc.*, Civ. A. No. 83-3504, 1987 WL 18743 (D.D.C. Oct. 1, 1987), rev'd, 897 F.2d 1159 (D.C. Cir.), cert. denied, 498 U.S. 950 (1990), the jury awarded compensatory damages of \$20 million and punitive damages of \$75 million to a boy born with limb reduction defects attributed to Bendectin. The district court allowed the compensatory verdict to stand, but granted remittitur as to the punitive verdict. The compensatory verdict was reversed on appeal.

20. See Sanders, *supra* note 17, at 357.

21. 293 F. 1013 (D.C. Cir. 1923).

22. See, e.g., *Black, supra* note 15, at 637-38; *Huber, supra* note 15, at 742-47. The *Frye* rule is still followed in many jurisdictions. See, e.g., *Christopherson v. Allied-Signal Corp.*, 939 F.2d 1106, 1110 (5th Cir. 1991) (en banc), cert. denied, 112 S. Ct. 1280 (1992). See generally *Black, supra* note 15, at 601 & n.23. *Frye* was also the basis of rejection of certain of plaintiffs' evidence in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992). See discussion of *Frye infra* notes 55-61 and accompanying text.

23. 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992).

admissibility inquiry to encompass sufficiency issues, or will address only part of the problem. Similar concerns are raised by proposals to change the rules of evidence to limit the use of expert testimony.²⁴

The problem of determining the sufficiency of evidence of causation is more directly addressed by proposals that courts use science boards, science panels or court-appointed experts to assist in resolving scientific issues.²⁵ Such proposals, however, except for the use of court-appointed experts, depart substantially from existing notions of civil jurisprudence because they involve delegation to experts of the traditional fact-finding functions of the lay trier of fact.

The thesis of this article is that measures such as the return to the *Frye* rule, or the use of science panels or science courts are unnecessary, because common law courts already possess the authority under the existing rules to "actively review"²⁶ scientific evidence by eliciting and scrutinizing the reasoning underlying scientific evidence and expert testimony and determining its validity and probative worth. As this article will demonstrate, much of the junk science that appears in toxic tort cases is readily apparent or easily uncovered by inquiry of which courts are quite capable.

If active review under the existing rules can uncover bad science, why do a significant number of courts take a lenient posture toward scientific evidence? There appear to be two major reasons for the deferential approach. First, some courts are philosophically indisposed to examine scientific reasoning or methodology, fearing that they are ill-

24. A working group of the Judicial Conference proposed the following amendment of Rule 702 of the Federal Rules of Evidence:

Testimony providing scientific, technical, or other specialized information, in the form of an opinion or otherwise, may be permitted only if (1) the information is reasonably reliable and will substantially assist the trier of fact to understand the evidence or to determine a fact in issue, and (2) the witness is qualified as an expert by knowledge, skill, experience, training, or education to provide such testimony. Except with leave of court for good cause shown, the witness shall not testify on direct examination in any civil action to any opinion or inference, or reason or basis therefor, that has not been seasonably disclosed as required by Rules 26(a)(2) and 26(e)(1) of the Federal Rules of Civil Procedure.

137 F.R.D. 83 (1991). The proposed changes seem more a shift in emphasis than a radical revision of the existing rule. *See also* Black, *supra* note 15, at 611-13 (proposing a modification of Rule 702 to require the court to determine the validity of reasoning as well as its reliability as a precondition to admitting scientific evidence).

25. See 2 AMERICAN LAW INST. REPORTERS' STUDY, ENTERPRISE RESPONSIBILITY FOR PERSONAL INJURY, APPROACHES TO LEGAL AND INSTITUTIONAL CHANGE 332-51 (1991). The ALI Reporters' Study recommendations were based on Brennan, *supra* note 15, and Troyen Brennan, *Helping Courts with Toxic Torts: Some Proposals Regarding Alternative Methods for Presenting and Assessing Scientific Evidence in Common Law Courts*, 51 U. PITTS. L. REV. 1 (1989).

26. For a discussion of "active review," see Black, *supra* note 15, at 674-77.

equipped to delve into scientific disciplines. As will be described below, however, scientific reasoning and legal factfinding employ the same rules of logic. Thus, lay judges need not fear that examination of scientific evidence to determine whether it is soundly reasoned and reliable is beyond their capabilities.

Moreover, the reasons for judicial control of evidence are more compelling where technical evidence is concerned than for non-technical evidence. Judges exhibit no hesitation in barring non-expert testimony based on hearsay and otherwise lacking in foundation even though juries could readily identify the flaws in such testimony with skilled cross-examination and argument by opposing counsel. Juries are less likely to identify the weaknesses in testimony cloaked in technical jargon from an expert with a lengthy list of credentials than in testimony on ordinary factual issues.²⁷ Thus, it is more important for the judge, who understands the legal requirements of proof, to discriminate between reliable and unreliable scientific evidence than between well founded and unfounded evidence on matters within the understanding of ordinary people.²⁸

A second reason for the lenient treatment of scientific evidence in some courts is the apparent desire to compensate for perceived inequities and deficiencies of the tort system. Much of the movement toward the adoption of lenient standards of admissibility and proof of causation in toxic torts has been prompted by the recognition of the difficulties faced by plaintiffs in meeting the traditional requirement that they prove, by a preponderance of the evidence, that their injuries were caused by chronic, low-level chemical or radiation exposures that were remote in time from the manifestation of injury. The paucity of scientific evidence on the causation of diseases such as cancer and birth defects, and the difficulty of distinguishing other identified or background risk factors for the disease, decrease the likelihood that deserving plaintiffs will be compensated. The level of concern about those difficulties was heightened by increasing scientific knowledge of the role of chemicals and radiation in diseases such as cancer and birth defects, as well as scientific speculation about

27. Courts that scrutinize scientific evidence more closely recognize that jurors are likely to be persuaded by the aura of infallibility that surrounds scientific evidence, or by the credentials and certainty expressed by the expert. *See Barefoot v. Estelle*, 463 U.S. 880, 926-28 (1983) (Blackmun, J., dissenting).

28. Courts' abandonment of the *Frye* standard increases the need for judicial scrutiny of scientific evidence because the *Frye* general acceptance standard assures that some evaluation of methods or theories other than that of the expert witness has occurred. Once courts unhinge the admissibility of scientific evidence from scientists' standards, it is incumbent on them to see that other safeguards are in place. *See Steven M. Egesdal, Note, The Frye Doctrine and Relevancy Approach Controversy*, 74 GEO. L.J. 1769, 1787 (1986) (suggesting the need to increase jurors' understanding of novel scientific techniques under the relevancy approach).

potential effects of the greatly accelerated dissemination of untested new chemicals in consumer products and the environment.²⁹ Taking their cue from the scientists,³⁰ legal scholars began to address the difficulties faced by plaintiffs in proving that exposure to toxic substances or chemicals caused their diseases or injuries,³¹ difficulties that can result in uncompensated injuries and the failure to adequately deter harmful activity.³² Lenient standards of admissibility and proof certainly facilitate plaintiffs' recoveries; further, they are consistent with courts' suspicions that mainstream scientists are too demanding in their requirements of proof, and that the unconventional scientists who testify that an exposure caused a plaintiff's disease may be correct.

More than a decade of scientific research into cancer incidence and causation, however, has failed to bear out the fears that prompted deferential review of causation evidence. Many of the assumptions that underlay the shift to more lenient standards for causation evidence in toxic torts are still unproven or are even contrary to current scientific thinking. The contribution of toxic synthetic chemicals and other hazards of the industrial age to cancer and other diseases and injuries is still an open question, but it appears unlikely that such substances cause anything approaching a majority of human cancer and birth defects.

As for the possibility that the unconventional expert may be right, even a superficial examination of much of the disputed evidence reveals that it amounts to speculation about possibilities that have not been tested or that fall far short of meeting the more probable than not standard of proof. Speculation about possibilities forms the beginning, not the endpoint, of factual inquiry, in either the scientific or legal realm. A causal explanation of disease or injury can be said to be probable only when it is supported by observations or data that distinguish between it and other possible explanations. When courts authorize or approve plaintiffs' verdicts without a factual basis for causal inference, they undermine traditional tort requirements for rational factfinding and the "more probable than not" standard of proof. The case for the abrogation of those standards has not been made, nor have courts given full consideration to the implications of such a radical change in the law.

29. See Bruce N. Ames, *Identifying Environmental Chemicals Causing Mutations and Cancer*, 204 SCIENCE 587, 588-89 (1979).

30. See, e.g., *id.* at 592 (recommending short-term mutagenicity testing to expedite identification of environmental mutagens and carcinogens).

31. See, e.g., Jeffrey Trauberman, *Statutory Reform of "Toxic Torts": Relieving Legal, Scientific and Economic Burdens on the Chemical Victim*, 7 HARV. ENVTL. L. REV. 177, 188 n.48 (1983) (citing law review articles and other writings).

32. See generally David Rosenberg, *The Causal Connection in Mass Exposure Cases: A "Public Law" Vision of the Tort System*, 97 HARV. L. REV. 849 (1984).

The purpose of this Article is to demonstrate that courts can and should actively review scientific evidence of causation in toxic tort cases. The next Part describes how courts have loosened the standards for expert testimony in an effort to compensate for the perceived problems faced by toxic tort plaintiffs. Part III then discusses active review and its relation to the rules of evidence and civil procedure and attempts to allay courts' fears that they are ill equipped to evaluate the basis of scientific opinion testimony. Part IV then describes the criteria against which the reliability of scientific evidence can be evaluated and then applies those criteria to the kinds of evidence offered on causation in toxic tort suits. Part V examines a sampling of recent cases that illustrate inadequate judicial scrutiny of scientific evidence, as well as cases that skillfully distinguish probative from nonprobative or insufficient evidence. Lastly, Part VI discusses in depth the factors that underlie courts' failure to examine adequately scientific evidence and shows that many of those concerns are unjustified or that, even where justified, the remedy of authorizing plaintiffs' verdicts that are unsupported by a factual foundation goes too far.

II. HARD CASES MAKE BAD LAW

In the 1960s and 1970s, mounting evidence on the harmful effects of chemicals such as asbestos, vinyl chloride, dioxin and many others, together with the dramatic increase in the use of new chemicals in products ranging from foods, to drugs and medical devices, to many other consumer products, raised concerns that chronic, low level exposures to those substances would lead, or might already have led, to widespread illness and injury.³³ As evidence mounted that exposure to substances such as asbestos and vinyl chloride could cause cancer and other debilitating or fatal conditions, the courts began to see an increasing number of toxic tort suits—tort actions seeking to recover for injuries attributed to toxic substances.

As numerous commentators have explained, proof of causation³⁴ has been the biggest stumbling block to recovery in toxic torts cases.³⁵

33. See R. Jeffrey Smith, *Government Says Cancer Rate Is Increasing*, 209 SCIENCE 998 (1980); Mostafa K. Tolba, *Chemicals in the Environment*, 1979 NAT'L PARKS & CONSERVATION MAG. 16. The controversy continues, as indicated by more recent publications. See Eliot Marshall, *Experts Clash over Cancer Data*, 250 SCIENCE 900 (1990); see also *infra* notes 332-38 and accompanying text.

34. This Article is addressed to issues of causation in fact, by which is meant the issue of whether there is an empirical linkage between the causative event and the claimed injury.

35. See Brennan, *supra* note 25, at 2; Daniel A. Farber, *Toxic Causation*, 71 MINN. L. REV. 1219, 1219-20 (1987); Jean M. Eggen, *Toxic Reproductive and Genetic Hazards in the Workplace: Challenging the Myths of the Tort and Worker's Compensation System*, 60 FORDHAM L. REV. 843, 861-64 (1992) (discussing causation problems in the worker's compensation system);

Both negligence and strict liability require the plaintiff to prove that the substance in question³⁶ caused the plaintiff's disease or injury.³⁷ That inquiry often involves a number of subissues,³⁸ including whether: (1) the toxic substance is capable of causing the harm complained of³⁹; (2) the plaintiff was exposed to the toxic substance in quantity sufficient to cause disease,⁴⁰ and (3) the toxic substance exposure caused the particular plaintiff's injury or disease.⁴¹ Proof of any of these propositions is likely to require expert testimony on scientific evidence.⁴²

Palma J. Strand, Note, *The Inapplicability of Traditional Tort Analysis to Environmental Risks: The Example of Toxic Waste Pollution Victim Compensation*, 35 STAN. L. REV. 575, 583-84 (1983); Note, *Tort Actions for Cancer: Deterrence, Compensation, and Environmental Carcinogenesis*, 90 YALE L.J. 840 (1981) (hereinafter Note, *Tort Actions for Cancer*).

36. Disputes over who produced the offending substance have also been cast as causation questions. These "indeterminate defendant" cases have arisen frequently in asbestos and DES litigation where the plaintiff may have difficulty identifying the producer of the substance to which the plaintiff was exposed, even where the causal connection between the substance and the injury is established. See Richard Delgado, *Beyond Sindell: Relaxation of Cause-in-Fact Rules for Indeterminate Plaintiffs*, 70 CAL. L. REV. 881 (1982); Eggen, *supra* note 35, at 890-91 & n.258.

37. Most courts require proof of causation to meet a "more likely than not" standard. See, e.g., *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1553 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992). See generally Bert Black & David E. Lilienfeld, *Epidemiologic Proof in Toxic Tort Litigation*, 52 FORDHAM L. REV. 732, 749-50 (1984). But see Black, *supra* note 15, at 659-69 (discussing the meaning of "reasonable medical certainty"). Additionally, most jurisdictions require the plaintiff to prove that her injuries would not have occurred "but for" the exposure to the toxic substance. Brennan, *supra* note 15, at 493-94. Where there are two or more contributing causes to a single harm, some courts will require proof only that the exposure was a "substantial factor" in causing the plaintiff's injury or that it "contributed to" the plaintiff's injury. See *Renaud v. Martin Marietta Corp.*, 749 F. Supp. at 1551 (plaintiff must prove that "the exposure caused, or contributed to, plaintiff's injuries"). Proof under the substantial-or-contributing-factor test nonetheless requires establishment of a "but for" causal relationship between the substance and the plaintiff's disease. See Bert Black et al., *Unravelling Causation: Back to the Basics*, 7 Toxics L. Rep. (BNA) 1061, 1063 (1993). A somewhat different formulation, perhaps more suited to the realities of toxic torts, is Calabresi's "causal linkage," that is, the belief that the causative event makes the occurrence of the injury result more likely. See Guido Calabresi, *Concerning Cause and the Law of Torts: An Essay for Harry Kalven, Jr.*, 43 U. CHI. L. REV. 69, 71 (1975).

38. See Black & Lilienfeld, *supra* note 37, at 737-38.

39. See Black, *supra* note 15, at 689. Although this framing of the question seems implicit, plaintiffs sometimes argue that evidence of causation of one type of harm is evidence of causation of other types of harm. *Id.*; see also *Christopherson v. Allied-Signal Corp.*, 939 F.2d 1106, 1115 (5th Cir. 1991) (en banc) (association of nickel and cadmium with small-cell carcinoma of the lung asserted as probative of causation of small-cell colon cancer), *cert. denied*, 112 S. Ct. 1280 (1992).

40. See Black & Lilienfeld, *supra* note 37, at 737-38. Courts sometimes frame the question more simply as whether the plaintiff was exposed to the toxic substance, and there is some divergence in the case law as to the specificity with which exposure must be proved. See *infra* notes 219-20 and accompanying text.

41. This statement, which appears all-inclusive, is intended to cover those aspects of causation-in-fact that remain after exposure and capability of the substance to cause harm ("general causation") are established, including primarily the issue of whether plaintiff's

Several characteristics of the typical toxic tort case diminish the prospects of recovery by deserving plaintiffs.⁴³ The long latency period between exposure and disease manifestation⁴⁴ decreases the likelihood that the plaintiff will even suspect the causal connection, as well as decreasing the likelihood that the plaintiff will be able to marshal the facts on issues such as exposure necessary to prove her case.⁴⁵ Typically there is no clinical evidence capable of linking the substance to the disease.⁴⁶ The situation is further complicated by the fact that exposure to the toxic

injury was the result of the toxic substance exposure or other causes. This issue is sometimes referred to as one of "individual causation" or "medical causation." *Renaud v. Martin Marietta Corp.*, 972 F.2d 304, 306 (10th Cir. 1992) (discussing medical causation); *see also Rosenberg, supra* note 32, at 855-56 (discussing "specific causation").

42. *See, e.g., In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990) (plaintiff's case depended upon expert testimony relating to exposure and causation), *cert. denied*, 111 S. Ct. 1584 (1991); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (expert testimony on exposure and individual causation), *aff'd*, 972 F.2d 304 (10th Cir. 1992). As a definitional matter, this Article will use the terms science and scientific evidence to encompass both science, in the sense of discovery of new factual information, and technology, which can be defined as application of established scientific principles to a particular problem. *See Howard T. Markey, Needed: A Judicial Welcome for Technology—Star Wars or Stare Decisis?*, 79 F.R.D. 209, 210-12 (1978). An additional assumption will be made that scientific evidence will be presented by expert witnesses, because that is most often the case.

43. *See generally Brennan, supra* note 25, at 20-26 (discussing cancer causation); *Strand, supra* note 35, at 578-86.

44. *See Strand, supra* note 35 at 580-81. More precisely, the lapse of time between exposure and the appearance of clinical symptoms may comprise both an induction period, the period of time between the exposure and disease initiation, and a latency period, the interval between disease occurrence and detection. *See KENNETH J. ROTHMAN, MODERN EPIDEMIOLOGY* 14-15 (1986). The period between first exposure and clinically detectable disease for many cancers is 20 to 30 years. *Ames, supra* note 29, at 587. Birth defects that are manifest at birth or soon thereafter would not exemplify this problem to nearly as great a degree.

45. Delay may, however, may increase the chance that epidemiologic evidence will be available. Nonetheless, latency also gives rise to problems under some formulations of statutes of limitation, although many jurisdictions employ the discovery rule to determine when the statute of limitations begins to run. *See Black & Lilienfeld, supra* note 37, at 780; *Strand, supra* note 35, at 580-81.

46. In cases involving "signature diseases," diseases that are almost exclusively associated with a toxic substance, the presence of the condition is highly probative of the causative agent. Examples of signature diseases are mesothelioma, associated almost entirely with asbestos exposure, and clear cell adenocarcinoma of the vagina, associated almost exclusively with diethylstilbestrol (DES) exposure in utero. *See Brennan, supra* note 25, at 21 & n.96. In most cases, however, either the toxic substance is no longer present when the disease manifests itself, as is the case with benzene and leukemia, or its presence, if persistent, is not the only or even most probable explanation of disease. *See Brennan, supra* note 15, at 502. An example of the latter is the almost ubiquitous presence of PCBs in human adipose tissue, apparently without effect in most cases. *See In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 843 (3d Cir. 1990) (discussing ATSDR studies), *cert. denied*, 111 S. Ct. 1584 (1991). *But see Landrigan v. Celotex Corp.*, 605 A.2d 1079, 1087 (N.J. 1992) (discussing the presence of asbestos near the tumor as probative of colon cancer causation).

substance, even at relatively high levels, may not result in disease in most persons.⁴⁷ Moreover, many of the diseases caused by toxic chemicals, particularly cancers and birth defects, occur in the general population.⁴⁸ The absence of any unequivocal linkage between the disease and the toxin, together with the absence of clinical tests that could establish a linkage, means that proof of causation, if it can be made out at all, must be made indirectly, from comparisons between exposed and unexposed groups, or from studies where surrogates such as animals or single-celled organisms are used. Further, there may be other known risk factors for the claimed injury, whose role in the disease process must be considered.⁴⁹

The obvious difficulties of proof in toxic tort cases provoked a flood of commentary and proposals for reform.⁵⁰ A number of commentators have focused specifically on limitations placed by courts on the kinds of

47. Occupational asbestos exposure in nonsmokers increases the risk of lung cancer by about a factor of five, from about 11 per 100,000, for nonsmoking industrial workers not exposed to asbestos, to about 58 per 100,000 for nonsmoking asbestos workers. See U.S. SURGEON GEN., U.S. DEP'T OF HEALTH & HUMAN SERVS., PUB. NO. 85-50207, HEALTH CONSEQUENCES OF SMOKING: CANCER AND CHRONIC LUNG DISEASE IN THE WORKPLACE 216 (1985); see also Rodolfo Saracci, *The Interactions of Tobacco Smoking and Other Agents in Cancer Etiology*, EPIDEMIOLOGIC REV. 175, 181-83 (1987).

48. See, e.g., Rubanick v. Witco Chem. Corp., 593 A.2d 733, 745 (N.J. 1991) (involving contention that PCB exposure caused colon cancer). Colorectal cancer is the second most common cancer in the United States. *Colonoscopy Recommended*, AM. MED. NEWS, Sept. 16, 1991, at 39, cited in Landrigan, 605 A.2d at 1082).

49. For example, although asbestos is recognized as a cause of lung cancer, *see supra* note 47, other causative factors such as smoking are well known. That fact often leads to contentions that the plaintiff's disease was caused by factors other than the toxic chemical exposure. For discussion of attributable risk and the problems of distinguishing among causes, see *infra* notes 206-18 and accompanying text.

50. Some commentators have proposed modification of the tort system's rules of liability, suggesting, for example, that courts recognize causes of action for tortiously created risk. See, e.g., Glen O. Robinson, *Probabilistic Causation and Compensation for Tortious Risk*, 14 J. LEGAL STUD. 779 (1985) [hereinafter Robinson, *Probabilistic Causation*]. Others have suggested that all victims of a disease attributable in part to toxic chemical exposure recover the fraction of their damages that corresponds to the proportion of disease incidence attributable to the toxic exposure. See, e.g., Delgado, *supra* note 36, at 892; Glen O. Robinson, *Multiple Causation in Tort Law: Reflections on the DES Cases*, 68 VA. L. REV. 713, 759 (1982); Rosenberg, *supra* note 32; cf. Farber, *supra* note 35, at 1221 (proposing compensation for the "most likely victim"). A number of courts have redefined damages or injury to include exposure, presumed subclinical injury, medical monitoring costs, or fear of cancer where clinically manifest disease or injury is absent. See 2 DORE, *supra* note 5, §2.02.

Other commentators have recommended the shifting burden of proving causation to defendants, once a threshold showing is made of the possibility of harm. See Note, *Tort Actions for Cancer*, *supra* note 35, at 855-62. Still others have suggested administrative compensation systems with reduced requirements for proof of causation. See Black & Lilienfeld, *supra* note 37, at 734 & nn.3-5 (discussing the Superfund Study Group's proposal for an administrative compensation scheme); see also E. Donald Elliott, *Why Courts? Comment on Robinson*, 14 J. LEGAL STUD. 799, 801 (1985).

evidence deemed admissible or sufficient to prove causation. One set of problems has been courts' reluctance to accept statistical evidence, such as epidemiologic studies, because statistical evidence does not provide mechanistic explanations of cause and because statistics do not provide a basis for distinguishing between persons in an exposed group whose disease was caused by the exposure from those whose disease was caused by background or other risk factors.⁵¹ Recognizing that epidemiologic evidence is often the best if not the only evidence linking a toxic substance exposure to disease, however, recent cases have been more accepting of epidemiologic evidence,⁵² in some cases evidencing quite a sophisticated understanding of epidemiologic evidence.⁵³

Other commentators have urged courts to liberalize the standards for admissibility of scientific evidence in general.⁵⁴ They have suggested that the traditional requirement under *United States v. Frye* that limits the admission of scientific evidence to that generally accepted in the relevant scientific discipline⁵⁵ may preclude recovery by deserving plaintiffs who

51. See, e.g., Black & Lilienfeld, *supra* note 37, at 767; Brennan, *supra* note 15, at 491-501.

52. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307, 313 (5th Cir.) (holding absence of "conclusive" epidemiologic evidence fatal to plaintiffs' case), modified, 884 F.2d 166, 167 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990). The Fifth Circuit subsequently modified *Brock*, stating that the plaintiffs' case was fatally flawed because of their failure to present "statistically significant" epidemiologic evidence. *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 884 F.2d 166, 167 (5th Cir. 1989) (denying plaintiffs' motion for rehearing en banc and modifying prior opinion), cert. denied, 494 U.S. 1046 (1990). Courts willing to accept statistical evidence as probative of the capability of a substance to cause harm have sometimes balked at accepting such evidence on the question of whether the substance caused the plaintiff's injury, on the basis the epidemiologic evidence cannot prove individual causation. See, e.g., *Landrigan v. Celotex Corp.*, 605 A.2d 1079, 1087 (N.J. 1992) (discussing the trial court's refusal to allow an epidemiologist to testify on individual causation). In *Landrigan*, the New Jersey Supreme Court, however, set forth a detailed summary of how epidemiologic reasoning could be applied to the question of individual causation and concluded that an epidemiologist could offer an opinion on that issue, provided the expert's qualifications and methodology withstood the trial court's scrutiny. *Id.* at 1087-89.

53. See, e.g., *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 946-49 (3d Cir. 1990) (discussing statistical significance); *Landrigan*, 605 A.2d at 1087 (discussing the concept of attributable risk derived from epidemiologic studies). Several courts have announced that epidemiologic evidence is the only sufficient evidence on the question of whether Bendectin causes human birth defects. See, e.g., *Brock*, 874 F.2d at 313-15.

54. Anne S. Toker, *Admitting Scientific Evidence in Toxic Tort Litigation*, 15 HARV. ENVTL. L. REV. 165 *passim* (1991); see also citations in *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 740 (N.J. 1991).

55. The traditional standard for determining the admissibility of novel scientific evidence was set forth in *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). The *Frye* court stated, in regard to evidence based on a forerunner of modern polygraph testing, that "the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field to which it belongs." *Id.* at 1014. The *Frye* test is most appropriately applied to the expert's methodology or reasoning, including but not limited to devices or techniques such as the breathalyzer or polygraph. See Black, *supra* note 15, at 627-29. It is sometimes applied to the expert's opinion or conclusions, however,

must rely on novel, yet valid and reliable evidence.⁵⁶ That line of reasoning was accepted in *Ferebee v. Chevron Chemical Company*,⁵⁷ in which the Court of Appeals for the District of Columbia Circuit upheld a jury verdict of liability based on expert opinion testimony on causation that did not enjoy general acceptance in the scientific community.⁵⁸

Ferebee coincided with a general move away from the *Frye* standard under the Federal Rules of Evidence toward the relevancy or reliability test articulated in *United States v. Downing*.⁵⁹ In *Downing*, the Third Circuit stated that the admissibility of scientific evidence should focus on the soundness and reliability of the expert's methodology, the strength of the connection between the evidence and the issues in the case, and the possibility of confusing or misleading the jury.⁶⁰ Acceptance of the

in such cases being stated to require that the expert's opinion or theory be generally accepted by the relevant scientific community. *See, e.g.*, Rubanick v. Witco Chem. Corp., 542 A.2d 975, 982 (N.J. Super. Ct. Law Div. 1988) (applying *Frye* analysis to scientific principle on which expert's opinion was based), *rev'd*, 576 A.2d 4 (N.J. Super. Ct. 1990), *modified*, 593 A.2d 733 (N.J. 1991); *see also* Black, *supra* note 15, at 629-38. Contrarily, some commentators have taken the position *Frye*'s general acceptance test should not be applied to an expert's reasoning or methodology, but only to particular techniques or devices. *See, e.g.*, Christopherson v. Allied-Signal Corp., 939 F.2d 1106, 1131-33 (5th Cir. 1991) (Rawley, J., dissenting), *cert. denied*, 112 S. Ct. 1280 (1992). Many jurisdictions still follow *Frye*. *See, e.g.*, Christopherson, 939 F.2d 1106.

56. *Frye* has proven to be a significant barrier to novel scientific theories and methodologies. Edward J. Imwinkelreid, *The Standard for Admitting Scientific Evidence: A Critique from the Perspective of Juror Psychology*, 28 VILL. L. REV. 554, 555-56 (1982-83). As Huber has pointed out, however, when the *Frye* inquiry is directed to the methodology and reasoning underlying scientific opinion, a novel opinion on causation will easily pass muster if it is based on well-established and properly conducted methods, such as epidemiologic studies. Huber, *supra* note 15, at 744.

57. *See Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984).

58. *Id.* at 1535-36. The *Ferebee* court did not reject *Frye* out of hand, however, but construed it as applicable only to novel techniques or methodologies, not scientific opinion testimony. *Id.* at 1535.

59. 753 F.2d 1224 (3d Cir. 1985); *see also* *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 856-60 (3d Cir. 1990), *cert. denied*, 111 S. Ct. 1584 (1991); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1242 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988); Rubanick v. Witco Chem. Corp., 593 A.2d 733, 746 (N.J. 1991).

60. In *Downing*, the Third Circuit articulated the proper test as follows:

In our view, Rule 702 [of the Federal Rules of Evidence] requires that a district court ruling upon the admission of (novel) scientific evidence . . . conduct a preliminary inquiry focusing on (1) the soundness and reliability of the process or technique used in generating the evidence, (2) the possibility that admitting the evidence would overwhelmingly confuse or mislead the jury, and (3) the proffered connection between the scientific research or test result to be presented and the particular disputed factual issues in the case.

Downing, 753 F.2d at 1238. The *Downing* reliability standard is inherently more flexible than *Frye* because it is not tied to "general acceptance." Nonetheless, courts recognize that acceptance in the expert community is an important indicium of reliability. *See, e.g., id.*

expert's techniques or methodology in the relevant scientific community is evidence of soundness, but need not be the sole basis for that determination.

Although *Frye* has been justifiably criticized as too simplistic and inflexible,⁶¹ the *Downing* standard is equally problematic when it is used to justify such minimal scrutiny of the reliability of scientific evidence, particularly of expert opinion testimony, that it amounts to no standard at all. The troublesome, deferential application of the reliability standard adopts the approach that if "qualified" experts are willing to testify that a causal relationship exists, the court is willing to uphold a plaintiff's verdict without examining whether a reasoned basis exists for the expert's opinion.⁶² This approach is undoubtedly the result of some courts' reluctance to delve into the reasoning underlying scientific evidence, a reluctance that results in deference to the expert with seemingly impressive credentials. The crucial determination then becomes whether the expert is qualified, a particularly weak screening device given the lenient standards for determining expert qualifications.⁶³

Deferential review is the gateway for the admission of junk science into the courts. When courts do not examine the reasoning of expert testimony, they are likely to accept medical opinion based on the facts in the case at hand, or supported by perhaps a few other case reports, facts

Thus, the *Frye* standard is related to reliability, though more limiting. See generally Imwinkelried, *supra* note 56.

61. Part of the difficulty with the *Frye* rule is the lack of consensus regarding the subject matter to which it applies. For example, is it the expert's opinion, the reasoning or methodology that underlies the opinion, or both that must be generally accepted? See *supra* note 55. The better rule would seem to be that the *Frye* general acceptance test applies to the expert's reasoning and methodology, but not to the opinion or conclusion derived from that methodology. Otherwise, the *Frye* rule effectively delegates part of the admissibility determination to the scientific discipline, obviating the need for the court to evaluate the expert's reasoning or methodology. On the other hand, as Black has pointed out, the general acceptance test of *Frye* is not an appropriate standard to apply to the uncertainty or accuracy (i.e., the reliability) of scientific methodology. See Black, *supra* note 15, at 629-57. The Ninth Circuit's opinion in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 951 F.2d 1128 (9th Cir. 1991), *cert. granted*, 113 S. Ct. 320 (1992), appears to commit this error, when it frames the admissibility standard regarding an unpublished, un-peer-reviewed reanalysis of epidemiologic data as follows: "Expert opinion based on a scientific technique 'is admissible if it is generally accepted as a reliable technique among the scientific community.'" *Id.* at 1129 (quoting *United States v. Solomon*, 753 F.2d 1522, 1526 (9th Cir. 1985)).

62. *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 429 U.S. 1062 (1984), is the leading case following this approach and is often cited by other courts taking similar approaches. In *Ferebee*, the Court of Appeals for the District of Columbia Circuit upheld a jury verdict for the plaintiff where testimony of causation was based on "tissue samples, standard tests, and patient examination." *Id.* at 1536. There is nothing in the opinion to suggest that the cited tests and examinations were capable of indicating the cause of the lung disease complained of, however.

63. FED. R. EVID. 702 provides that a witness may be qualified as an expert "by knowledge, skill, experience or training."

that cannot establish causation because the coincidence of exposure and disease may be the result of chance.⁶⁴ In some cases, courts accept as sufficient medical or similar opinions supplemented by reference to animal studies, chemical structure-activity analyses, mutagenicity testing, or other similar lines of reasoning that are subject to a large degree of uncertainty.⁶⁵ Affirmative epidemiologic evidence of a statistically significant association between the alleged causative agent and human disease is absent.⁶⁶ As a practical matter, only those cases based on studies in human populations of the association of suspected toxic substances and disease—e.g., epidemiologic studies or highly unusual disease clusters—have proven to be sound as new scientific information developed.⁶⁷

A reliability analysis should not result in uncritical acceptance of junk science.⁶⁸ Tort jurisprudence requires that there be a rational basis

64. See, e.g., *Ferebee*, 736 F.2d at 1535.

65. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 735-36 (N.J. 1991) (expert based opinion that PCBs caused plaintiff's colon cancer on animal test reports and other cancer cases at Witco). The Bendectin litigation has been characterized by plaintiffs' cases based on animal testing, structure-activity relationships, *in vitro* testing, and reanalysis of data from epidemiologic studies that failed to show statistically significant increased risks. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990). See generally *Sanders, supra* note 17. Brennan in particular has urged courts to admit animal studies and other methods used in cancer and other medical research. *Brennan, supra* note 25, at 41-57; see also Michael D. Green, *Expert Witnesses and Sufficiency of the Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and the Bendectin Litigation*, 86 NW. U. L. REV. 643, 680-81 (1992).

66. Occasionally, plaintiffs may offer a "reanalysis" of existing epidemiologic data. See *infra* notes 350-58.

67. It may be tempting to characterize the argument made herein as establishing a threshold requirement of epidemiologic evidence to support a toxic tort case. A number of commentators have characterized *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990), and cases that have followed it as creating such a threshold in Bendectin cases. See, e.g., Green, *supra* note 65, at 679-82. The intent of this Article, however, is to show why, given the present state of toxicological science, anecdotal evidence, animal test results, and other evidence offered when positive human evidence is missing are generally unreliable and insufficiently probative in the typical toxic torts case. The kind of analysis proposed herein can be applied to new information as it develops, without the rigidity of a *per se* rule about specific kinds of evidence.

68. In *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988), Judge Weinstein excluded the causation opinion testimony of several of plaintiffs' witnesses because he concluded that their testimony, which relied on animal tests and studies of industrial exposures, and which failed to consider and eliminate other causal explanations, was "insufficiently grounded in any reliable evidence." *Id.* at 1248-51. Although Judge Weinstein cited Rule 703 as the basis of his ruling, *see id.* at 1243-55, it is clear that he recognized the uncertainty associated with causal inferences derived from animal studies or human studies where exposures differed widely from plaintiffs', particularly where the experts ignored more relevant studies and alternative causal explanations. *See id.* at 1250. Under the analysis proposed in this Article, the factors cited by Judge Weinstein would be part of a reliability

for judicial findings of fact.⁶⁹ The relevancy or reliability standard's "soundness and reliability" inquiries bear directly on whether there is a rational basis for findings of fact and whether the evidence is sufficient to meet the more probable than not standard of proof.⁷⁰ Active review facilitates the inquiries necessary to decide those issues, while deferential review avoids them. Courts cannot and should not avoid those responsibilities by deferring to "qualified" experts.

III. ACTIVE REVIEW OF SCIENTIFIC EVIDENCE

A. Active Review and the Rules of Evidence

The active review contemplated by this article and being conducted by some courts is a process in which the court conducts two inquiries. First, the court examines the evidentiary basis and reasoning of scientific opinion testimony and determines whether there is a rational basis for the opinion. The evidentiary basis of the opinion, as well as the expert's reasoning, can be probed by the proponent of the testimony, the opponent, or the court,⁷¹ and will often be assisted by the defendants' experts.

The second inquiry focuses on the sufficiency of the admissible evidence to meet the plaintiff's burden of proof. This inquiry goes to the reliability or accuracy of the evidence and requires that the plaintiff present admissible evidence from which a reasonable juror could find that it is more probable than not that the defendant caused the plaintiff's

analysis. *See infra* notes 310-14 and accompanying text; *see also* Black, *supra* note 15, at 674-76.

69. *See, e.g.*, *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. at 1250.

70. FED. R. EVID. 104 requires the court to determine questions of admissibility of evidence. *See, e.g.*, *Egger v. Burlington N.R.R.*, No. CV89-159-BLG-JFB, 1991 U.S. Dist. LEXIS 19240 (D. Mont. Dec. 18, 1991). Generally, the proponent of evidence must demonstrate by a preponderance of the evidence that the evidence in question is admissible. *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. at 1239. Under FED. R. CIV. P. 50, 56, the court must determine the sufficiency of the evidence on a motion for summary judgment, a motion for a directed verdict, or a motion for judgment notwithstanding the verdict. Generally the standard for granting any of the foregoing (for defendant) is that no reasonable juror could find or have found for the plaintiff. *See, e.g.*, *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1555 (D. Colo. 1990) (granting summary judgment to defendants), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

71. FED. R. EVID. 705 provides: "The expert may testify in terms of opinion or inference and give his reasons therefor without prior disclosure of the underlying facts or data, unless the court requires otherwise. The expert may in any event be required to disclose the underlying facts or data on cross-examination."

As a practical matter, the court may have to do some translation of the language of the scientific field or of legal expressions into commonly understood terms. The court may also be guided by court-appointed experts who serve as witnesses or advisors. For an example of court-appointed experts serving as advisors to the judge, see *Renaud*, 749 F. Supp. 1545.

disease or injury. The same tools used to probe the underlying reasoning of the evidence can be used to inquire into its accuracy, but the question of whether the evidence is sufficiently accurate to satisfy legal standards is, of course, a legal question.

It is important to note that active review is not strict scrutiny.⁷² The plaintiff need not show that her evidence is stronger than the defendant's or that it meets some high level of certainty. The plaintiff's scientific evidence need only be such that a rational factfinder could conclude from the testimony that it is more likely than not that the defendant caused the plaintiff's injury.⁷³ Only when the factual basis and reasoning underlying the expert's opinion on causation do not meet that minimum level of rationality and accuracy should the evidence be excluded.

Active review is not tied to any particular formulation of the standards for admissibility of expert testimony. It is, however, more easily related to the "reliability" determination embraced by a number of courts⁷⁴ than it is to the general acceptance rule of *United States v. Frye*.⁷⁵ The *Frye* rule forecloses the occasion for the court to examine the reasoning underlying the expert's method; however, it leaves questions such as the applicability of a generally accepted method to a particular case, the way in which a generally accepted method was carried out in a particular case,⁷⁶ and the sufficiency of the evidence to be addressed under other criteria. Thus, even if the United States Supreme Court upholds the application of the *Frye* rule in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,⁷⁷ it will not eliminate the need for courts to actively review scientific expert testimony.⁷⁸

72. But see, e.g., Peter A. Bell, *Strict Scrutiny of Scientific Evidence: A Bad Idea Whose Time Has Come*, Toxics L. Rep. (BNA) 1014 (1992). A more apt comparison would be to the "hard look" doctrine of administrative law, that is, the view articulated by Judge Leventhal that courts reviewing the decisions of a technical agency, such as the Environmental Protection Agency, should review the evidence on which the agency's decision is based to determine "whether the agency decision was rational and based on consideration of the relevant factors." Ethyl Corp. v. EPA, 541 F.2d 1, 34-36 (D.C. Cir.) (en banc), cert. denied, 426 U.S. 941 (1976). Judge Leventhal's views are not without detractors. See *id.* at 66-67 (Bazelon, C.J.). A non-technical, lay jury's decisions would seem to justify greater scrutiny than those of a regulatory agency with technical expertise.

73. See, e.g., *Renaud*, 749 F. Supp. 1545.

74. See *supra* notes 59-60.

75. See *supra* note 55.

76. Cf., e.g., *United States v. Jacobetz*, 955 F.2d 787 (2d Cir. 1992) (the value of DNA testing depends on whether accepted protocols were followed in the specific case), cert. denied, 113 S. Ct. 104 (1992).

77. 951 F.2d 1129 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992).

78. The *Frye* rule at least creates a threshold for evaluation of the evidence that may serve to curb courts' tendencies to uncritically admit all arguably relevant evidence. The reliability standard nonetheless can serve an appropriate screening function if the court actually conducts a reliability analysis.

B. Active Review and Scientific Reasoning

One of the factors that seems to dissuade courts from scrutinizing scientific evidence more carefully is the belief that the differences between scientific and legal inquiries into causation are such that courts are poorly equipped to examine and evaluate science.⁷⁹ Actually, in determining whether there is a link between an event and a later harm, law and science use identical reasoning processes. Differences between scientific and legal institutions, goals and policies, however, obscure that commonality.

Judge Markey has succinctly stated an essential distinction between science and technology on the one hand, and law on the other:

The differences between the judicial and scientific-technological processes are profound and pervasive. Failure to recognize that difference has led to judicial expressions of frustration and an unfortunate tendency to rest judicial decisions on current, often transient, "truths" and "facts" of science and technology. The purpose of science is to learn physical facts. The purpose and function of technology is to provide a means of using that learning. All that is important and necessary, but that's all it is—learning and using physical facts.

The purpose and function of law is to resolve disputes and to facilitate a structure for the organization of a just society—in a word, to provide justice.⁸⁰

As Markey suggests, science and law do differ in important ways. The culture, institutions and processes by which scientific knowledge is developed and refined are very different from those of law.⁸¹ The development of scientific knowledge involves observation, hypothesis

79. See, e.g., *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984). The *Ferebee* court stated:

Judges, both trial and appellate, have no special competence to resolve the complex and refractory causal issues raised by the attempt to link low-level exposure to toxic chemicals with human disease. On questions such as these, which stand at the frontier of current medical and epidemiologic inquiry, if experts are willing to testify that such a link exists, it is for the jury to decide whether to credit such testimony.

Id. at 1534.

80. Markey, *supra* note 42, at 210, quoted in *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 741 (N.J. 1991).

81. See Sheila Jasanoff, *What Judges Should Know About the Sociology of Science*, 32 JURIMETRICS J. 345 (1992); David Kaye, *Proof in Law and Science*, 32 JURIMETRICS J. 313, 317-18 (1992). The discomfort many scientist-experts experience in the adversarial setting of legal adjudication is largely due to scientists' perception that the law requires unequivocal statements on matters that are not clear cut from the scientist's perspective. Further, they are uncomfortable with the legal system's insistence on decisions, often before adequate evidence is available from a scientific perspective. For further discussion of the differences between the processes of legal and scientific inquiry, see Huber, *supra* note 15, at 739-42 (1992).

building, testing, generalizing, and consensus building.⁸² Legal factfinding, on the other hand, is adversarial, confrontational, and directed toward a definitive result in the case at hand. Concern for consistency from case to case plays a lesser role in law⁸³ than in science.⁸⁴

Unfortunately, these institutional and methodological differences obscure the reality that factfinding, that is, science in its broadest sense, is a necessary part of legal decisionmaking. Legal decisionmaking has additional policy components beyond the purely factual, so that it may attach different consequences to the same facts than would a scientist. Thus, the starting point for the analysis of the relationship between science and law on the issue of causation is a delineation of the factual and nonfactual components of legal concepts of cause.

To be sure, causation issues in tort law have nonfactual, policy-laden elements, as exemplified by the legal concept of proximate cause.⁸⁵ All tort theories include some notion of "cause-in-fact" as a prerequisite to liability,⁸⁶ however, and where cause-in-fact is concerned, science and law are attempting to answer the same questions. Further, law, like science, accepts only rational or reasoned findings of fact.⁸⁷ Most importantly, scientific and legal factfinding employ the same logic.⁸⁸

82. See Black, *supra* note 15, at 615-27.

83. In *Wells v. Ortho Pharmaceutical Corp.*, 615 F. Supp. 262 (D. Ga. 1985), *aff'd in part, modified on other grounds*, 788 F.2d 741 (11th Cir.) (modifying damage award), *cert. denied*, 479 U.S. 950 (1986), the court held that plaintiff had proved that her daughter's birth defects were caused by the mother's prenatal use of a spermicide, despite FDA approval and scientific consensus that spermicides do not cause birth defects. *See id.* at 266. But see *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307, 315 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989), *cert. denied*, 494 U.S. 1046 (1990). In *Brock*, the court expressed the hope that its ruling would have "a precedential effect on other cases pending in this circuit which allege Bendectin as the cause of birth defects." *Id.* at 315.

84. That is not to say that science is fixed and unchangeable. Scientific knowledge is always open to revision as new information comes to light that is inconsistent with previously understanding. The point, however, is that scientific reasoning requires that a scientific explanation accommodate and be consistent with all the available data at any point in time.

85. The concept of proximate cause is generally recognized as encompassing policy questions of how closely the defendant's tortious conduct must be related to the plaintiff's injury for the defendant to be held liable. *See, e.g.*, Richard W. Wright, *Responsibility, Risk, Probability, Causation, Naked Statistics and Proof: Pruning the Bramble Bush by Clarifying the Concepts*, 73 IOWA L. REV. 1001, 1011-12 (1988). Viewed in that light, the proximate cause requirement is a limitation on liability where defendant's conduct was the actual cause of plaintiff's injury. *Id.*

86. *See id.* Of course, the way in which the factual question is framed, as well as the burden of proof and evidentiary standards has policy overtones. *See Eggen, supra* note 35, at 899-904 (suggesting shift of burden of proving causation); Nancy L. Firak, *The Developing Policy Characteristics of Cause-in-Fact: Alternative Forms of Liability, Epidemiologic Proof and Trans-Scientific Issues*, 63 TEMP. L. REV. 311, 313 (1990) (arguing that courts' acceptance of epidemiologic evidence is a policy choice rather than a factual conclusion).

87. *See Wright, supra* note 85, at 1011-12; *see also In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1250 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487

Much of the early commentary about the differences between science and law in toxic torts concerned courts' discomfort with statistical evidence of causation. Commentators have attributed that discomfort in part to courts' preferences for mechanistic causal explanations and their reluctance to rely heavily or entirely on statistical evidence.⁸⁹ Courts and lay persons typically think about causal issues in terms of *how* things happen and statistical evidence does not explain how events occur.⁹⁰

When mechanistic thinking about cause is extended to the area of toxic substance disease causation, it immediately encounters a large, perhaps insurmountable, stumbling block. Scientists know very little about how, in a mechanistic sense, toxic substances cause diseases such as

U.S. 1234 (1988). The ubiquitous legal requirement that there exist a rational or reasonable basis for findings of fact evidences the underlying assumption that reasoning and logic must connect evidence to conclusions.

88. David Kaye has demonstrated that science and law use the same logical rules in proving facts. *See Kaye, supra* note 81. He concludes: "[W]hen it comes to proving facts, the logic of law and that of science are one and the same. At an abstract level, the rules of inference can be given the same formal representation." *Id.* at 317; *see also* Lee Loevinger, *Standards of Proof in Science and Law*, 32 JURIMETRICS J. 323, 328 (1992).

In regard to the role of social science in overturning *Plessy v. Ferguson*, Kenneth B. Clark has stated:

The development of science as an approach to the determination of truth involved the development of methods for the control of errors in human observation, judgment, biases, and vested interests. These were the factors which seemed to have distorted man's concept of, or blocked his contact with, the "truth" or "facts" of experience. When they are operative, man's "common knowledge" becomes inconsistent with "scientific knowledge." When they are controlled or for some other reason non-operative, "common knowledge" and "scientific knowledge" are coincident—both reflecting the nature of reality, truth, or facts, as these are knowable to the human senses and intelligence.

Science is essentially a method of controlled observation and verification for the purpose of reducing human errors of observation, judgment, or logic. Science begins with observation and ends by testing its assumptions against experience. It is not a creation of another order of reality. In a very basic sense there cannot be a "legal fact" or a "fact of common knowledge" which is not at the same time a "scientific fact." Whenever this appears to be true, one or the other type of "fact" is not a fact.

Kenneth B. Clark, *The Desegregation Cases: Criticism of the Social Scientist's Role*, 5 VILL. L. REV. 224, 233 (1959).

89. *See e.g.*, Brennan, *supra* note 15, at 478-91. Brennan refers to mechanistic conceptions of cause as "corpuscularianism," after the writings of various philosophers of science. *See id.* at 478-79.

90. The understanding of how a cause produces an effect makes us more comfortable with the conclusion that causation occurred. Richard Wright puts it this way:

Usually, the issue [of proving causation] is what has happened—including how it happened and who did it—although sometimes the issue is what is expected to happen—for example, the expected reduction in future income as an element of damages. That is, proof generally involves either causal explanation or causal prediction.

Wright, *supra* note 86, at 1049.

cancer or injuries such as birth defects.⁹¹ Nonetheless, they may know a considerable amount about whether toxic substances cause disease or injury through inferences drawn from statistical associations and other indirect means.⁹² Thus, the shift in thinking required for courts to come to grips with current scientific knowledge had more to do with abandoning a felt need for an explanatory process that increases comfort with the causal inference than it did with redefining causation.

Courts' discomfort with statistical evidence has gone beyond the absence of mechanistic explanations, however.⁹³ Statistical evidence by definition provides information only about the incidence of disease in groups. Where there are other possible causes of disease, statistical evidence cannot determine which individuals' diseases within the exposed group were caused by background or other factors.⁹⁴ It can only

91. Brennan, *supra* note 25, at 20-25 (discussing scientific evidence of cancer causation).

92. *Id.*

93. See Black & Lilienfeld, *supra* note 37, at 744-50; Brennan, *supra* note 15, at 483-93. Brennan states that courts' refusal to consider and accept statistical evidence reflects and is consistent with courts' traditional reliance on mechanistic causal explanations. See Brennan, *id.* at 491-92.

94. Extensive or complete reliance on epidemiologic proof and other statistical evidence is not without its detractors. See, e.g., Michael Dore, *A Commentary on the Use of Epidemiologic Evidence in Demonstrating Cause-in-Fact*, 7 HARV. ENVTL. L. REV. 429 (1983); Wright, *supra* note 86, at 1049-67 (arguing that particularistic evidence is required to prove actual causation). Dore reiterates the commonly held view that epidemiologic evidence is proof not of actual, individual causation, but only of risk. See Dore, *supra*, at 435. Regarding the use of epidemiology in proving risk (apparently meant as the ability of a substance to cause harm), Dore states:

Within the limitations just discussed, epidemiologic evidence can demonstrate the relative level of risk to which the defendant's activities exposed the members of the plaintiff's group. This risk, of course, does relate to the individual plaintiff. Courts that fail to distinguish the issue of risk from that of actual causation may accordingly, but erroneously, permit the evidence of risk to establish causation. Epidemiologists do not design their studies to resolve issues of individual biological causation, however, and the courts must strictly limit the use of such studies for this purpose.

The limitations on epidemiology's ability to prove individual causation stem from its general and statistical nature. Epidemiologic studies are general in that they deal with sources of disease in groups of people rather than particular individuals. Being statistical, they quantify the probabilities, or risks, that members of a group will contract certain diseases under certain conditions. The only individual cause-and-effect relationship that epidemiologic evidence can show is that the defendant's conduct increased the plaintiff's risk of injury to some statistically measurable extent. *It cannot answer the critical question whether the defendant's conduct actually injured the plaintiff.*

Id. at 436 (citations omitted). Dore and other detractors of statistical evidence frame the question incorrectly, however. The issue in toxic torts is whether there is evidence from which an inference can be made that it is more probable than not that the exposure caused the plaintiff's disease. As others have pointed out, the statistical evidence provided by epidemiology is probative of that issue. See Black & Lilienfeld, *supra* note 37, at 764-69 (combining relative risk with more-probable-than-not standard of proof); Khristina L. Hall

provide an estimate of the likelihood that an individual's disease was caused by the toxic substance in question.⁹⁵ Thus, courts' concerns are not unreasonable. The more likely than not standard of proof, however, implicitly contemplates the marshaling of facts that ultimately prove liability in terms of probabilities.

Uncomfortable with factual indeterminacy, some courts rejected statistical evidence entirely, demanding evidence that is particular to the plaintiff.⁹⁶ Other courts have accepted statistical evidence on issues such as whether a toxic substance is capable of causing harm, but not on the question of whether it caused the plaintiff's harm.⁹⁷ A number of recent cases, however, have recognized the necessarily statistical nature of proof at all levels in toxic torts, and accepted statistical evidence as probative of individual causation, at least where there is evidence indicating a greater than 50% likelihood that the toxic substance caused the plaintiff's disease.⁹⁸ A number of recent decisions evidence a sophisticated understanding of epidemiologic evidence and its relation to legal standards of proof.⁹⁹

The remaining areas where science seems to fit poorly with legal problems are largely the result of failure to distinguish legal standards of proof from factual issues. Courts are concerned that they must decide cases based on the information available, which may not be complete enough to satisfy the requirements of a particular scientific discipline.¹⁰⁰ Some courts perceive scientists as generally requiring higher levels of certainty than does the law.¹⁰¹ That perception may be correct in some instances, particularly in areas such as epidemiology, where standard

& Ellen K. Silbergeld, *Reappraising Epidemiology: A Response to Mr. Dore*, 7 HARV. ENVTL. L. REV. 441, 445-46 (1983). Indeed, signature diseases, which are usually not perceived as presenting difficult individual causation issues, are simply cases in which the statistical evidence is very persuasive because the background incidence of disease is very low compared to the incidence in the exposed population.

95. See *supra* notes 94 and accompanying text.

96. See Brennan, *supra* note 15, at 492 & nn.114-15.

97. See, e.g., *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992) (reversing the trial court's rejection of opinion testimony on individual causation based on epidemiology).

98. See, e.g., *id.* at 1087.

99. See, e.g., *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 946-56 (3d Cir. 1990) (discussing the statistical significance of epidemiologic data); *Landrigan*, 605 A.2d at 1085-87 (discussing the significance of relative risk and attributable fraction).

100. See, e.g., *Ferebee v. Chevron Chem. Corp.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984). The *Ferebee* court held that a treating physician could testify to his opinion that a cause and effect relationship existed between the insecticide paraquat and Ferebee's pulmonary fibrosis even if such a relationship had not been "clearly established" by animal or epidemiologic studies. *Id.* at 1535.

101. In *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 737, 740-41 (N.J. 1991), the New Jersey Supreme Court made several references to the "extraordinarily high level of proof" required by the scientific method. Defendant's witness apparently played into that concern, however unwittingly. See *id.* at 737.

protocols for statistical analysis of relative risk data typically require a 95% level of certainty that an observed increased in risk is not due to chance.¹⁰² Scientists do not require a high degree of certainty for all purposes, however. Risk assessment for purposes of regulation is based on highly uncertain risk estimates. Additionally, scientists often use highly tenuous or uncertain assumptions in making decisions about further research.¹⁰³

The issue of how much uncertainty is acceptable is a legal requirement to be applied to the evidence once the uncertainty attending the scientific evidence is established.¹⁰⁴ Where the law requires the plaintiff to prove her case by a preponderance of the evidence, current standards permit the plaintiff to win if sufficient evidence is available, but not prevail if it is not available.¹⁰⁵ Scientific evidence can be evaluated against those standards, irrespective of whether the scientific discipline would be satisfied or not with the available level of certainty.¹⁰⁶ Moreover, the fact that scientists may require a different level of certainty is not a good reason to dispense with science's requirement of a reasoned analysis, a requirement common to law and science. Unfortunately, some courts throw the baby out with the bathwater by rejecting scientific reasoning altogether when they perceive scientists' requirements for certainty to be too stringent.¹⁰⁷

102. Black & Lilienfeld, *supra* note 37, at 757 n.104; *see infra* notes 364-67 and accompanying text; *see also* DeLuca, 911 F.2d at 946-49 (discussing statistical significance in epidemiology).

103. For example, as discussed in Wells v. Ortho Pharmaceutical Corp., 615 F. Supp. 262 (D. Ga. 1985), *aff'd in part, modified on other grounds*, 788 F.2d 741 (11th Cir.) (modifying damages), cert. denied, 479 U.S. 950 (1986), the "Oeschli study" raised suspicions about the possibility of an association between spermicides and birth defects and recommended further study. *Id.* at 284. Further studies with greater statistical power failed to confirm that suspicion. *See id.*

104. *See* Black, *supra* note 15, at 600 (discussing reliability as a legal question).

105. From that perspective, science and law seem to have parallel requirements because each refuses to reach an affirmative conclusion that causation exists until an acceptable level of certainty is attained, even though the law and the scientific discipline may require different levels of certainty. The relationship between legal and scientific notions of sufficiency of proof is perhaps less clear, however, than is the identity of the logic employed by each. *See* David Kaye, *On Standards and Sociology*, 32 JURIMETRICS J. 535 (1992); Lee Loevinger, *On Logic and Sociology*, 32 JURIMETRICS J. 527 (1992). *Compare* Kaye, *supra* note 81, *with* Loevinger, *supra* note 88.

106. That is not to say that scientists' perceptions of the appropriate level of certainty should be ignored. Requirements such as epidemiologists' practice of requiring a 95% confidence level often have their roots in years of experience in the discipline. With epidemiologic studies in particular, there may be undetected systematic bias in selection of the comparison groups, including the possibility of undetected confounding factors, that are not taken into account in the statistical analysis. *See* ROTHMAN, *supra* note 44, at 89-96; Black & Lilienfeld, *supra* note 37, at 737-38; *infra* text accompanying notes 346-49.

107. *See, e.g.*, Rubanick v. Witco Chem. Corp., 593 A.2d 733 (N.J. 1991), *discussed* *infra* notes 287-300 and accompanying text.

Courts can and should evaluate the underlying reasoning of scientific evidence and measure its reliability or uncertainty against legal standards of sufficiency to meet the applicable burden of proof. The following part of this article attempts to facilitate that process by explicating the bases on which courts can recognize and reject invalid or unreliable evidence, matters on which the differences between science and law are a matter of degree, not kind. Thus, courts need not fear that delving into science and technology will be entirely a foray into alien territory.

IV. ACTIVE REVIEW OF CAUSATION EVIDENCE IN TOXIC TORTS

A. Validity, Reliability, and the Determination of Probative Value

Whether courts operate under the *Frye* rule, the "reliability" standard of *United States v. Downing*,¹⁰⁸ or some other formulation of the rules governing scientific expert testimony, the question courts must answer when they evaluate scientific evidence is, "How probative is it?"¹⁰⁹ That question includes two subissues, however: validity and reliability.¹¹⁰ Validity is the issue of whether the evidence is capable of producing the kind of information sought; thus, it is essentially equivalent to the concept of relevance as used in the rules of evidence.¹¹¹ Reliability connotes the likelihood of a correct or accurate result,¹¹² and thus encompasses notions of certainty or accuracy.¹¹³ Reliability is

108. 753 F.2d 1224 (3d Cir. 1985).

109. So framed, that question corresponds to the determination of "reliability" under *United States v. Downing*. See *supra* notes 59-63 and accompanying text.

110. See Black, *supra* note 15, at 599-600 (discussing validity as part of the reliability determination).

111. FED. R. EVID. 401.

112. See *supra* note 110.

113. This Article thus adopts and expands on the analytical framework proposed by Black, although it uses the terms validity and reliability in a slightly different way. See generally Black, *supra* note 15.

Black defines validity as "that which results from sound and cogent reasoning," and reliability as meaning "that a successful outcome, or correct answer, is sufficiently probable for a given situation." *Id.* at 599-600. Thus, he frames validity as a scientific question, and reliability as a legal one. *Id.* at 600. Validity is to be determined largely by reference to widespread acceptance in the scientific community of the underlying reasoning. *Id.* at 637-38. Black also recognizes, however, that some aspects of the validity analysis relate to the specifics of a particular case that must be examined apart from the test of general acceptance. See *id.* at 657-58 (discussing *Downing* court's evaluation on remand of the applicability of research on eyewitness identification to the facts at hand).

As defined by Black and as used herein, validity is a subissue of reliability, rather than a separate and independent factor, since invalid reasoning or methodology cannot

therefore the ultimate indicator of the probative value and sufficiency of evidence, either alone or in combination with other evidence, to meet the more probable than not standard of proof.

Consider, for example, a diagnostic blood test for a viral blood disease. Without the blood test, the disease can be diagnosed only by elaborate procedures. A simple test is desired for screening large numbers of blood samples for the presence of the virus. A virologist might speculate about any number of parameters that might be indicative of the presence of the virus. None of the possible indicators could be used as a diagnostic test, however, until validated by testing that demonstrates a correspondence between the indicator (a "positive" test) and the presence of the virus. This example illustrates the more general principle that where the physical connections between observed and inferred facts are hidden from direct observation, it is necessary for the inferred connection (e.g., between the indicator and the virus) to be validated through trials or tests that independently measure the properties or characteristics that are ostensibly connected.¹¹⁴

A valid method may nonetheless be insufficiently reliable for evidentiary purposes; that is, the method may be incapable of producing the desired information to an acceptable level of certainty. Using again the example of a test for an asymptomatic virus, the test might have a high rate of false positives or false negatives, or both. Thus, although persons who are test positive are more likely than those who test negative to actually have the virus in their blood, the test may be too inaccurate or unreliable for the purpose for which it is administered.¹¹⁵ Similarly, if the question of whether someone is infected with the virus were a factual

produce reliable or accurate results. *See id.* at 599-606, 613. Reliability is the criterion that courts tend to apply to expert scientific evidence; thus, the proposed analysis fits within recognized criteria for evaluating scientific evidence. *See supra* notes 59-60 and accompanying text.

The definitional structure used herein departs, however, from the usage of the terms validity and reliability in social science research. In social science disciplines, reliability describes the reproducibility of the results and validity describes the degree to which the phenomenon measured corresponds to the phenomenon sought to be measured. Thus, this Article's use of reliability to encompass the accuracy as well as the reproducibility of an outcome encompasses some issues that would be characterized as validity issues under the social science rubric.

114. In toxic torts, validity issues are present when a physician or other expert witness testifies on causation based on patient examination even though there is no clinical basis for linking individual cases to a particular causative agent. *See infra* text accompanying notes 197-205.

115. If, for example, the test were used to determine whether donated blood is safe for transfusions, a significant rate of false negatives would be of much greater concern than a correspondingly high rate of false positives.

question in a legal setting, our hypothetical test might be insufficiently reliable to satisfy the legal standard of proof.¹¹⁶

Validity or reliability questions may arise when methodology that has proved valid and reliable is applied in new circumstances. Invalid application of valid methodology may result from extending a method or line of reasoning to purposes for which it has not been validated.¹¹⁷ In toxic torts, this question arises in connection with whether the conclusions derived from toxicological research on animals or single-celled organisms are applicable to humans.¹¹⁸

Uncertainty or reliability questions may also result from the improper application of valid and reliable methodology. Failure to properly calibrate an instrument such as a breathalyzer, or other concerns related to how a method is applied in a particular case, may increase the likelihood of erroneous results.¹¹⁹ Assume, for example, that in the hypothetical virus test, the incidence of false negatives increases with the length of time that the patient's blood samples are stored before the laboratory test is run. The inferences drawn from a test run by a laboratory that stores its blood samples longer than the optimum time for the test would be subject to a greater variation and uncertainty than results from a laboratory that runs its tests promptly.

A subset of questions regarding "reliability as applied," particularly where the methodology involves calculations from raw data, concerns the quality and quantity of the underlying data. In toxic torts, the data on which estimates of exposure to a toxic substance are based are often sketchy or subject to large uncertainties. Those uncertainties make the inferences of causation that depend on the exposure data similarly uncertain and unreliable.

116. If the question whether someone is infected with a virus were part of the prosecution's proof in a criminal action, that fact would have to be proven beyond a reasonable doubt, so that any significant rate of false positives would likely render it "unreliable" for that purpose. Proof in a civil action would have to satisfy a more probable than not standard, so that a somewhat higher level of false positives, possibly up to 49%, could be tolerated. Courts sometimes refuse to admit evidence that nominally satisfies the applicable standard of proof, however.

117. The issue of the generalizability of a study is characterized as one of external validity. See ROTHMAN, *supra* note 44, at 95-96 (epidemiologic studies).

118. See, e.g., *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1273 (E.D.N.Y. 1985) (questioning the use of human epidemiologic studies of workplace dioxin exposures and animal studies as evidence of effects in Vietnam veterans), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

119. See MCCORMICK ON EVIDENCE § 209, at 513 (Edward W. Clearly ed., 3d ed. 1972) (discussing factual predicate for admitting chemical testing for alcohol intoxication). Courts differ, however, in their approach to whether the manner in which a method is applied goes to the weight of the evidence rather than its admissibility, and is therefore a jury question. See, e.g., *United States v. Jakobetz*, 955 F.2d 786 (2d Cir.) (DNA testing), *cert. denied*, 113 S. Ct. 104 (1992).

B. Validity and Reliability of Causation Evidence in Toxic Torts

Analysis of the kinds of evidence at issue in the typical toxic torts case illuminates many of the problems with such evidence that tend to be obscured when cases are considered as a whole. The following discussion is intended to suggest some ways of looking at such evidence to distinguish probative evidence from junk science; thus, the discussion deals separately with the subissues of the ability of the toxic substance to cause disease (general causation), exposure, and the causation of the plaintiff's disease (individual causation). It would be well to keep in mind, however, that the ultimate question on causation is whether the evidence allows a reasoned conclusion that exposure to the toxic substance in question, rather than other known or unknown factors that also cause such disease, caused the plaintiff's condition.

As noted previously, the characteristics of toxic tort cases impose limitations on the ability to establish causal connections between exposure and disease. The latency periods typical of toxic tort injuries, the absence in most cases of a unique signature injury associated with a toxic substance, the fact that injury does not occur in every instance of exposure, and the absence of clinical indicators that discriminate among causes of a particular individual's disease all tend to obscure toxic injury causation.¹²⁰ In simple terms, the typical toxic tort case looks something like this: The plaintiff believes she has been exposed to a toxic chemical. She has a disease that is commonplace, or at least not unknown, in the general population. The current progress of the disease bears no relation to the continuation of exposure, and the exposure may have long since terminated. There is no diagnostic or clinical test that can determine what caused her disease.

How can such a plaintiff prove that a toxic substance caused her disease? Because of the absence of clinical indicia of cause, the plaintiff must always make out her case indirectly. First, she needs evidence that the substance can cause the condition from which she suffers and of the circumstances under which disease causation is reasonably likely to occur. The coincidence of exposure and disease in the same individual, while necessary, can never be sufficient to prove the capability of the substance to cause disease. Similar problems attend the use of anecdotal case reports or evidence of clusters of disease that have not been subjected to statistical analysis because a certain amount of coincidence and toxic chemical exposure or even clustering of a disease can occur as the result of random chance.¹²¹

120. See *supra* notes 43-49 and accompanying text.

121. Chance may lead to disease clusters rather than disease uniformly distributed throughout a large population. Anecdotal reports and clusters of disease are important in

Second, she needs to establish that she is within the class of persons to which inferences from the general causation evidence should be applied. This second, particularistic causation component of proof, which is discussed later in this article, usually involves two parts: proof of sufficient exposure to permit the inference that the general causation evidence is applicable to her and a demonstration that other causal explanations, including background causes, are less likely causes than the toxic substance exposure.¹²²

1. ABILITY OF THE TOXIC SUBSTANCE TO CAUSE DISEASE (GENERAL CAUSATION)

a. Causal Inferences from Human Disease

On the issue of general causation, systematic studies that can account for the effects of chance are necessary to allow a causal inference to be drawn from data on exposure and disease incidence in humans. Epidemiologic studies, which involve comparisons of disease incidence in exposed and unexposed human populations,¹²³ are based on this line of reasoning. A higher incidence of disease in the exposed population, if parameters of the study are such that the differential rates of disease are unlikely to be due to chance or other confounding factors,¹²⁴ may be indicative of a causal relationship between exposure to the toxic substance and disease.¹²⁵ Scientists have long accepted epidemiologic studies as indicative of causal relationships and courts have more recently begun to do so. Epidemiologic studies are the basis of causation findings

the identification of possible causal links that should be investigated further, however. See Brennan, *supra* note 25, at 21.

122. The issue of whether other causal explanations are less likely is referred to herein as the issue of individual causation.

123. Other commentators have described epidemiologic studies and their relation to proof of causation of human disease. See Black & Lilienfeld, *supra* note 37. See generally 2 DORE, *supra* note 5, §§ 25.01-05. Epidemiologic studies will be described in more detail in the discussion of distinguishing among causes. See *infra* notes 180-90 and accompanying text; see also *infra* notes 342-58, 364-76 and accompanying text (discussing limitations of epidemiology and statistical significance).

124. See 2 DORE, *supra* note 5, §§ 25.02[4], 25.03.

125. Cause is an inference drawn from epidemiologic studies; the studies themselves can only directly prove an association between exposure and disease incidence. Epidemiologists use the Henle-Koch-Evans postulates or other similar premises as criteria for arriving at biological inferences of causation from epidemiologic studies. Black & Lilienfeld, *supra* note 37, at 762-64. The Henle-Koch-Evans postulates are addressed to the magnitude of the risk elevation in the exposed group and other factors tending to increase the plausibility of a biological relationship between the exposure and disease. See *id.*

in asbestos injury claims, and have served as important evidence of the lack of causation in the Bendectin cases.¹²⁶

Epidemiologic studies are expensive to conduct and are subject to a number of limitations on the size of the effect they can detect.¹²⁷ Thus, it is sometimes argued that case reports and clusters of disease constitute sufficient evidence of the capability of a substance to cause toxic injury.¹²⁸ Case reports and disease clusters are sometimes sufficient to raise suspicions and stimulate investigation of toxic chemicals as causative agents.¹²⁹ Benzene was identified as a leukemogenic agent through clinical studies of case reports beginning in the late 1800s,¹³⁰ and vinyl chloride was more recently recognized as carcinogenic through the appearance of clusters of angiosarcoma of the liver in plant workers in the early 1970s.¹³¹ Those examples, however, are typified, in the case of benzene, by very high exposures and the accumulation of evidence over decades,¹³² or by the unexpected appearance of an otherwise very unusual disease.¹³³ Such identifications through case reports and clusters, however, depend on at least a rough sense that the incidence of the disease in the exposed group exceeds the background rates,¹³⁴ even if the reports of unusually high incidence are not initially subjected to the same rigorous statistical analysis as is typical of an epidemiologic study.¹³⁵ Moreover, those initial clusters or unusual case reports will often suggest other places to look for additional evidence, such as workplace exposures involving the same substance, or other users or consumers of the suspect chemical.¹³⁶ The absence of similarly affected individuals among other

126. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307, 313 (5th Cir.) (plaintiffs could not succeed without epidemiologic evidence), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990).

127. See Brennan, *supra* note 25, at 53 & n.228; Green, *supra* note 65, at 653.

128. See *Rubanick v. Witco Chem. Corp.*, 542 A.2d 975 (N.J. Super. Ct. Law Div. 1988), rev'd, 576 A.2d 4 (N.J. Super. Ct. App. Div. 1990), modified, 593 A.2d 733 (N.J. 1991); *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984).

129. See Brennan, *supra* note 25, at 502.

130. JOHN GRAHAM ET AL., IN SEARCH OF SAFETY: CHEMICALS AND CANCER RISK 119-23 (1988).

131. See David D. Doniger, *Federal Regulation of Vinyl Chloride: A Short Course in the Law and Policy of Toxic Substances Control*, 7 ECOLOGY L.Q. 497, 500 (1978).

132. See *supra* text accompanying note 130.

133. Doniger, *supra* note 131, at 500.

134. See GRAHAM ET AL., *supra* note 130, at 119-23 (discussing relative risk estimates derived from clinical studies and epidemiologic studies of disease in benzene-exposed workers).

135. A causal argument based on observation of an otherwise unknown group of symptoms in breast implant recipients may be possible if recent reports of such symptoms are borne out. See Rigdon, *supra* note 6, at B1.

136. The suspicions aroused by the initial reports of angiosarcoma of the liver in B.F. Goodrich's vinyl chloride plant were quickly confirmed by reports from other companies. See Doniger, *supra* note 131, at 500.

populations with similar exposures would suggest that the cluster is a statistical accident rather than a true cluster.

Case reports and apparent disease clusters are likely to be argued in toxic tort cases in circumstances where they do not have even minimal indicia of reliability. In *Renaud v. Martin Marietta Corp.*,¹³⁷ the plaintiffs argued that the existence of four cases of childhood cancer in Friendly Hills, an area in which only two would have been expected, was evidence that the substances allegedly in their water supply had caused their cancers.¹³⁸ Plaintiffs' experts agreed, however, that the Friendly Hills population was too small to yield meaningful results. Moreover, another expert's opinion was that four cases of childhood cancer was within the expected range for the community.¹³⁹

b. Animal Studies¹⁴⁰

Animal studies, other biological assay methods and chemical structure-activity relationships, all of which are used by toxicologists to estimate human risk from toxic chemicals,¹⁴¹ are much more problematic than epidemiologic studies in the toxic tort context. The use of such methods in risk regulation is based on unproven assumptions about the applicability of the results of such studies to humans, assumptions that are subject to a large degree of uncertainty and in some cases skepticism in the scientific community.¹⁴²

Animal studies are based on the theory that substances that cause harmful effects in animals are likely to cause similar harmful effects in humans.¹⁴³ That thesis is supported by observations that many substances that cause harmful effects in one species also cause harmful effects in other species.¹⁴⁴ All but one of the chemicals identified by epide-

137. 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

138. *Id.* at 1554-55.

139. *Id.* at 1551. Dr. Steven Piantodosi's epidemiological study on the incidence of cancer in children in Friendly Hills indicated that the difference between expected and observed incidence rates was not statistically significant. *Id.*

140. Animal testing has been treated as either admissible or inadmissible. See Jack L. Landau & W. Hugh O'Riordan, *Of Mice and Men: The Admissibility of Animal Studies to Prove Causation in Toxic Tort Litigation*, 25 IDAHO L. REV. 521 (1988-89). As is discussed *infra* notes 151-68 and accompanying text, the issue of animal testing should be addressed as one of how probative is animal testing of causation of human disease, that is, as a question of sufficiency rather than of relevance.

141. Brennan, *supra* note 25, at 21-23, 44. See generally Chemical Carcinogens: A Review of the Science and Its Associated Principles, 50 Fed. Reg. 10,371 (Office of Science & Technology Policy 1985) [hereinafter OS&TP, Chemical Carcinogens].

142. Brennan calls these issues "trans-scientific." See Brennan, *supra* note 25, at 23. Such issues are not always inherently unprovable, although it may be impractical to do so.

143. See Brennan, *supra* note 15, at 504-06.

144. See, e.g., James E. Huff & Joseph K. Haseman, *Exposure to Certain Pesticides May Pose Real Carcinogenic Risk*, CHEMICAL & ENGINEERING NEWS, Jan. 7, 1991, at 33, 34

miologic studies as causing cancer in humans have also proven to be carcinogenic in one or more animal species.¹⁴⁵ Thus, there appears to be some correlation between carcinogenicity in animals and carcinogenicity in humans. Similar observations and findings have been made with respect to other kinds of toxic effects, including teratogenic effects.¹⁴⁶

As any observer of the popular media knows, however, animal testing for diseases such as cancer, which has a long latency periods, and for which even low incidence rates are of concern, are conducted under conditions that are very different from the usual human exposure scenario.¹⁴⁷ Animal studies of carcinogenicity typically utilize doses at or near the maximum level tolerated by the animal.¹⁴⁸ That practice is necessitated by the need to detect effects in relatively small groups of test subjects, in a relatively short period of time. Those same concerns also have led to protocols using animal strains bred for their susceptibility for tumor formation.¹⁴⁹ Additionally, the route of administration may differ from the likely human exposure route.¹⁵⁰

The prediction of effects in humans from animal testing involves a number of extrapolations—from animal species to humans, from one route of administration to another, and most acutely, from a high-dose exposure in which the animals are typically subjected to the maximum dose they can tolerate (the MTD),¹⁵¹ to a low-dose chronic exposure.¹⁵² Each of those extrapolations introduces uncertainty into the predictive value of animal testing in proving causation of human disease.¹⁵³

(reporting that information on carcinogenicity of 8 of 54 known carcinogens was first obtained in animal studies).

145. OS&TP, Chemical Carcinogens, *supra* note 141, at 10,411; J.F. Robens et al., *Methods of Testing Carcinogenicity*, in PRINCIPLES AND METHODS OF TOXICOLOGY 251, 253 (A. Wallace Hayes ed., 2d ed. 1989).

146. See Jeanne M. Manson & L. David Wise, *Teratogens*, in CASARETT AND DOULL'S TOXICOLOGY 226, 240 (Mary O. Amdur et al. eds., 4th ed. 1991) [hereinafter CASARETT & DOULL].

147. Ames, *supra* note 29, at 589; see Landau & O'Riordan, *supra* note 140, at 545.

148. OS&TP, Chemical Carcinogens, *supra* note 141, at 10,377; Bruce N. Ames & Lois S. Gold, *Cancer Prevention Strategies Greatly Exaggerate Risks*, CHEMICAL & ENGINEERING NEWS, Jan. 7, 1991, at 28, 29; see also Kent R. Stevens & Michael A. Gallo, *Chronic Toxicity Studies*, in PRINCIPLES AND METHODS OF TOXICOLOGY, *supra* note 145, at 237, 238-39.

149. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,377.

150. Animal testing may involve skin application, oral gavaging or injection, *see id.* at 10,413-14, rather than the usual human exposure routes of inhalation, ingestion or dermal contact. *See id.*

151. See Ames & Gold, *supra* note 148, at 29. Test animals such as rodents live only one to two years, though they may receive test doses throughout the majority of their lifespans. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,413, 10,414.

152. See Landau & O'Riordan, *supra* note 140, at 543-48.

153. The use of animal test results as proof of toxic effects in humans can be regarded as raising validity issues, because it is questionable whether results in one species can be extrapolated to another at all. See Black, *supra* note 15, at 677-79; Green, *supra* note 65, at 654-56. This issue is treated here as one of uncertainty or inaccuracy, however, because

Differences in species can have a dramatic impact on the effects of a toxic substance,¹⁵⁴ as can routes of administration.¹⁵⁵

The high dose exposure scenario of typical animal testing protocols raises several concerns. One concern relates to the model used to extrapolate the results of high dose exposures to the much lower doses encountered by humans. The lack of a complete mechanistic understanding of cancer causation precludes the adoption of any particular extrapolation model with a high degree of certainty.¹⁵⁶ For example, one

even if validity is assumed, the uncertainty attending extrapolation of results from animal studies to humans will usually render them insufficiently probative to support a plaintiff's verdict. Additionally, most scientists regard animal studies as having some validity in predicting human disease, and such studies are widely used for regulatory purposes. See OS&TP, *Chemical Carcinogens*, *supra* note 141. Animal studies vary in their predictive value for humans, however, depending on the nature of the effects being studied and the number of species in which toxic effects of a substance have been confirmed. Models that may improve predictive capabilities are being developed for quantitative interspecies extrapolations. See Robert A. Scala, *Risk Assessment*, in *CASARETT & DOULL*, *supra* note 146, at 985, 993 (discussing potency correlations of animal and human carcinogens). Thus, it seems appropriate to address the extrapolation of animal test results to humans as an issue of the degree of certainty that attends that extrapolation in a given instance, rather than to assert as a general proposition that animal tests results can or cannot in any instance be validly extrapolated to humans. *Id.*

154. Animal testing is of limited value in a context where false positives are a concern, as is the case with toxic torts, a generalization that cuts across the various types of effects for which such studies are conducted. For example, Manson and Wise report that of 38 substances with positive teratogenic findings in humans, only one was negative in all animal species studied, thus producing a low rate of "false negatives" for human teratogenicity. Manson & Wise, *supra* note 146, at 240. In contrast, of 165 substances studied with no teratogenic finding in humans, only 47, or 29%, were negative in all laboratory animal test species. *Id.* Similar uncertainties occur in animal testing for carcinogenicity. Although all but a few of the 30 or so known human carcinogens (substances or industrial processes) are also carcinogenic in at least one animal species, there are many more substances that have exhibited carcinogenicity in animals that are not known to be human carcinogens. *Id.* It is not uncommon for a substance to exhibit carcinogenicity in one species and not in another. See Scala, *supra* note 153, at 992. For example, in one study of almost 1000 chemicals, only 76% of rat carcinogens were positive in mice, and 70% of mouse carcinogens were positive in rats. *Id.* Studies of carcinogenic potency of the same substances in animals and humans have yielded good correlations for some substances, but animal data overpredicts human response by as much as a factor of 500 for vinyl chloride. See Landau & O'Riordan, *supra* note 140, at 536 (citing Michael D. Hogan and David G. Hoel, *Extrapolation to Man*, in *PRINCIPLES AND METHODS OF TOXICOLOGY*, *supra* note 145, at 879). That difference may be due to incomplete data on human cancer, but it cannot be ignored.

155. For example, EPA's "level of regulatory concern" for inhalation of chromium has been stated as 1.9×10^{-6} mg/day, as compared to 0.1 mg/day for exposure by ingestion. See, e.g., Final Exclusion, 53 Fed. Reg. 29,038, 29,040-41 (EPA 1988) (tbls. 1 & 2) (evaluating petition for delisting of hazardous waste). The inhalation level of regulatory concern was thus set 50,000 times lower than the ingestion level, apparently due to demonstrated respiratory tract carcinogenicity of inhaled chromium as compared to lower risks through other routes of exposure. See Robert A. Goyer, *Regulatory Toxicology*, in *CASARETT & DOULL*, *supra* note 146, at 623, 639.

156. See generally Robert A. Scala, *Risk Assessment*, in *CASARETT & DOULL*, *supra* note 153, at 985, 990-91.

possible set of assumptions is that no dose of a carcinogen is completely risk free and that the disease incidence rate will be directly proportional to the dose. Those assumptions lead to a linear extrapolation model.¹⁵⁷ Another possibility, which apparently applies to some carcinogens, is that at very low levels, a toxic chemical exerts no adverse effects and that such effects appear only when a threshold level of exposure is exceeded.¹⁵⁸ The set of assumptions adopted in a particular instance can lead to vastly different predictions of the effects of low dosage exposures, sometimes as much as several factors of ten.¹⁵⁹

The accuracy of risk extrapolations from exposure of animals to the MTD has recently been called into further question by prominent researchers in the field of carcinogenesis.¹⁶⁰ Bruce Ames, the developer of the "Ames test" for mutagenicity,¹⁶¹ now argues that risk estimates obtained under such circumstances are largely due to toxic effects of the test chemical, rather than factors that might operate at lower doses in human.¹⁶² Thus, the results from animal studies may not be predictive of human carcinogenicity under the usual exposure scenario.

Despite the sparse knowledge of mechanisms of cancer causation, toxicologists have identified a number of steps in the carcinogenesis process, including DNA alteration, DNA expression, and promotion and progression to neoplastic or cancerous tumors. They have also identified two general classes of carcinogens. See Gary M. Williams & John H. Weisburger, *Chemical Carcinogenesis*, in CASARETT & DOULL, *supra* note 146, at 127, 129-31, 170-85. DNA-reactive carcinogens are those that appear to initiate cancer through chemical alteration of DNA. *Id.* at 170. Epigenetic carcinogens, on the other hand, do not necessarily react with DNA and exert carcinogenic effects through other pathways such as by promoting the growth of dormant cancer cells. *Id.* at 185.

157. See Scala, *supra* note 153, at 990-91 (discussing a linearized, multi-stage, nonthreshold model); see also OS&TP, Chemical Carcinogens, *supra* note 141, at 10,438-39. This model assumes that there is no threshold dose below which cancer does not occur, but recognizes the multi-step nature of carcinogenesis. See *id.* Threshold doses are common for acute effects of toxins, i.e., those that occur within a short time after exposure. See Curtis D. Klaasen & David L. Eaton, *Principles of Toxicology*, in CASARETT & DOULL, *supra* note 146, at 12, 38. However, whether carcinogens are subject to thresholds is an unresolved issue. See Scala, *supra* note 153, at 990-91. The choice of model can make a difference of several orders of magnitude in risk levels. As a matter of policy, the conservative linearized multi-stage extrapolation model is often used to estimate the upper bound on risk. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,438-39.

158. See Williams & Weisburger, *supra* note 156, at 154. This assumption appears a likely model for carcinogens classed as promoters rather than as initiators. See David J. Hanson, *Dioxin Toxicity: New Studies Prompt Debate, Regulatory Action*, CHEMICAL & ENGINEERING NEWS, Aug. 12, 1991, at 7, 13 (EPA reconsidering model for dioxin carcinogenicity low-dose extrapolation to allow for threshold).

159. See *supra* note 158; see also OS&TP, Chemical Carcinogens, *supra* note 141, at 10,439.

160. Ames & Gold, *supra* note 148, at 29.

161. See *supra* note 6 and accompanying text.

162. See *supra* note 139 and accompanying text; see also Bruce N. Ames et al., *Ranking Possible Carcinogenic Hazards*, 236 SCIENCE 271 (1987). The authors believe the toxicity effect to be particularly true for carcinogens that are not DNA reactive. See Bruce N. Ames & Lois S. Gold, *Too Many Rodent Carcinogens: Mitogenesis Increases Mutagenesis*, 249 SCIENCE

Quite a number of toxic torts plaintiffs have offered animal studies in support of their contentions that the substances in question can cause harm in humans.¹⁶³ In some cases courts have been willing to entertain such evidence,¹⁶⁴ while others have found it inadmissible¹⁶⁵ or insufficient.¹⁶⁶ The closer examination of the assumptions and methodologies involved in animal testing, however, reveals that the extrapolation of animal test results to humans is too uncertain—the potential for error is too high—for animal testing alone to support an inference that it is more probable than not that a substance causes cancer or birth defects in humans at a specified level of exposure.¹⁶⁷ There is considerable doubt about the inference that an animal carcinogen is a human carcinogen at all. Even if that hurdle is assumed away, however, the uncertainties that attend the interspecies and high-dose to low-dose extrapolations necessary to extend animal test results to human exposure scenarios are simply too large. Policy considerations in the regulatory arena dictate or at least support the use of models that overpredict rather than underpredict risks levels.¹⁶⁸ Risk estimates based on unproven dose-response models, where the choice of model may alter results by a thousand times or more, however, are not consistent with a more likely than not standard of proof.

c. Biological Screening Methods

Even greater problems attend the application of biological screening methods such as short term assays. Short term assays are designed to detect mutagenic effects and cancer-initiating or promoting properties of

970 (1990). Ames and coworkers direct their argument to the allocation of scarce resources for cancer prevention purposes. Their concern about the predictive value of animal testing and other protocols for determining carcinogenicity bears on the general causation issue in toxic torts, however.

163. See, e.g., *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D.Colo. 1990), aff'd, 972 F.2d 304 (10th Cir. 1992); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985) (Agent Orange and dioxin), aff'd, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988).

164. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991).

165. See, e.g., *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985), aff'd, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988).

166. See, e.g., *Brock*, 874 F.2d 307.

167. In conjunction with positive epidemiologic studies, positive animal results may be relevant. In such instances, however, some would argue that such evidence is cumulative and of such low probative value as to warrant its exclusion. See, e.g., Landau & O'Riordan, *supra* note 67, at 551-54.

168. Douglas G. Camp et al., *Pesticide Regulation Is Sound*, CHEMICAL & ENGINEERING NEWS, Jan. 7, 1991, at 44, 46.

substances.¹⁶⁹ Mutagenicity,¹⁷⁰ for example, is used as a predictor of carcinogenicity because many known carcinogens are also mutagens.¹⁷¹ Mutagenicity is also considered to be an indicator of potential for causing birth defects.¹⁷² Short term assays' predictive capabilities have been validated by reference to animal carcinogenesis,¹⁷³ however, so they are subject to all of the uncertainties of animal testing, as well as additional uncertainties introduced by the procedure itself.¹⁷⁴ Moreover, because such tests involve single-celled organisms, they are less likely than *in vivo* testing in animals to represent the response of humans.

Another kind of evidence, typically offered as evidence of causation of birth defects or other noncancerous disease or injury, is *in vitro* testing, tests involving exposure of isolated groups of cells or organs¹⁷⁵ to suspect chemicals. To test for teratogenesis (birth defects), fetal cells or embryos may be used. These tests are fraught with uncertainties, however, related to whether and to what degree the chemical in question would reach or react with the sensitive cells or organs in a whole organism.¹⁷⁶ They are also fraught with the same uncertainties relating to interspecies extrapolation as is animal testing generally.

d. Chemical Structure-Activity Analysis

Chemical structure-activity analysis is a kind of scientific reasoning by analogy that is based on the recognition that similarities in chemical structure sometimes correspond to similarities in biological activity.

169. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,403.

170. Mutagenicity refers to the alteration of the genetic material of a cell.

171. Ames, *supra* note 29, at 589.

172. *Id.* at 587.

173. *Id.* at 588.

174. Single-celled organisms are, of course, farther removed biologically from humans than are mammals such as mice and rats which are typically used in animal testing. See OS&TP, Chemical Carcinogens, *supra* note 141, at 10,404 (listing commonly used assays). Short-term assays are utilized to select substances for chronic (i.e., long-term) animal testing. See *id.* at 10,408. The OS&TR review states the utility of short-term assays succinctly:

Short-term tests are presently limited in their ability to predict the presence or absence of carcinogenicity and cannot supplant data from long-term animal studies or epidemiologic data since the tests do not necessarily screen for all potential means of cancer induction and do not necessarily mimic all reactions that would occur *in vivo*.

Id. at 10,376.

175. In *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989), *cert. denied*, 494 U.S. 1046 (1990), the plaintiff's evidence included limb bud tests as evidence of teratogenicity of doxylamine, the active ingredient of Bendectin.

176. See *id.* at 314 (discussing the possibility that doxylamine breaks down in the human body and does not reach limb buds in unaltered form).

Observations based on such reasoning are the impetus of much drug research, as well as research on the hazardous effects of chemicals.¹⁷⁷ The relation of chemical structure to biological activity is highly uncertain, however, at least where the effects of only one or two compounds similar to the one in question are known.¹⁷⁸ No two chemicals have the same structure, so the question is always whether the similarities are more important than the differences in predicting toxicological properties. Structure-activity analysis is used primarily to select candidates for additional study. In most cases, reasoning based on structure-activity relationships will fall far short of the reliability required to satisfy a more probable than not standard of proof.

e. The Insufficiency of Animal Test Results, Short-term Assays,
In Vitro Testing and Structure-Activity Relationships to Prove
General Causation.

As can be seen from the foregoing discussion, animal test results, biological screening methods and chemical structure-activity relationships are insufficiently reliable, even if arguably valid, to permit inferences to be drawn that it is more probable than not that a substance can cause a disease in humans. It is also important to understand that even when all such indicators are positive, they are still insufficient for that purpose, given the present state of science.

Toxicological research into the causes of human disease, when direct evidence in humans is unavailable, proceeds according to a hierarchy of reasoning, from the least costly and time-consuming, and least predictive methods, to the most costly, time-consuming methods

177. See Williams & Weisburger, *supra* note 156, at 156-57 (discussing chemical structure-activity analysis as the first step in carcinogenicity assessment).

178. Saffiotti summed up the state of predictions from structure-activity analysis as follows:

There is a moderately substantial base of empirical data that permits conclusions about carcinogenic potential on the basis of molecular structure, *at least on the basis that* certain groupings of atoms (functional groups) in some molecules may impart carcinogenic properties. *The predictive power of such correlations has, however, been unsatisfactory so far, and the general consensus of the scientific community appears to be that chemical structure has limited value in identifying carcinogens and is to be used in carcinogenic hazard assessment only as corroborative supporting evidence.*

Umberto Saffiotti, *Identification and Definition of Chemical Carcinogens: Review of Criteria and Research Needs*, 6 J. TOXICOLOGY & ENVTL. HEALTH 1029, 1043 (1980).

It is unusual for all chemicals in a class to be carcinogenic, or for all carcinogenic members of a class to be equally potent. When 60 structural analogs of thalidomide were studied for teratogenicity, only three were found to exhibit that property. Manson & Wise, *supra* note 146, at 228. Structure-activity relationships are used in toxicology research primarily to select candidates for short-term assays and other more extensive tests. See, e.g., Williams & Weisburger, *supra* note 156, at 156-67.

that are believed to correspond most closely to human response. Thus, the investigation of the toxicological properties of a chemical is likely to start with the analysis of available information about chemicals with similar structures—chemical structure-activity analysis.¹⁷⁹ The most likely candidates identified by structure-activity analyses are then subjected to biological assays such as mutagenicity testing or *in vitro* testing on cell groups. Lastly, animal testing will likely be conducted on chemicals that exhibit toxic effects in the short-term screening procedures. Structure-activity analysis and short-term screening are not the end points of the evaluation process, even in a regulatory context, because they are recognized as significantly less valid and reliable than animal testing. Whether considered separately or in the aggregate, the methods that do not involve observations of disease in humans are too likely to lead to an erroneous conclusion to satisfy the traditional burden of proof.

2. CAUSATION OF PLAINTIFF'S DISEASE (INDIVIDUAL CAUSATION)

Even where the capability of a substance to cause disease can be shown, the plaintiff will still need to prove that the toxic substance caused his disease. The primary difficulty facing such a plaintiff is the need to differentiate between the exposure and background causes as explanations of the injury. Much speculation masquerading as science appears in connection with this issue.

To understand what kinds of evidence are probative of individual causation, we must first make reference to the kinds of evidence probative of the capability of the substance to cause harm, namely, epidemiologic evidence or possibly other human evidence of sufficient reliability. That evidence will identify one or more diseases that are believed to be causally associated with exposure to a toxic substance. Such studies will also typically be based on or identify certain levels or ranges of exposures. The plaintiff must argue that the association established through the statistical study applies to him and that background and other risk factors are less likely causal explanations.

a. Epidemiologic Reasoning

The best case for the plaintiff is the situation in which an epidemiologic study has identified an association between exposure to a toxic substance and a disease. For example, a number of studies have shown an association between asbestos exposure and lung cancer. The study will produce an estimate of the increased incidence of disease

179. See Williams & Weisburger, *supra* note 156, at 156-57 (describing the decision point approach to carcinogen testing).

associated with the exposure, such as relative risk.¹⁸⁰ Relative risk, illustrated by the formula below, is the ratio of disease incidence in the exposed population to disease incidence in the unexposed population in the study.¹⁸¹

$$\text{Relative Risk} = \frac{\text{incidence in exposed group}}{\text{incidence in unexposed group}}$$

Black provides an example in which the disease rates in exposed and unexposed groups are 50 and 5 per 100,000 population respectively.¹⁸² The relative risk in that case is 50/5 or 10, indicating that the exposure increases the disease rate to ten times that of the background rate.

How can a toxic tort plaintiff use such information? At a minimum, it would seem obvious that the plaintiff must have a disease identified as associated with the toxic chemical exposure. Nonetheless, it is not uncommon for plaintiffs' experts to assert that evidence that a substance causes any cancer is evidence that it can and has caused other cancers.¹⁸³ Although substances that are discovered to cause one type of cancer may cause other types of cancer as well, that possibility does not permit a prediction of what those other cancers, if any, are likely to be.¹⁸⁴ The

180. See Black & Lilienfeld, *supra* note 37, at 758.

181. *Id.* at 758 & n.105. Relative risk estimates can also be generated from case control studies, which compare the incidence of exposure in cases and controls that do not exhibit the disease in question. See ROTHMAN, *supra* note 44, at 63-64.

182. See Black & Lilienfeld, *supra* note 37, at 758 & n.105.

183. In *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988), the opt-out plaintiffs alleged, based in part on animal tests, that Agent Orange exposure had caused a number of different kinds of cancer, including Hodgkin's disease and cancer of the ileum, chronic skin rashes, infertility, *id.* at 1239, 1252-54, gastrointestinal disorders, miscarriages and other birth defects, *id.* at 1231, various behavioral disorders, including memory loss, increased irritability, anger, depression and others, weight loss, various liver disorders, including abnormal liver function, hepatitis and cirrhosis, and elevated triglycerides and cholesterol, *id.* at 1235-36. Plaintiffs relied on animal and workplace exposure studies to link Agent Orange or its contaminant, dioxin, to their injuries. *Id.* at 1236. The court noted, however, that plaintiffs' liver injuries differed "substantially" from those reported in the studies. *Id.* at 1236. Plaintiffs' reports of skin rashes or chloracne many years after exposure were also inconsistent with the immediate and transient relationship of chloracne and dioxin exposure. *Id.* at 1260.

Renaud v. Martin Marietta Corp., 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992), also typifies such cases. Plaintiffs alleged that a number of injuries, including childhood cancers (one case of leukemia), kidney cancer, seizure disorders, and congenital heart defects resulted from exposure primarily to hydrazines, also classified as animal carcinogens. *Id.* at 1547. Plaintiff also alleged exposure to several other chemicals. *Id.*

184. Nor does it suggest what the relative risk ratios would be for this possible, but unproven disease. This question is thus one of both general and individual causation.

problem is compounded when the plaintiff's general causation evidence is not based on human evidence, but on animal studies or other less reliable methods that only generally suggest a possible carcinogenic effect;¹⁸⁵ in such cases, the evidence may not allow an identification of any particular disease associated with the substance.

If plaintiff has a disease associated with the toxic substance exposure, demonstrating that it is more likely than not that plaintiff's condition was caused by exposure usually will require the plaintiff to demonstrate that her exposure was in the range found to be associated with an increased risk of disease.¹⁸⁶ Alternatively, plaintiff might be able to show that the results of the study could be extrapolated to lower doses.¹⁸⁷

If the plaintiff can demonstrate sufficient exposure to argue that the relative risk factor in the study applies to him, he still must differentiate background causes where the disease occurs in the background population. Under the traditional more likely than not rule, the plaintiff will prevail if the relative risk identified in the epidemiologic study is greater than two. That is because relative rate greater than two corresponds to more than doubling of the disease rate, permitting the inference that more than half of the cases of the disease in the exposed population were caused by the exposure. When applied to the plaintiff, the inference can be made that it is more likely than not that the exposure caused the

185. Such was the problem in *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991), discussed *infra* text accompanying notes 250-85. See also *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733 (N.J. 1991), discussed *infra* text accompanying notes 287-300.

186. Exposures are often at issue. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). Much of the evidence at issue in *Paoli* concerned whether plaintiffs' PCB exposures had exceeded normal background levels in the general population. See *id.* at 860-61; see also *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545, 1555 (D. Colo. 1990) (dismissing suit because of plaintiffs' inability to present evidence of exposure establishing a *prima facie* case), aff'd, 972 F.2d 304 (10th Cir. 1992). Proof of exposure is discussed *infra* notes 214-49 and accompanying text.

187. It is widely accepted that dose-response relationships exist for toxic substances, that is, that disease incidence increases with increased exposure. The existence of a dose-response relationship is considered to be evidence that the association of toxic substance and disease incidence is causal. See *Black & Lilienfeld*, *supra* note 37, at 762-63 (discussing the Henle-Koch-Evans postulates).

High-dose to low-dose extrapolations based on epidemiologic data are subject to some of the same limitations as those discussed in connection with animal studies. See *supra* notes 151-59 and accompanying text. At some point, the application of epidemiologic studies to persons whose exposure levels were very different from those in the study raises a validity issue. This issue arises in cases where plaintiffs offer epidemiologic evidence based on relatively high occupational exposures as probative of the effects of lower level environmental exposures. See, e.g., *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1241 (E.D.N.Y. 1985) (rejecting epidemiologic studies based on industrial exposures), aff'd, 818 F.2d 187 (2d Cir. 1987), cert. denied, 487 U.S. 1234 (1988).

plaintiff's disease.¹⁸⁸ Put another way, when the relative risk is greater than two, the fraction of all disease in the exposed group attributable to the exposure is greater than 50%.¹⁸⁹ The causal connection inferred from the presence of a signature disease represents the application of this principle when the relative risk is very large and the corresponding risk attributable to the exposure may be ninety percent or more.¹⁹⁰

b. Mechanistic Explanation

When the plaintiff's exposure evidence is weak, and when general causation evidence is based on animal studies, *in vitro* testing and the like, experts may attempt to bolster the plaintiff's case through speculation in the guise of a mechanistic explanation of causation.¹⁹¹ For example, an expert may offer testimony about the "one hit" theory of causation, namely, the theory that cancer can be caused by only one molecule of the

188. See Black & Lilienfeld, *supra* note 37, at 767-69. Relative risks greater than two are the basis of causation findings in cases involving claims for lung cancer from asbestos exposure. See *infra* notes 212-15 and accompanying text. The principle has been cited with seeming approval in a number of cases. See, e.g., DeLuca v. Merrell Dow Pharmaceuticals, Inc., 911 F.2d 941, 958-59 (3d Cir. 1990).

189. Attributable risk can be viewed as the proportion of a disease that is statistically attributable to a risk factor. Black & Lilienfeld, *supra* note 37, at 760-61. The attributable risk in an exposed population can be calculated as follows:

$$\text{Attributable Risk} = (\text{Relative Risk} - 1) / (\text{Relative Risk})$$

Where the relative risk is 2.0, attributable risk or attributable fraction is $(2.0-1.0)/2.0$, or 0.5. See Black & Lilienfeld, *supra* note 37, at 761 & n.123.

The reasoning illustrated in the text accompanying this note can result in recovery by 100% of exposed persons with the disease where the relative risk is greater than two, even though up to 50% would almost certainly have contracted the disease without exposure. Further, when there is a relative risk less than 2.0 but greater than 1.0, no plaintiffs will recover even though the epidemiologic study indicates causation of a group constituting less than half the cases. The unfairness of such results has led commentators to suggest that when an epidemiologic study indicates any increased risk, all exposed individuals with the indicated disease should recover proportionately to the magnitude of the increased risk. See, e.g., David Rosenberg, *supra* note 32; cf. Robinson, *Probabilistic Causation*, *supra* note 50, at 783 (recommending compensation for risk of future disease); Gregory L. Ash, Note, *Toxic Torts and Latent Diseases: The Case for an Increased Risk Cause of Action*, 38 KAN. L. REV. 1087, 1102-03 (1990).

190. For example, the relative risk of mesothelioma for asbestos exposure is on the order of 46, see Brennan, *supra* note 25, at 39 n.166 (citing A.D. McDonald & J.C. McDonald, *Malignant Mesothelioma in North America*, 46 CANCER 1650 (1980)), resulting in an attributable risk of over 97%.

191. In Brock v. Merrell Dow Pharmaceuticals, Inc., 874 F.2d 307 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989), *cert. denied*, 494 U.S. 1046 (1990), Dr. McBride, one of plaintiff's experts on the issue of whether Bendectin causes limb defects, testified about his theory of how Bendectin could cause such defects. *Id.* at 314-15. The court characterized the doctor's theory of causation as "nothing more than unproven medical speculation lacking any sort of consensus." *Id.*

toxic substance acting on only one cell, which then becomes cancerous.¹⁹² The witness may then explain that cancerous changes are thought to involve chemical alteration of DNA, the genetic material of cells, alterations that can be brought about by interaction with toxic chemicals.¹⁹³ Sometimes the witness expounds on the one hit theory by explaining that even very low concentrations of toxic chemicals involve exposure to trillions of molecules, and thus many opportunities for cancerous or mutagenic changes.¹⁹⁴ This kind of argument has much superficial appeal because it represents a common line of reasoning in cancer research or risk assessment.¹⁹⁵ When offered in proof of the likelihood that a low-level exposure, rather than background factors, caused the plaintiff's disease, however, the one hit theory and its corollaries amount to nothing more than speculation. The validity of the theory in a given case will rarely have been tested. Moreover, unless a mechanistic explanation offers a way to distinguish between causation by the toxic substance and causation by background factors,¹⁹⁶ it adds nothing to the proof of the plaintiff's case. Mechanistic explanations are not a substitute for statistical information such as epidemiologic studies when the background rate of the plaintiff's disease is significant.

192. See, e.g., *Peteet v. Dow Chem. Co.*, 868 F.2d 1428, 1433 (5th Cir.), *cert. denied*, 493 U.S. 935 (1989). A variant of this line of reasoning, mechanistically related to it, is the assertion that there is no safe level of a carcinogen, that is, that there is no threshold dose below which cancer does not occur. *See supra* notes 157-59 and accompanying text; *see also* *Sterling v. Velsicol Chem. Corp.*, 647 F. Supp. 303, 482-83 (W.D. Tenn. 1986), *aff'd in part, rev'd in part*, 855 F.2d 1188 (6th Cir. 1988); *Rubanick v. Witco Chem. Corp.*, 576 A.2d 4, 12 (N.J. Super. Ct. App. Div. 1990), *modified*, 593 A.2d 733 (N.J. 1991);. The U.S. Supreme Court discussed the one-hit theory and its relation to the issue of threshold exposures in *Industrial Union Department, AFL-CIO v. American Petroleum Institute*, 488 U.S. 607, 636 & n.41 (1979).

193. See, e.g., 2 Transcript of Hearing at 191-95 (testimony of Dr. Marvin Legator), *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (Civ. A. No. 87-Z-42), *aff'd*, 972 F.2d 304 (10th Cir. 1992); *id.* at 275-79 (testimony of Dr. David Ozonoff).

194. *Id.*

195. See generally OS&TP, Chemical Carcinogens, *supra* note 141, at 10,387-88, 10,438; *see also* *supra* notes 157-59. Cancer researchers recognize that DNA alteration represents only one mode of chemical carcinogenesis. Some carcinogens have been demonstrated not to react with DNA. These are generally grouped under the general heading of epigenetic carcinogens, which include promoters, that is, agents that facilitate the growth of dormant cancer cells into tumors. *See Williams & Weisburger, supra* note 156, at 185-86. These carcinogens typically require high doses and sustained exposure to exhibit carcinogenicity. *Id.* at 185.

The "one hit" theory is consistent with the linear extrapolation model or no-threshold model discussed *supra* notes 157-59 and accompanying text. The appeal of this kind of reasoning is also fundamentally related to the desire to understand *how* causation occurs.

196. Nor is even a proven mechanism likely to provide much information on the likelihood that an exposure caused disease unless the mechanistic explanation leads to a clinical test capable of distinguishing toxic chemical exposure from background or other causes.

c. Medical Opinion Evidence

Plaintiffs often offer medical opinion evidence on individual causation; indeed, courts sometimes express a preference for testimony by a treating or examining physician.¹⁹⁷ Such testimony may be essential to establish the diagnosis of the plaintiff's disease, his medical history, and the presence or absence of other possible risk factors for the disease.¹⁹⁸ The problems with medical opinion evidence arise when epidemiologic evidence of general causation is weak or absent¹⁹⁹ and the evidence of the capability of a substance to cause harm consists only of animal studies, mutagenicity testing, *in vitro* studies, or chemical structure-activity relationships.²⁰⁰ Those kinds of evidence are a weak basis for concluding that the toxic substance causes any human disease at all.²⁰¹ They are an extremely uncertain basis for making quantitative predictions that would allow a comparison of exposure risks and background risks. Indeed, the cases that have approved such evidence as sufficient to support a plaintiff's verdict have tended to ignore the absence of evidence²⁰² that would permit the plaintiff to distinguish background risks.

Plaintiffs who lack an epidemiologic basis for their proof of causation nonetheless frequently offer medical testimony from a treating physician or other expert who opines that the plaintiff's disease was caused by toxic substance exposure.²⁰³ This form of opinion is evident in

197. See *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984); *Wells v. Ortho Pharmaceutical Corp.*, 615 F. Supp. 262, 272-73 (N.D. Ga. 1985), *aff'd in part, modified in part*, 788 F.2d 741 (11th Cir.), *cert. denied*, 479 U.S. 950 (1986); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1235 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

198. See *infra* notes 206-12 and accompanying text.

199. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 862 (3d Cir. 1990) (opinion evidence based on, *inter alia*, animal tests and industrial exposures should not have been excluded), *cert. denied*, 111 S. Ct. 1584 (1991); *Ferebee*, 736 F.2d at 1535 (treating physicians testified).

200. See *supra* note 179 and accompanying text.

201. *Id.*

202. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d at 862 (approving proffered medical causation evidence). Such evidence is not a valid means of distinguishing other causal explanations when the causes of a majority of background cases are unknown. See *Rubanick v. Witco Chem. Co.*, 593 A.2d 733, 735-36 (N.J. 1991) (animal studies of PCBs, personal history, and higher than expected incidence of cancer at place of work admissible on causation of colon cancer); *infra* notes 264-76, 293-300 and accompanying text.

203. See, e.g., *Eggar v. Burlington N.R.R.*, No. CV89-159-BLG-JFB, 1991 U.S. Dist. LEXIS 19240 (D. Mont. Dec. 18, 1991); *In re Joint E. and S. Dist. Asbestos Litig. (Maiorana)*, 758 F. Supp. 199 (S.D.N.Y. 1991), *rev'd*, 964 F.2d 92 (2d Cir. 1992); *Wells v. Ortho Pharmaceutical Corp.*, 615 F. Supp. 262 (N.D. Ga. 1985), *aff'd in part, modified in part*, 788 F.2d 741 (11th Cir.), *cert. denied*, 479 U.S. 950 (1986); *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992).

The problems associated with medical testimony or disease causation have been discussed at length in Black, *supra* note 15, at 659-81. Depending on the circumstances, courts or the parties may espouse the view that a physician's testimony is preferred on the

*Renaud v. Martin Marietta Corp.*²⁰⁴ and other recent cases. When there is no clinical test that establishes a cause or distinguishes among possible causes, however, such "intuition" can only be characterized as speculation. This kind of speculation could easily be unmasked by inquiring into the reasoning behind the witness's opinion.²⁰⁵

d. Differential Diagnosis

The problem of distinguishing other causes takes a somewhat different form when there are other known risk factors for the plaintiff's disease. The effort to distinguish and eliminate other known risk factors is sometimes called "differential diagnosis."²⁰⁶ Diseases such as cancer that can result from toxic chemical exposure may also be associated with other identified risk factors, such as smoking, diet, lifestyle, as well as undifferentiated background risk associated with radiation and biological processes. Thus, the obvious question, particularly when the plaintiff does not exhibit other known risk factors such as diet, smoking, or a family history of the same cancer, is whether the absence of other risk factors increases the likelihood that the plaintiff's disease was caused by exposure to the toxic substance.²⁰⁷

issue of causation. The District of Columbia Circuit affirmed a jury's finding of liability in *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529, 1535-36 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984), based on the testimony of treating physicians. Plaintiffs often offer such testimony.

Interestingly, defendants sometimes object when a treating physician does not testify on causation. *See, e.g., Landrigan*, 605 A.2d at 1083 (trial court ruled that an epidemiologist could not testify on individual causation, nor could a nontreating physician offer an opinion based on epidemiologic evidence). Medical evidence, including testimony of a treating physician, may provide evidence of diagnosis of plaintiff's disease or injury or of the existence of other risk factors for the disease. *See id.*

204. 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

205. *See Ferebee v. Chevron Chem. Corp.*, 736 F.2d 1529 (D.C. Cir.), *cert. denied*, 469 U.S. 1062 (1984); *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733 (N.J. 1991). The *Ferebee* court based its affirmation of the plaintiff's verdict on the treating physicians' use of "tissue samples, standard tests, and patient examination." 736 F.2d at 1536. The court did not appear to consider whether those tests were in any way indicative of the cause of *Ferebee's* condition.

206. This usage of the term seems to be a misnomer. STEDMAN'S MEDICAL DICTIONARY (25th ed. 1990), defines differential diagnosis as "the determination of which of two or more diseases with similar symptoms is the one from which the patient is suffering, by a systematic comparison and contrasting of the clinical findings." *Id.* at 428. In toxic torts, the term is applied to the determination of which of two or more factors caused the plaintiff's disease, the diagnosis of which is not in question.

207. "Differential diagnosis" is an argument that cuts both ways. Plaintiffs are likely to make a differential diagnosis argument that the toxic exposure is the likely cause of the plaintiff's disease because other known risk factors are absent. *See, e.g., Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992) (discussing risk factors for colon cancer). Defendants are likely to point out that plaintiff has failed to eliminate other known risk factors that are applicable to the plaintiff. *See, e.g., In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1253 (E.D.N.Y. 1985) (plaintiff's experts "fail to show how the myriad illnesses at issue are more likely to have been caused by Agent Orange than by something

Plaintiffs often argue and courts sometimes accept the notion that the absence of other risk factors increases the likelihood that the plaintiff's disease was caused by the toxic exposure at issue.²⁰⁸ The validity of that kind of reasoning, however, rests on two unstated, and usually untested, assumptions. First, such reasoning treats toxic exposure and the other risks as alternatives. In other words, it assumes that the disease was caused by the toxic exposure *or* some other cause, such as the other identified risk factors.²⁰⁹ Second, it assumes that most causes of the disease in question are known; otherwise, the elimination of other risk factors would not significantly increase the likelihood that the toxic exposure was the cause of the plaintiff's disease.

The assumption that risk factors, including the toxic exposure, represent alternative causes is true only if the various risks are additive. Additivity is only one of several ways in which risk factors for the same disease may relate. The combined effects may be the same, greater, or less than the sum of the effects as measured separately.²¹⁰ Additive effects represent the absence of interaction between risk factors.²¹¹ Thus, each factor adds an incremental level of risk to the background risk that is independent of the presence or absence of other risk factors. Additive

else"), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988). In actuality, the issue of other causes is present in every case that involves a disease with a significant background risk. Almost certainly, there are as yet unidentified risk factors that affect background disease incidence. See ROTHMAN, *supra* note 44, at 12. The assumption of uniform risk factors in the background population reflects ignorance of what those factors are. *Id.* Only with signature diseases such as mesothelioma and clear cell adenocarcinoma that are rare in the absence of exposure to an identified carcinogen can this issue be avoided. See *supra* note 190 and accompanying text.

208. See generally Landrigan v. Celotex Corp., 605 A.2d 1079, 1087-88 (N.J. 1992); Rubanick v. Witco Chem. Corp., 542 A.2d 975 (N.J. Super. Ct. Law Div. 1988), *rev'd*, 576 A.2d 4 (N.J. Super. Ct. App. Div. 1990), *modified*, 593 A.2d 733 (N.J. 1991).

209. Attribution of causation in an illness with several possible causes is often a complex task. As the reader will recall from the discussion at the beginning of this section, the relative risk data from epidemiologic studies can be used to determine the fractions of cases of a disease in an exposed population that are attributable to the exposure and to the background causes. See *supra* notes 187-90 and accompanying text.

210. See ROTHMAN, *supra* note 44, at 311-26. Combined effects that are less than additive are considered antagonistic, while combined risks that are greater than the sum of the separate effects are considered synergistic. *Id.* at 318-20.

211. *Id.* at 313-15. Rothman states that although some epidemiologists treat the foregoing definitional scheme as arbitrary, the use of an additivity assumption for independent action has practical consequences in interpreting and utilizing epidemiologic data. *Id.* at 316-17. For example, the combined relative risks of oral contraceptives and hypertension for thrombotic stroke are greater than the sum of the relative risks of each. *Id.* at 316. When these greater than additive risks are considered to be synergistic, the practical conclusion is that a woman should consider her blood pressure history in deciding whether to use oral contraceptives, a result that seems intuitively correct. *Id.* at 316-17.

risks are properly treated as alternative risks in a causation analysis.²¹² Often, however, the information necessary to support that assumption is not available.

Risk factors whose combined effects are greater than additive are considered interactive or synergistic.²¹³ In this situation, each risk factor enhances the risk contributed by the other factor so that the total incidence of the disease is greater than the sum of the incidence attributable to each factor separately, sometimes approaching a multiplicative effect. Perhaps surprisingly, the presence or absence of other risk factors that are multiplicative does not increase or decrease the fraction of disease attributable to the toxic exposure. Thus, when the causes are synergistic, as with smoking and asbestos and lung cancer,²¹⁴ it is incorrect to pose the question as one of whether the disease was caused by one factor or another.²¹⁵

212. For example, smoking increases the incidence of lung cancer by a factor of about 10 and occupational exposure to asbestos increases risk by about a factor of 5. U.S. SURGEON GEN., *supra* note 47, at 216-17. If the background incidence of lung cancer among nonsmokers who do not have occupational exposures to asbestos is normalized to 1.0, nonsmoking asbestos workers would have a lung cancer incidence of 5.0, while smokers would have a lung cancer rate incidence of 10.0. For smoking asbestos workers, if the risks were noninteractive (and therefore additive), the lung cancer incidence would be about 14, which is the sum of the background case, the four cases added by asbestos exposure, and the nine cases expected to be added as a result of smoking. In this situation, smoking and asbestos would represent alternative causes, since any one plaintiff's case would probably be caused by one factor or the other, not by both acting together. Thus, while a nonsmoking asbestos worker could argue that the probability that asbestos caused his cancer was 80% (4/5), the smoking asbestos worker could point to only a 28% probability (4 cases out of 14 total) that his lung cancer was caused by asbestos (the other 10 cases being the result of background causes or cigarette smoking).

213. ROTHMAN, *supra* note 44 at 315-16.

214. *Id.* at 312.

215. Because the effects of smoking and asbestos are multiplicative for lung cancer, the population of smoking asbestos workers described in *supra* note 212 is expected to have lung cancer incidence of 5 times 10, or 50, rather than the 15 cases predicted by adding the separate risks. See U.S. SURGEON GEN., U.S. DEP'T OF HEALTH & HUMAN SERVS., *supra* note 47, at 216-17. The fraction attributable to asbestos in this case is 40/50 (or 0.8, subtracting the 10 cases that would have occurred as the result of smoking alone or without exposure to either factor from the 50 total cases), the same attributable fraction obtained when the effects of asbestos alone are considered. Counterintuitively, the fraction of lung cancer cases among smoking asbestos workers for which smoking can be considered causative is 45/50 (or 0.9, obtained by subtracting the 5 cases that would have occurred due to asbestos alone or in the absence of either exposure, from the total cases). Those results do not mean, however, that arguments cannot be made that the plaintiff's smoking constituted contributory negligence or an intervening cause that should reduce the asbestos manufacturer's liability or eliminate it altogether.

Defendants sometimes make such arguments. See, e.g., *In re Brooklyn Navy Yard Asbestos Litig.* (Joint E. & S. Dist. Asbestos Litig.), 971 F.2d 831 (2d Cir. 1992); *In re Manguno* (Acosta v. Babcock & Wilcox), 961 F.2d 533 (5th Cir. 1992); *Borman v. Raymark Indus.*, 960 F.2d 327 (3d Cir. 1992) (upholding plaintiff's verdict and affirming trial court's refusal to charge the jury on apportionment of damages between smoking and asbestos). In *Manguno*, the jury returned a verdict for defendant asbestos manufacturers. *Id.* at 534.

There are several scenarios under which this issue could arise, but the actual cases tend to be grouped into two extremes. Where epidemiologic data are available that address the contributions of both the toxic substance and other causal factors, the plaintiff's attributable risk and probability of causation by the toxic substance can be determined by calculating attributable fractions, whether the risks are additive, multiplicative or antagonistic. From those calculations, it can be determined whether the plaintiff can satisfy the more likely than not standard of proof on causation.²¹⁶

The other extreme is represented by toxic tort cases where multiple risk factors are treated in a vague or qualitative fashion and data are not available to support a quantitative analysis. The plaintiff's expert may opine that because other known risk factors are absent in plaintiff's case, it is the expert's opinion that plaintiff's disease was caused by the toxic substance exposure.²¹⁷ This argument has superficial appeal. It can only be valid, however, when risk factors that account for most cases of the plaintiff's injury and their interactions are understood. Although some risk factors for cancers and birth defects have been identified, the causes of background incidence of most birth defects and cancers remain unknown. Even if the identified risk factors are alternative, independent causes, as this line of analysis assumes, the expert's opinion distinguishes among factors that make up only a small part of the total picture, while ignoring the probability that the plaintiff's injury may stem from the same unidentified factors that are responsible for most cases of the injury. Whether the differential diagnosis argument is made on behalf of plaintiff or defendant, it adds little to the resolution of the case when it is based on vague, qualitative assumptions about alternative causes.²¹⁸

The defendants argued, and the jury apparently accepted, that plaintiff's smoking cast sufficient doubt on the proposition that asbestos caused the plaintiff's lung cancer. The Fifth Circuit reversed on the basis of improper jury instructions and remanded for a new trial. *Id.* at 535-36.

216. Often there will not be sufficient data to determine whether the risk factors are additive, synergistic, or antagonistic. In such cases the point is arguable, but the better approach would seem to be to ignore other risk factors whose interactions with the toxic substance exposure are unknown and assume that the relative risk associated with the toxic substance exposure applies whether the other risk factors are present or not.

217. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 735-36 (N.J. 1991).

218. The defendant's argument that the existence of known risk factors dilutes the likelihood of causation by the chemical exposure has a point, however, when the plaintiff's general causation evidence is based on medical opinion, animal studies, cellular assays and other such evidence. In *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 839-40 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991), the plaintiffs claimed that a variety of commonplace ailments were due to PCB exposure. Available epidemiologic data failed to demonstrate any connection between PCBs and the claimed physical injuries. The defendants' objection that plaintiffs failed to rule out the known causes of such commonplace ailments as high cholesterol and high blood pressure should have been well taken by the court. See *id.* at 861-62.

3. PROOF OF EXPOSURE

a. Inferences from Similar Circumstances.

While the foregoing discussion has focused on proving the harmful effects of exposure, the issues of whether an exposure occurred and if so, of what magnitude, are often present in toxic tort cases.²¹⁹ In several recent cases, the duration and magnitude of the plaintiffs' exposure was subject to a great deal of uncertainty.²²⁰ How can the plaintiff prove exposure? In instances where the plaintiff has a signature disease, the disease itself constitutes strong evidence of exposure because it rarely occurs in the absence of exposure.²²¹ More commonly, however, the plaintiff's injury is not so distinctive that it can be reliably attributed to a toxic substance exposure, even where the substance is known to increase the risk of the condition. Where the claimed injury is one that is not attributable almost solely to a toxic substance, exposure data is meaningful only if it is quantitative. The plaintiff must prove the magnitude of her exposure to a degree of certainty that supports the inference that the evidence linking a toxic chemical to disease is applicable to her.

Some cases, particularly those involving workplace exposures, involve situations that are known to involve exposure to a toxic substance.²²² The severity of the plaintiff's exposure can be inferred from the length of time he was present in the environment. The situation would be similar where there is an ongoing exposure, such as a drinking

219. See, e.g., *Paoli R.R. Yard*, 916 F.2d 829 (evidentiary issue of the degree of plaintiffs' exposure to PCBs), cert. denied, 111 S. Ct. 1584 (1991); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (case dismissed because of plaintiffs' inability to make out a prima facie case of exposure), aff'd, 972 F.2d 304 (10th Cir. 1992).

220. See *Christopherson v. Allied-Signal Corp.*, 939 F.2d 1106 (5th Cir. 1991) (en banc), cert. denied, 112 S. Ct. 1280 (1992). In *Christopherson*, plaintiff's husband worked in a plant that produced nickel/cadmium batteries. *Id.* at 1108. He did not work in the production area, however, but visited the area intermittently. *Id.* There also appears to have been no direct evidence on the nature of fumes to which Christopherson was exposed during those visits. *Id.* at 1113. Apparently a fellow employee's affidavit alleged that Christopherson was exposed to airborne particles of nickel and cadmium, but it is not clear that the employee could have known the chemical composition of the fumes. See *Christopherson v. Allied-Signal Corp.*, 902 F.2d 362 (5th Cir. 1990), rev'd, 939 F.2d 1106 (5th Cir. 1991) (en banc), cert. denied, 112 S. Ct. 1280 (1992).

221. In signature disease cases, the existence of the disease may suffice to prove that sufficient exposure occurred and that the exposure caused the plaintiff's injury. Cf. *Renaud*, 749 F. Supp. at 1553 (plaintiffs could have attempted to prove exposure through epidemiologic study).

222. Occupational exposure of insulation installers to asbestos is an example of that situation. See, e.g., *In re Joint E. & S. Dist. Asbestos Litig.* (Johns-Manville Corp.), 129 B.R. 710, 868 (E.D. & S.D.N.Y. 1991) (Weinstein, J.) (discussing proof of asbestos exposure in occupational and other settings), vacated, 982 F.2d 721 (2d Cir. 1992).

water exposure, that can be measured. In such cases, inferences about past conditions can be drawn from the present ones.²²³

Sometimes there is clinical evidence of the toxic substance that will suffice to prove the plaintiff's exposure.²²⁴ Substances such as asbestos²²⁵ and PCBs²²⁶ remain in tissues indefinitely, and can be detected by appropriate analytical tests. Other substances may result in subclinical changes that can be detected through appropriate testing.²²⁷

b. Inferences from Modeling

The reasoning underlying the proofs of exposure outlined above is readily apparent and is of the kind already familiar to courts in other contexts. Cases involving possible past exposures where contemporaneous measurements cannot be taken or inferred, and for which there are no available analytical tests, present a much more difficult case. In such cases, experts may use various models to estimate exposure. Models are mathematical formulas that are designed to provide estimates of facts that cannot be measured directly.²²⁸ They range from simple formulas such as the model discussed below that purports to describe the ratio of blood or serum PCB levels to adipose tissue levels, to complex computer programs used to model groundwater contaminant migration and air pollutant dispersion.

Models, at least conceptually, sometimes begin with theories or hypotheses about how different kinds of data might be related. A scientist would be unlikely to propose a model that was not at least plausible, based on her understanding of how the phenomena in question

223. See, e.g., *Sterling v. Velsicol Chem. Corp.*, 855 F.2d 1188 (6th Cir. 1988) (chlorinated hydrocarbons detected in 12 to 15 drinking water wells).

224. Clinical evidence will not necessarily be sufficient to prove that the exposure caused injury, however, because exposure does not always result in disease. See *supra* note 47 and accompanying text.

225. See, e.g., *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992) (clinical data, such as asbestos in or near a tumor, may support a finding of specific causation).

226. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 841 & n.10 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). Dioxin is also retained in fat. Brennan, *supra* note 25, at 51 (citing Peter C. Kahn et al., *Dioxins and Dibenzofurans in Blood and Adipose Tissue of Agent Orange-Exposed Vietnam Veterans and Matched Controls*, 259 JAMA 1661 (1988)).

227. See Robert R. Lauwerys, *Occupational Toxicology*, in CASARETT & DOULL, *supra* note 148, 947, 954-66; see also OS&TP, *Chemical Carcinogens*, *supra* note 141, at 10,386, 10,441 (discussing DNA alteration and relation to carcinogenesis); Brennan, *supra* note 15, at 502 n.174. Clinical testing for exposure-induced damage may become more important as knowledge accumulates regarding molecular alterations from chemical exposures. See OS&TP, *Chemical Carcinogens*, *supra* note 141, at 10,441.

228. For a discussion of groundwater contaminant migration modeling, see Allen Kezsombi & Alan V. Goldman, *The Boundaries of Groundwater Modeling Under the Law: Standards for Excluding Speculative Expert Testimony*, 27 TORT & INS. L.J. 109 (1991). See also Itzchak E. Kornfeld, *Comment to the Boundaries of Groundwater Modeling Under the Law: Standards for Excluding Speculative Expert Testimony*, 28 TORT & INS. L.J. 59 (1993).

are connected. Plausibility alone, however, like the mechanistic theories of cancer causation, is often insufficient to eliminate other plausible but untested models or theories. Thus, a model must be validated before an expert or a court should assume that it has any probative value.

Validation of a model involves testing the model's predictive capability in circumstances where the expected results can be independently measured.²²⁹ In the case of the blood/adipose tissue partition model or the much more complex groundwater contaminant migration modeling, actual concentrations can be measured and compared with values predicted by the model. From such data, it can be determined whether the model has any predictive value and how accurate or reliable those predictions are. If the model proves valid and reasonably accurate, then it is reasonable to apply it to other situations similar to the one for which the model has been tested.

A number of cases, however, have involved modeling of exposures where the models have not been subjected to the most cursory validation, sometimes in the face of data that contradict the model. In *In re Paoli Railroad Yard PCB Litigation*,²³⁰ one of the ways the plaintiffs attempted to prove exposure to PCBs was by showing that their PCB levels were elevated above background levels. Because PCBs accumulate in fatty tissue and are not quickly eliminated from the body, it is possible to measure PCB levels in blood or tissue samples from individuals and compare them to norms for the general population.²³¹ The Agency for Toxic Substances and Disease Registry (ATSDR) had done a study on blood PCB levels in Paoli residents and concluded that the residents' serum PCB levels did not differ significantly from those of the general population.²³² Plaintiffs contended that the ATSDR study's conclusions regarding background PCB levels in the general population were erroneous. Their experts sought to show that the plaintiffs' adipose or fat PCB levels exceeded norms found in the Environmental Protection Agency's National Human Adipose Tissue Study.²³³ Because most of the plaintiffs' PCB levels were determined by blood tests alone, plaintiffs'

229. See Kezsbom & Goldman, *supra* note 228, at 117 (noting that the plaintiffs in *Sterling v. Velsicol* never verified their model against real-world data). Kornfeld, who has a contrary view of *Sterling v. Velsicol*, states that models must be scrutinized to determine whether they replicate real-world data and have been calibrated. See Kornfeld, *supra* note 228, at 68.

230. 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991).

231. See *id.* at 839-41. PCBs are polychlorinated biphenyls, once commonly used in electrical transformers.

232. Nor did they differ significantly among the residents according to higher soil concentrations in the residents' yards, number of years in the vicinity, or residence in more or less highly contaminated areas. *In re Paoli R.R. Yard PCB Litig.*, 706 F. Supp. 358, 364-65 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991).

233. *Id.* at 371. The study was conducted between 1970 and 1983.

expert used his own formula to calculate their adipose tissue PCB levels, which he then compared with results reported in the national study.²³⁴

As the trial court recognized, however, neither of the plaintiffs' experts on this issue²³⁵ cited any basis for the claimed relationship between PCB concentrations in blood and adipose tissue.²³⁶ Moreover, where blood and adipose tissue PCB levels were measured in the same plaintiffs, the results did not bear out the ratios asserted by plaintiffs' experts.²³⁷ Thus, although a relationship between blood and adipose tissue levels of PCBs is plausible, the direct ratio posited by plaintiffs' witnesses was not validated and was demonstrably inaccurate as indicated by comparison of predicted levels with actual measurements in plaintiffs who had both tests. Conclusions based on models that have not been validated by actual measurement,²³⁸ or worse, which are contradicted by actual measurements, are based on invalid reasoning and should be rejected by the courts.

Use of unvalidated models is not the only concern about models. Models inherently involve approximations—generalizations about physical phenomena and estimations that must be made because the actual situation cannot be studied directly.²³⁹ Those approximations inevitably introduce inaccuracies into the model's predictions. At some point, the uncertainty or inaccuracy may become so large that the model's results are too unreliable to prove the fact on which they are offered.

Modeling of groundwater and surface water contamination was at issue in *Renaud v. Martin Marietta Corp.*,²⁴⁰ in which the plaintiffs claimed that contamination of the public water supply had caused childhood cancers and other diseases.²⁴¹ Plaintiffs' case on exposure involved the issue of whether contaminants released at the Martin Marietta facility had reached their taps through the Denver culinary water distribution system. Because the circumstances that created the discharges at issue had changed in the years preceding the suit, the plaintiffs relied on

234. *Id.* at 370-71. The Third Circuit opinion implies that the NHATS study used the formula that relates blood and adipose tissue levels. *See Paoli*, 916 F.2d at 841 n.10.

235. Dr. Ian C.T. Nesbit, a Ph.D. physicist and consultant, and Dr. Robert K. Simon, an industrial hygienist, toxicologist, and forensic analytical chemist, testified on the same issue. *See Paoli*, 916 F.2d at 840, 847.

236. *Paoli*, 706 F. Supp. at 372..

237. *Id.*

238. In *Paoli*, one of plaintiffs' experts, Dr. Herbert Allen, purported to calculate the levels of airborne PCBs to which plaintiffs had been exposed based on soil PCB concentrations, using a formula, i.e. a model, he had devised. 916 F.2d at 839. Dr. Allen's predictions, however, were higher than the measurements actually taken. *See infra* text accompanying notes 257-59.

239. *See Kezsomb & Goldman, supra* note 228, at 109, 116-19.

240. 749 F. Supp. 1545 (D. Colo. 1990), *aff'd*, 972 F.2d 304 (10th Cir. 1992).

241. *Id.* at 1547.

hydrological modeling of ground and surface water movement as proof that contaminants had reached the water distribution system.²⁴²

Defendants argued that there were serious flaws in the modeling, most notably that plaintiffs had erred by failing to consider all relevant factors when deriving the decay coefficient for the contaminants.²⁴³ Further, the experts had not taken into account the possibility that chlorination at the water intake plant had destroyed or greatly reduced the concentration of contaminants.²⁴⁴ Although the court stated that “[t]he issues of which factors should have been considered and what impact each should have been given” were questions for the jury, it seems clear that the factors that plaintiffs’ experts ignored would have had a large impact on the concentrations predicted by the fate and transport modeling.²⁴⁵ Models are always subject to dispute over the factors that are included or excluded, and thus cannot be judged by too rigorous a standard. Nonetheless, where experts have excluded significant factors that would tend to produce results at odds with their conclusions, the court should exclude the modeling results unless there is some more direct way to demonstrate the model’s validity and accuracy.²⁴⁶

Another concern with modeling is that uncertain input data affect the reliability of the results of all kinds of exposure modeling. If the input data are very limited, there will be uncertainty about how representative those data are, and those uncertainties will produce corresponding uncertainties about the modeling results, no matter how good the model is. The greatest uncertainty about the exposure modeling in *Renaud* resulted from such a scarcity of data.²⁴⁷ In that case, the plaintiffs’ exposure estimate, obtained through the transport modeling described above, was based on a single loading concentration for the contaminants. The court recognized that the single data point on which the modeling was based could not be said to be representative of the 11-year period

242. *Id.* at 1549.

243. *Id.* at 1552. A decay coefficient was necessary because the chemicals at issue were known to undergo degradation in the environment. *See id.* at 1549.

244. *See Letter from Dr. Hannah Pavlik, Ebasco Environmental, to Judge Zita Weinshienk, U.S. District Court (Aug. 29, 1990).* Dr. Pavlik, a geochemist and hydrogeologist, was retained as a court-appointed expert witness. *Renaud*, 749 F. Supp. at 1553. Her report also indicated a number of other methodological and factual flaws in the hydrological modeling. *See Letter from Dr. Hannah Pavlik, supra*, at 9.

245. *Renaud*, 749 F. Supp. at 1552; *see Letter from Dr. Hannah Pavlik, supra* note 244, at 16.

246. The *Renaud* plaintiffs’ experts could have conducted experiments on the effects of chlorination on the chemicals in question. The only available information on that issue appeared to be the defendants’ own tests, however, which were consistent with their position. *See Letter from Dr. Hannah Pavlik, supra* note 244, at 9.

247. Data that is based on unsound assumptions, unverifiable assumptions, or erroneous input results in a manifestation of the “garbage in, garbage out” phenomenon. *See Kezsomb & Goldman, supra* note 228, at 116.

over which releases occurred.²⁴⁸ It found the single data point to be a fatal flaw in the plaintiffs' exposure case.²⁴⁹

V. DIVERGENCE OF OPINION

A. Deferential Review and the Accumulation of Errors

The foregoing Part has examined separately the problems associated with proof of exposure, capability of the substance to cause harm, and distinguishing among causes and concluded that the evidence deemed acceptable in many toxic tort cases is often grossly inadequate to prove the propositions on which it is offered. It is also important to examine how those issues are brought together in real cases, keeping in mind that the ultimate causation issue is whether exposure to a toxic substance caused the plaintiff's disease. This part discusses several recent cases that are particularly troubling when viewed as a whole, because the plaintiff's cases can, at best, be characterized as consisting of possibilities and speculation strung together in ways that fall far short of the legal requirements of proof.

The Third Circuit's decision in *In re Paoli Railroad Yard PCB Litigation*²⁵⁰ represents one of the most troubling decisions on the admissibility and sufficiency of challenged scientific evidence. *Paoli* involved an action by 38 neighbors and employees of an electric railcar maintenance facility contaminated by PCBs.²⁵¹ The action, which was brought against owners and operators of the site, and suppliers of PCBs and transformers, made claims for various injuries and for medical monitoring costs necessary to protect against latent disease.²⁵²

Defendants' motions to exclude plaintiffs' evidence of exposure to PCBs and other causation evidence, and for summary judgment were granted by the trial court.²⁵³ The Third Circuit, however, reversed, finding that the trial court had improperly excluded sufficient evidence to survive summary judgment.²⁵⁴ The case represents a virtually complete catalog of the unprobative and insufficient kinds of proof identified in this article.

Two factual issues on which scientific evidence was crucial were: (1) whether plaintiffs received any exposure above background levels of

248. *Renaud*, 749 F. Supp. at 1552-53.

249. *Id.*

250. 916 F.2d 829 (3d Cir. 1990), *rev'd* 706 F. Supp. 358 (E.D. Pa. 1988), *cert. denied* 111 S. Ct. 1584 (1991).

251. *Id.* at 832.

252. *Id.* at 849.

253. *Id.* at 835.

254. *Id.* at 862.

PCBs that could be attributed to the Paoli railyard; and (2) whether PCBs are capable of causing the ailments of which plaintiffs complained or of which they believed they were at risk.²⁵⁵ The primary problem with plaintiffs' evidence on exposure involves unvalidated methodology through which plaintiffs attempted to show that their exposure to PCBs exceeded background levels.

Plaintiffs offered several forms of evidence in addition to the blood/adipose tissue calculations discussed in the preceding Part,²⁵⁶ in support of their contentions of higher than background levels of PCB exposure. First, they offered the testimony of Dr. Herbert Allen, an environmental chemist who used a formula of his own devising to calculate airborne exposure levels from PCB levels measured in neighborhood soils.²⁵⁷ Nothing in either the district court's or the appellate court's opinions, however, suggests that Dr. Allen's formula had been tested, that is, that its validity had been demonstrated by comparing the airborne concentrations predicted by the formula and actual, measured air concentrations.²⁵⁸ In fact, the district court opinion states that the "actual measurements that were taken showed an amount much lower than [Allen] calculated."²⁵⁹

Plaintiffs also offered the testimony of Dr. Deborah Barsotti, a toxicologist employed by the Agency for Toxic Substances and Disease Registry, who claimed to have correlated gas chromatography tracings of PCBs in the plaintiff's blood with tracings from soil samples from the Paoli railyard.²⁶⁰ Dr. Barsotti, however, was apparently unable to support her general statements with reference to any specific plaintiffs' blood samples or soil samples.²⁶¹ Later, she apparently conceded that the equipment she had used was not capable of yielding the results she claimed.²⁶²

Plaintiffs' evidence on general causation and on distinguishing background causes was hardly more probative. On the question of whether PCBs are capable of causing the kinds of illnesses complained of, the plaintiffs were faced with various studies that had failed to find a correlation between PCB exposure and significant human disease. One such study was the ATSDR's study, *Toxicological Profile for Selected*

255. See *id.* at 860-62.

256. See *supra* notes 230-38 and accompanying text.

257. *Paoli*, 916 F.2d at 838-39.

258. *Id.* at 839, 842; *In re Paoli R.R. Yard PCB Litig.*, 706 F. Supp. 358, 370 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), *cert. denied*, 111 S. Ct. 1584 (1991).

259. *Paoli*, 706 F. Supp. at 370.

260. *Paoli*, 916 F.2d at 839.

261. *Id.* at 842.

262. *Id.*

PCBs.²⁶³ As summarized in the Foreword, the ATSDR study found that only skin lesions and liver effects that were not associated with "clinically detectable disease" had been observed in PCB-exposed workers. The study also concluded that adverse effects had not been observed in persons with non-occupational exposures.²⁶⁴

The Third Circuit, however, found various plaintiffs' witnesses' testimony sufficient to create a jury question.²⁶⁵ Some of this testimony can only be described as conclusory.²⁶⁶ Several of plaintiffs' experts relied on animal studies and on studies involving incidents of accidental ingestion of a mixture of PCBs and PCDFs,²⁶⁷ the "Yusho" and "Yu Cheng" incidents that occurred respectively in Japan in 1968, and Taiwan in 1979.²⁶⁸ Lastly, the plaintiffs offered the testimony of Dr. William J. Nicholson, a professor of community medicine who had performed a "meta-analysis" of existing (negative) epidemiologic studies²⁶⁹ and concluded that his analysis showed that PCBs can cause liver, biliary tract and gall bladder disorders.²⁷⁰

The evidence on general and individual causation was clearly insufficient to permit the inference that would satisfy the more likely than not standard of proof. Animal studies, as discussed earlier, are at best subject to large uncertainties when extrapolated to humans, particularly for effects of chronic, low-level exposures.²⁷¹

The use of the Yusho and Yu Cheng studies were not challenged as invalid,²⁷² but their use for the purposes of proving that PCBs cause significant adverse effects in humans involves invalid reasoning.²⁷³ The

263. *Paoli*, 706 F. Supp at 365 (citing Toxicological Profile for Selected PCBs (draft Nov. 1987)).

264. *Id.*

265. *Paoli*, 916 F.2d at 862. Dr. Barsotti testified that exposure to PCBs at Paoli was a substantial factor in causing plaintiffs' elevated triglycerides, cholesterol, and liver enzyme levels. *Id.* at 839. According to the court, she based her conclusions on her inspection of the Paoli railyard and her review of various reports and studies, and of soil samples from the railyard. *Id.*

266. Dr. Barsotti's affidavits on causation for each plaintiff were identical for the first fourteen pages, with only two additional paragraphs listing the alleged injuries and concluding they were caused by PCBs. *Id.* at 842-43. Thus, she never explained her reasoning beyond the fact that plaintiffs were exposed to the railyard PCBs, studies have indicated possible injuries from PCBs, and plaintiffs have injuries.

267. PCDFs are polychlorinated dibenzofurans, chemically related to PCBs. *See id.* at 839.

268. *Id.* at 840.

269. *Id.* at 841. A meta-analysis combines the results of epidemiologic studies to increase the total sample size and reanalyzes the data. *Id.*

270. *Id.*

271. *See supra* notes 140-68 and accompanying text.

272. *See In re Paoli R.R. Yard PCB Litig.*, 706 F. Supp. 358, 366 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), *cert. denied*, 111 S. Ct. 1584 (1991).

273. *Id.* at 368.

studies identified harmful effects from two incidents involving the ingestion of oil mixtures containing PCBs and PCDFs. Assuming that the conclusions regarding the Yusho and Yu Cheng incidents were accurate, the results can at most be said to prove the following proposition: The harmful effects observed in the Yusho and Yu Cheng incidents were caused by either: (1) PCBs; (2) PCDFs; or (3) PCBs and PCDFs in combination.²⁷⁴ The studies do not allow the conclusion that PCBs alone can cause the effects observed in the study. Additionally, because PCDFs are regarded as more toxic than PCBs,²⁷⁵ the proposition that plaintiffs argued from the study is not supported by the study.²⁷⁶

The meta-analysis of epidemiologic studies presents a somewhat different problem. Meta-analyses are not outside the scope of recognized scientific methodology.²⁷⁷ The defendants did raise questions, however, about the way in which Dr. Nicholson analyzed the existing data, contending that he omitted data that were inconsistent with his conclusions.²⁷⁸ Thus, the trial court could have examined the bases on which Dr. Nicholson included and excluded data to determine whether

274. This analysis assumes that there is a basis for concluding that the effects were not caused by other unnamed constituents.

275. See *In re Paoli R.R. Yard Litig.*, 916 F.2d 829, 839, 843 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). PCDFs are polychlorinated dibenzofurans. *Id.* at 839.

276. The Third Circuit noted that Dr. Herbert Allen testified that activities at the railyard such as welding and cutting contaminated equipment could have converted PCBs into dioxins and PCDFs. *Id.* at 839. Although the court characterized this testimony as "particularly significant." *Id.* It appears to offer only the most tenuous evidence of an unquantifiable possibility, since there is no indication that plaintiffs tested for or offered physical evidence of dioxin or PCDF contamination or exposure.

277. See *id.* at 857 (citing C. David Naylor, *Two Cheers for Meta-analysis: Problems and Opportunities in Aggregating Results of Clinical Trials*, 138 CAN. MED. ASS'N J. 891, 894 (1988)).

278. *Id.* at 845. The Third Circuit regarded the issue of how the meta-analysis was conducted as one of credibility, and therefore for the jury, although it qualified that conclusion with the statement that the meta-analysis would be excludable if no reasonable person could believe the study. *Id.* at 858. The district court appears to have excluded the meta-analysis primarily on relevance grounds, although it first discussed the standards for admitting novel scientific evidence in connection with the meta-analysis. *In re Paoli R.R. Yard Litig.*, 706 F. Supp. 358, 372-73 (E.D. Pa. 1988), *rev'd*, 916 F.2d 829 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). It deemed the study irrelevant because the diseases the study attributed to PCB exposure were not those claimed by plaintiffs. *Id.* at 373. The Third Circuit rejected that rationale, however, because it also held that plaintiffs were entitled to proceed under a medical monitoring claim; it deemed the meta-analysis relevant to a determination of the risks of future disease to which the plaintiffs were exposed.

The issue of when evidence based on a technique that may have been improperly applied may be excluded, as opposed to submitted to the jury, is a current controversy in evidence law. If, however, it was apparent that the meta-analysis was a result-oriented manipulation of data from other studies, the trial court would have been justified in excluding it, even under the Third Circuit's standard. See *infra* notes 357-66 and accompanying text (discussing "data dredging" in connection with meta-analysis and reanalysis of epidemiological studies).

there were logical criteria, systematically applied, in combining and evaluating the data from previous studies.

Paoli also involved questions related to expert testimony on individual causation.²⁷⁹ Several witnesses appear to have asserted that based on test results indicating the presence of PCBs in the railyard and surrounding area, the presence of PCBs in plaintiffs' blood, plaintiffs' medical records, and the literature on effects of PCBs, they could state "to a reasonable medical certainty"²⁸⁰ that plaintiffs' various ailments were caused by exposure to PCBs.²⁸¹

Reference to the evidence on which these conclusions were purportedly based reveals that those conclusions amounted to nothing more than speculation. Animal studies and studies of incidents involving several chemicals provide only uncertain evidence that the substance will cause human disease at all. Animal studies simply do not produce results that permit reliable conclusions about the likelihood of human disease at particular exposure levels. The Yusho and Yu Cheng studies involved ingestion of much larger quantities of PCBs than the *Paoli* plaintiffs were exposed to, and that exposure involved another, probably more toxic chemical. In fact, *Paoli*, like other recent cases,²⁸² involves a scenario in which plaintiffs complaining of a variety of common ailments²⁸³ attempt to attribute those ailments to exposure to a toxic chemical for which human effects have not been demonstrated or are different from those

279. *Paoli*, 916 F.2d at 862. In some cases, the witnesses seemed to combine the question of whether PCBs can cause disease with the question of whether the assumed exposure caused a particular plaintiff's disease. See, e.g., *id.* at 839 (testimony of Dr. Deborah Barsotti); *id.* at 840 (testimony of Dr. Harry Shubin).

280. *Id.* at 851.

281. In some cases, the diagnoses themselves (not only their causes) were at issue. For example, Dr. Arthur Zahalsky opined that the plaintiffs suffered from immune system injuries. *Id.* at 840. At his deposition, however, Zahalsky admitted that testing (of his own design) required to validate his opinion had not been carried out. *Id.* at 843. The speculative nature of Zahalsky's contentions regarding whether PCBs can cause immune system damage is illustrated by his admission that he had not tested any of the plaintiffs. *Id.* at 843. The court also quotes Zahalsky as stating that "if his tests should support such a conclusion, 'then I will have done something with the clinical immunologists that has not yet been done.' " *Id.* (quoting Zahalsky).

Diagnoses relating to increased fear of illness and emotional distress seem particularly lacking in support. Dr. Deborah Barsotti apparently offered her opinion of this issue without having met or examined any of the plaintiffs, having spoken to only one plaintiff on the telephone. *Id.*

282. See, e.g., *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990) (childhood cancer, including leukemia; kidney cancer; seizure disorders), *aff'd*, 972 F.2d 304 (10th Cir. 1992); *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223 (E.D.N.Y. 1985) (plaintiffs alleged infertility, miscarriage, birth defects, cancer, emotional disturbance, and other commonplace ailments), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

283. *Paoli*, 916 F.2d at 839 (elevated blood pressure, triglycerides, and cholesterol, elevated liver enzymes, and emotional distress).

complained of by the plaintiffs. Even if one assumes that some level of exposure above background has occurred, a proposition that appears doubtful in *Paoli* and other cases,²⁸⁴ the question remains as to whether the individual plaintiffs' conditions were caused by the exposure or by the more commonplace causes of such diseases in the general population, whether known or unknown, a question that cannot be answered without evidence demonstrating that a substance can cause the plaintiff's disease and indicating increased disease incidence at the levels to which plaintiffs were exposed.²⁸⁵ Animal testing and studies involving high level exposure to several toxic substances simply cannot provide that information.

The logical extension of the plaintiffs' position is that virtually anyone with one or more of a whole host of commonplace ailments who may have come into contact with toxic substances should be able to recover from the entity responsible for the toxic substance. Such a proposition clearly goes too far; yet it is difficult to draw any principled distinctions about who should or should not recover if the reasoning of the *Paoli* plaintiffs and their experts is accepted.²⁸⁶

Concerns about the evidentiary basis of causation in toxic torts are not limited to cases involving large numbers of plaintiffs and an array of alleged injuries. Cases involving one or a few plaintiffs may raise similar concerns. Further, a court may tend to view such cases in isolation, even though extension of the case's underlying logic may lead to results similar to those in cases such as *Paoli*, namely, that there is no principled way to

284. See *supra* notes 260-89 and accompanying text.

285. On remand, the district court held a series of evidentiary hearings in accordance with the Third Circuit's directive. The rulings on the summary judgment motion and other evidentiary rulings can be found at *In re Paoli R.R. Yard PCB Litig.*, No. 86-2229, 1992 U.S. Dist. LEXIS 16287 & 18427-37 (E.D. Pa. Oct. 21, 1992) (numerous rulings); *In re Paoli R.R. Yard Litig.*, No. 86-2229, 1992 U.S. Dist. LEXIS 17602 (E.D. Pa. Nov. 13, 1992) (defendant granted summary judgment with respect to plaintiffs' property damage claims). The plaintiffs presented a somewhat different group of witnesses than it had prior to the first summary judgment and appeal. Judge Kelly analyzes the testimony of each of those witnesses in separate opinions. See *Paoli*, 1992 U.S. Dist. LEXIS 18427-37.

286. Similar concerns undoubtedly underlay Judge Weinstein's opinion in the Agent Orange "opt-out" cases, discussed *infra* text accompanying notes 319-21. The plaintiffs sought recovery for many commonplace ailments including infertility, birth defects, miscarriages, liver disorders, skin rashes, and gastrointestinal disorders. It seems obvious that many instances of those conditions would have occurred among Vietnam veterans and their families without any causal relation to their service in Vietnam. As Judge Weinstein observed: "There are roughly 7500 new cases of Hodgkin's Disease per year in the United States. The fact that seventeen of these persons happen to be Agent Orange plaintiffs proves nothing about the origin of their condition." *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. at 1253 (citation omitted). Judge Weinstein goes on to state: "[Plaintiffs' experts Doctors Singer and Epstein] fail to show how the myriad illnesses at issue are more likely to have been caused by Agent Orange than by something else." *Id.*

distinguish persons whose disease was caused by a toxic substance exposure from those whose diseases were not so caused.

The New Jersey Supreme Court's decision in *Rubanick v. Witco Chemical Co.*,²⁸⁷ another case involving injuries claimed to have resulted from PCB exposure, is an example of that scenario. Plaintiffs, the relatives of two Witco employees who had died of colon cancer, claimed that their decedents' cancers were caused by PCB contamination at the site.²⁸⁸ Their expert on causation was Dr. Balis, a Ph.D. biochemist with extensive research experience on colon cancer.²⁸⁹ Dr. Balis cited studies indicating that PCBs cause cancer in animals, reports on the effects of PCBs on animals and humans, a high rate of cancer at Witco "during the relevant period," and the personal history of one of the deceased, indicating the absence of other risk factors.²⁹⁰ The difficulties of drawing inferences from animal studies and studies of the effects of PCBs in humans paralleled the difficulties in *Paoli*.²⁹¹ Further, the absence of other risk factors is of little probative value where the causes of most instances of a disease are unknown.²⁹²

The New Jersey court focused disapprovingly on the trial court's finding that Dr. Balis' theory of causation was not generally accepted in the scientific community.²⁹³ Citing the need for a more liberal standard for determining the reliability of scientific theories of causation in toxic torts cases,²⁹⁴ the court held that "a scientific theory of causation that has not yet reached general acceptance may be found to be sufficiently reliable if it is based on sound, adequately-founded scientific

287. 593 A.2d 733 (N.J. 1991).

288. *Id.* at 735.

289. *Id.*

290. *Id.* at 735-36.

291. See *supra* notes 260-89 and accompanying text.

292. There also seems to have been considerable doubt about the extent of Rubanick's exposure to PCBs. The New Jersey Supreme Court quotes Dr. Balis' summary of the evidence of exposure as follows:

that there was some thirty-five thousand parts per million PCBs in the soil around there, that he would come home covered with this stuff and the material was oozing out of his clothes, according to I guess it was his wife's testimony, it was something, and I think that report that he lifted these heavy drums and slopping around in this muddy PCB mix, and you also showed me some document about the State of New Jersey, some agency complaining about contamination from that stuff.

Rubanick, 593 A.2d at 736 (quoting testimony of Dr. Balis). One of the defendant's witnesses stated, "[T]here is no evidence that I have seen to this date that would definitively suggest that the individual actually did have extensive exposure to PCBs." *Id.* (quoting testimony of Dr. Fahey).

293. The trial court also found that although Dr. Balis was qualified to offer an opinion on human carcinogenesis generally, he was not a physician and thus was not qualified to offer an opinion on the cause of a specific person's cancer. *Id.* at 737.

294. *Id.* at 740.

methodology involving data and information of the type reasonably relied on by experts in the scientific field."²⁹⁵ The court went on to state that the theory must be offered by an expert with "a demonstrated professional capability to assess the scientific significance of the underlying data and methodology, and to explain the bases for the opinion reached."²⁹⁶

Remanding the case for further proceedings on the admissibility of Dr. Balis' testimony, the court admonished the trial court not to scrutinize the expert's methodology itself to determine its soundness, however, but to refer to the opinions of comparable experts in the field.²⁹⁷ The problem with that recommendation is that it was unclear that the expert used any methodology at all. The trial court's opinion makes it clear that Dr. Balis testified in terms of possibilities, not probabilities.²⁹⁸ Moreover, the trial court had the benefit of the testimony of the defendant's witnesses, who opined that the scientific literature did not support the plaintiffs' expert's opinion that PCBs can cause cancer in humans.²⁹⁹ The court itself had read the articles cited by the plaintiff's expert and concluded that they "do not say what plaintiff's expert concludes."³⁰⁰

Rubanick, like *Paoli*, is a case in which an appellate court appears to approve of toxic tort causation testimony based on uncertain exposure levels, animal testing and other indicators of possible carcinogenicity, despite the absence of any evidence suggesting a connection between PCB exposure and the plaintiff's specific disease. The court cites the appropriate criteria, including reliability, but neither conducts nor allows the trial court to conduct a reliability analysis. Rather than accepting the trial court's findings, which were based on testimony of other experts, the appellate court interjects its own assessment. In reality, the New Jersey Supreme Court's rationale is based almost entirely on the qualifications of the expert, since it harks back to the witness's qualifications as a point of

295. *Id.* at 747-48.

296. *Id.* at 748. The New Jersey Supreme Court emphasized the need for the court to scrutinize the expert's "status" and to direct the jury's attention to "factors that bear relevantly on the expert's credibility." *Id.* at 750.

297. *Id.* at 747-48.

298. When plaintiff's counsel questioned Dr. Balis on the issue of whether the scientific community accepts PCBs as human carcinogens, Balis replied that the people he contacted would not say "probable," but rather that it was a "high possibility." *Rubanick v. Witco Chem. Corp.*, 542 A.2d 975, 983 (N.J. Super. Ct. Law Div. 1988), *rev'd*, 576 A.2d 4, 14 (N.J. Super. Ct. App. Div. 1990) (discussing Dr. Balis' comment), *modified*, 593 A.2d 733 (N.J. 1991).

299. *Id.* at 983.

300. *Id.*

reference for determining the reliability of the data on which the expert relies as well as his methodology.³⁰¹

The problems of deferential review are not restricted to environmental exposure cases; they also occur in products liability cases where the costs of erroneous findings of liability in terms of withdrawal of useful products and disincentives to new product development are perhaps more apparent. *Ferebee v. Chevron Chemical Co.*,³⁰² perhaps the paradigm decision involving deferential review, involved injuries allegedly caused by a pesticide.³⁰³ Plaintiff's causation case was essentially based on the testimony of treating or examining physicians who claimed to have seen a few similar cases. It therefore illustrates the pitfalls of medical testimony based on the coexistence of exposure and disease.

The Bendectin cases are based on a more complex assemblage of evidence and testimony, consisting of chemical structure-activity analysis, *in vitro* testing, animal studies and purported reanalyses of existing epidemiologic studies.³⁰⁴ A review of one of the early cases decided in favor of plaintiffs, *Oxendine v. Merrell Dow Pharmaceuticals, Inc.*³⁰⁵ reveals much of the same evidence that plaintiffs have argued in other cases alleging birth defects caused by Bendectin. The chemical structure-activity analysis consisted of the observation that one of Bendectin's ingredients is an antihistamine and that some antihistamines are teratogenic.³⁰⁶ The *in vitro* and *in vivo* animal test results cited by plaintiffs' witness Dr. Done are subject to the same concerns for high rates of false positives that were discussed previously. The remaining evidence

301. *Rubanick*, 576 A.2d at 7-8, 14. The New Jersey Supreme Court's opinion in *Landrigan v. Celotex Corp.*, 605 A.2d 1079 (N.J. 1992), appears to have tightened the standard for admission of expert testimony. In regard to the issue of whether epidemiologic studies could provide the basis for an expert's opinion on causation, the court stated that such studies "must have been 'soundly and reliably generated' and be 'of a type reasonably relied on by comparable experts in the particular field.' " *Id.* at 1087 (quoting *Rubanick*, 593 A.2d 733). The court went on to state: The court must also examine the manner in which experts reason from the studies and other information to a conclusion [T]hat conclusion must derive from a sound methodology that is supported by some consensus of experts in the field. *Id.* Nonetheless, the court remanded the case for further proceedings on the issue of whether the plaintiff could satisfy the more-probable-than-not standard of proof despite relative risk data that showed less than a doubling of background risk. *Id.* at 1088.

302. 736 F.2d 1529 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984).

303. *Id.* at 1531-32.

304. See generally Sanders, *supra* note 17.

305. 506 A.2d 1100 (D.C. 1986).

306. *Id.* at 1104.

consisted primarily of Dr. Done's unpublished reanalysis of a previous epidemiologic study, which involved selective elimination of data.³⁰⁷

As will be discussed in more detail in Section VI.B.2 of this Article, reanalyses of epidemiologic studies are particularly susceptible to manipulation to achieve a preconceived result. The reanalyses offered by plaintiffs stand in contrast to a large body of epidemiologic evidence that has failed to confirm a statistically significant association between Bendectin and birth defects. Further, the fact that the studies offered by plaintiffs have been unpublished and therefore not subjected to peer review,³⁰⁸ lends further support to other courts' decisions to exclude them.³⁰⁹

B. Active Review Exemplified

In contrast to the deferential, uncritical review accorded expert testimony in *Paoli, Rubanick and Oxendine*, there are a growing number of decisions that utilize active review to make discerning judgments about scientific evidence. Judge Weinstein has been widely criticized for his exclusion of plaintiffs' evidence in the Agent Orange "opt out" cases,³¹⁰ but basically, he got it right. Plaintiffs claimed a wide variety of commonplace ailments, cancers and birth defects as injuries due to Agent Orange or more specifically, the contaminant dioxin.³¹¹ Their evidence consisted of animal studies, and workplace exposure studies that apparently did not indicate an association of dioxin or Agent Orange exposure with the diseases complained of.³¹² There was simply no evidence from which a fact-finder could conclude that any of the plaintiffs suffered from conditions attributable to Agent Orange rather than the causes of such disease in the general population, a fact recognized by the court.³¹³ That conclusion is valid even without taking into account the many studies of

307. See *id.* at 1107-08; see also *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 946-49, 954-57 (3d Cir. 1990) (discussing the failure of Done's reanalysis of other studies to meet traditional statistical significance criteria).

308. See, e.g., *Daubert v. Merrell Dow Pharmaceuticals*, 951 F.2d 1128 (9th Cir. 1991), cert. granted, 113 S. Ct. 320 (1992); *Lynch v. Merrell-Nat'l Lab.*, 830 F.2d 1190, 1194-96 (1st Cir. 1987) (discussing reanalyses by Dr. Done and Dr. Shanna Swan).

309. At least one published reanalysis of epidemiologic data has found no association between Bendectin and birth defects. See *infra* notes 373-76 and accompanying text.

310. See, e.g., *Brennan, supra* note 25, at 9 n.40, 53-56; *Green, supra* note 65.

311. See *Green, supra* note 65, at 659.

312. See *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1236 (E.D.N.Y. 1985) (liver disorders in animal and industrial studies differed from those reported by plaintiffs), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. 1234 (1988).

313. *Id.* at 1239.

Vietnam veterans put before the court that failed to show any increased incidence of serious disease.³¹⁴

The Bendectin litigation has also produced opinions that discerningly review scientific evidence; two such cases are *Brock v. Merrell Dow Pharmaceuticals, Inc.*³¹⁵ and *Lynch v. Merrell-National Laboratories*.³¹⁶ Bendectin has been the subject of over 2000 suits for birth defects allegedly caused by in utero exposure to the anti-nausea drug. Plaintiffs have sought recovery for a variety of malformations, but a number of the cases have involved limb reduction defects.³¹⁷ In *Brock*, the court based its reversal of a jury verdict for plaintiffs on the absence of any statistically significant epidemiologic evidence of an association between Bendectin and birth defects.³¹⁸ Both the *Brock* and the *Lynch* courts concluded that the plaintiffs' *in vitro* testing and animal studies evidence was insufficient, the *Lynch* court noting particularly the inability of *in vivo* and *in vitro* animal studies to prove causation in humans "in the absence of confirmatory epidemiologic data," which it contrasted with a number of studies that failed to find an association between Bendectin and birth defects.³¹⁹

In what has become one of the more controversial aspects of the Bendectin litigation, both courts rejected reanalyses of existing epidemiological studies that purported to show an association between Bendectin exposure and birth defects. The *Lynch* court noted the plaintiffs' failure to file any description of the expected testimony of Dr. Shanna Swan, whose reanalysis of epidemiologic data was offered by plaintiffs. The court went on to examine the basis of Swan's opinion from testimony in other litigation, observing that Swan's control group consisted of children with genetic birth defects, a group that had a lower than background risk for certain types of birth defects, raising the

314. More recent information concerning the hazards of dioxin does not significantly change the analysis. New data on chemical workers exposed to dioxin for more than a year, with more than twenty years' latency, has indicated a 46% increase in all cancers (although not any individual cancer). David J. Hanson, *supra* note 158, at 7, 10. According to Marilyn Fingerhut, the author of the study, the serum levels of dioxin correlated well with duration of exposure. *Id.* at 10. The Fingerhut study is consistent, however, with the Ranch Hand study of Vietnam veterans, which has not detected an increase in cancer at the lower exposure levels experienced by veterans, *id.* at 9, although it has recently revealed significant increases in body fat and diabetes that correlated with dioxin concentration, *id.* at 9. This information does not appear to provide a basis for distinguishing background causes from dioxin for most of the ailments claimed to result from dioxin in the Agent Orange litigation.

315. 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990).

316. 830 F.2d 1190 (1st Cir. 1987).

317. See *Sanders, supra* note 17, at 398-99.

318. See *Brock*, 874 F.2d at 314-15.

319. *Lynch*, 830 F.2d at 1194.

question whether genetic defects might make that control group less susceptible to non-genetic defects such as limb reduction. A lower susceptibility in the control group would skew the relative risk observed for the Bendectin-exposed group.³²⁰

In contrast, the *Brock* court's rationale focused on the fact that the elevated risk found in the reanalysis conducted by Dr. Jay Glasser lacked statistical significance.³²¹ As will be discussed in more detail in the following Section, reanalyses of epidemiological data are susceptible to inadvertent and inadvertent introduction of bias. Although the statistical significance point is arguable, the reanalyses were unpublished and therefore lacked the safeguards against biased or result-oriented data selection that peer-reviewed publication would have provided.

VI. ACTIVE REVIEW: THE ANTIDOTE FOR JUNK SCIENCE

As the foregoing Sections have demonstrated, active review of scientific evidence and expert testimony can go far to eliminate the arbitrary and unfair results that can result from the acceptance of junk science in toxic torts cases. Courts nonetheless cite a number of reasons for deferential review of scientific evidence, including the lack of any special expertise and, perhaps more significantly, the belief that traditional tort law, with its typical reliance on established science, is inadequate to redress toxic injuries of the industrial age. Those reasons, however, do not stand up to careful examination.

A. Courts' Ability to Review Scientific Evidence

As noted previously, one of the concerns regarding scrutiny of scientific evidence is the belief that courts lack the ability to understand scientific evidence and therefore should not deprive the jury of the opportunity to consider possibly relevant and probative evidence. It should be evident from the foregoing discussion, however, that courts are quite capable of determining whether there is a reasoned basis, grounded in fact, for expert opinion, as well as a level of reliability consistent with the applicable standard of proof.

The Third Circuit and the New Jersey Supreme Court, who authored the *Paoli* and *Rubanick* decisions respectively, have demonstrated their understanding of complex scientific evidence. In *DeLuca v.*

320. *Id.* The court also addressed a reanalysis of epidemiologic studies by Dr. Alan Done. Neither the Swan nor the Done study had been published. The court also questioned the bases for exclusion of certain data in each. *Id.* at 1194-96.

321. The Glasser study yielded a relative risk of 1.49, with a confidence interval of 0.17 to 3.0. *Brock*, 874 F.2d at 312. Because the confidence interval included the value 1.0, which represents no increased risk, the result did not satisfy statistical significance criteria. See *infra* notes 367-82 and accompanying text (discussing statistical significance).

Merrell Dow Pharmaceuticals, Inc.,³²² the Third Circuit discussed the epidemiologic evidence on Bendectin and birth defects. At issue was the admissibility of a meta-analysis of existing epidemiologic studies. The meta-analysis in question³²³ did not meet the level of statistical significance³²⁴ typically required for epidemiologic studies.³²⁵ The court's view of the meta-analysis was overly generous,³²⁶ but the opinion demonstrates the court's understanding of the concepts of statistical significance and the effects of bias in epidemiologic studies.³²⁷ Similarly, when confronted with a causation issue on which epidemiologic evidence was offered, the New Jersey Supreme Court also evidenced a sophisticated understanding of such evidence. In *Landrigan v. Celotex Corp.*,³²⁸ the court discussed the proffered epidemiologic evidence and the causal inferences to be drawn from it, as well as the concept of attributable risk.³²⁹

322. 911 F.2d 941 (3rd Cir. 1990).

323. One issue concerning admissibility is publication of studies; one meta-analysis of Bendectin epidemiologic studies has been published, unlike the study offered in DeLuca by Dr. Alan Done. See Thomas R. Einarson et al., *A Method for Meta-Analysis of Epidemiological Studies*, 22 DRUG INTELLIGENCE & CLINICAL PHARMACY 813 (1988); Sanders, *supra* note 17, at 341 n.182.

324. See *DeLuca*, 911 F.2d at 946-48.

325. *Id.* at 955-56.

326. On remand, the district court once again dismissed the plaintiffs' case on summary judgment. See *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 791 F. Supp. 1042 (D.N.J. 1992). The court excluded Dr. Alan Done's reanalysis of epidemiologic studies, finding that Done's calculations and presentation of his results contained numerous errors and his methodology could not be discerned or replicated by the other experts of either plaintiffs or defendants. *Id.* at 1047-48. Plaintiffs' other expert was Dr. Shanna Swan, who has also appeared in other Bendectin cases, including *Daubert v. Merrell Dow Pharmaceuticals, Inc.* In *DeLuca* on remand, she apparently commented on the Done reanalysis based on Done's representations of how it was performed. *Id.* at 1047.

327. The Third Circuit reversed the trial court's summary judgment for defendants and directed the trial court to evaluate the reliability of the proffered evidence "with an eye to all the risks of error posed" by it. *DeLuca*, 911 F.2d at 955. The court further stated, "The root issue . . . is what risk of what type of error the judicial system is willing to tolerate." *Id.* The court suggested that additional expert testimony on statistical significance would be helpful, but also stated a preference for admitting evidence with probative value and "dealing with the risk of error through the adversary process." *Id.* at 956. The ultimate issue, however, is whether that evidence would be sufficient to support a jury finding that Bendectin "more likely than not caused the [plaintiff's] birth defects." *Id.* at 958. The court characterized that requirement as requiring a relative risk greater than two in the exposed population. *Id.*

328. 605 A.2d 1079 (N.J. 1992).

329. *Id.* at 1085-88. The *Landrigan* court, however, refused to hold that a relative risk of 2.0 or greater is required to prove that individual causation was more probable than not. *Id.* at 1087. The court's assumption appears to have been that by ruling out other causes, plaintiff's expert could opine that causation of plaintiff's disease was more probable than not, a proposition that depends on how well developed the evidence is on other risk factors and the relationships among them. *Id.*

No doubt there are times when expert witness testimony and scientific evidence are obscure. The details of statistical significance calculations could undoubtedly lose all but the most mathematically inclined and dedicated lay observer. But judges need not examine expert testimony and scientific evidence at that level of detail. Courts can and should, however, require the proponent of such evidence to demonstrate to the court that the evidence is valid and reliable, that is, that it makes sense and is sufficiently likely to produce an accurate result.

B. Overcompensating for the Deficiencies and Inequities of the Tort System

Another reason courts cite for lenient review of expert testimony is perceived inequities and deficiencies of the tort system. The Sections below examine several aspects of the perception and show that there, too, the cited reasons do not justify the remedy.

1. THE BELIEF THAT MOST CANCERS AND BIRTH DEFECTS ARE CAUSED BY TOXIC PRODUCTS AND ENVIRONMENTAL POLLUTANTS.

Rubanick and *Paoli* illustrate the perception that many, if not most, cancers and birth defects are caused by toxic substances introduced into the environment in products or as waste injuries that they believe will go uncompensated if traditional evidentiary standards are applied. The New Jersey Supreme Court's opinion in *Rubanick* states that concern explicitly:

There are undeniable indications that persons do in fact suffer grave and lethal injury as a result of the wrongful or tortious exposure to toxic substances. Those indications do not spring simply from conjecture; they conform to our common experience and informed intuition. Judge Petrella noted in his opinion below that "[i]t has been widely considered that PCBs are a carcinogenic substance." Our common sense, with some empirical support, tells us of the deleterious effects of PCBs.³³⁰

The Third Circuit expressed similar beliefs in *Paoli*.³³¹ Those perceptions appear to be based on widely quoted statements that most

330. *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 747 (N.J. 1991) (citations omitted).

331. In addressing whether Pennsylvania would recognize a medical monitoring claim, the court noted the need to "accommodate a society with an increasing awareness of the danger and potential injury caused by the widespread use of toxic substances." *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 850 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991). The court went on to note:

The necessity of addressing problems of toxic exposure becomes particularly important with the continued widespread use of chemicals in American industrial and agricultural development. One commentator has pointed out that there are approximately 50,000 hazardous waste sites nationwide. In

cancers are caused by environmental factors.³³² A more careful reading of the sources of such sweeping statements, however, reveals that the

all, over 65,000 chemicals are in commercial use today which have not been tested for their effects on human health or the environment. According to varying estimates, workplace exposure to hazardous substances alone accounts for from five percent to as much as thirty-eight percent of all cancers.

Id. at 850 n.22 (quoting Leslie Gara, Note, *Medical Surveillance Damages: Using Common Sense and the Common Law to Mitigate the Dangers Posed by Environmental Hazards*, 12 HARV. ENVTL. L. REV. 265 (1988)).

332. See *id.* In 1978, David Doniger wrote the following:

From comparisons of different rates of different cancers throughout the world, the World Health Organization and other prominent institutions and individual experts have concluded that 60 to 90 percent of all human cancers are caused by exposure to chemical substances (and, to a lesser extent, radiation) present in our air, workplaces, food, water, and the rest of our environment.

Doniger, *supra* note 131, at 509.

Doniger was not alone in his concern about chemicals and carcinogenesis. *See id.* and references cited therein. Professor Bruce Ames, the developer of the "Ames" mutagenicity test, expressed similar concerns in a 1979 publication recommending mutagenicity assays as methods for identifying mutagens and carcinogens:

A variety of data supports the hypothesis that environmental factors are a major cause of cancer. Epidemiologic studies show different rates of incidence for certain types of cancer in different parts of the world. For example, in Japan there is an extremely low rate of breast and colon cancer and a high rate of stomach cancer, whereas in the United States the reverse is true. When Japanese immigrate to the United States, within a generation or two they show the high colon and breast cancer rates and low stomach cancer rates characteristic of other Americans. Known environmental mutagens that can cause human cancer include cigarette smoke tar, ultraviolet light, x-rays, and asbestos, and the list of human chemical carcinogens is steadily lengthening.

Ames, *supra* note 29, at 587 (citations omitted). Those concerns were prompted in part by the rapid increase in production and exposure of the workforce and the general public to synthetic chemicals. These concerns were summarized by Ames:

Clearly, many more chemicals will be identified as human mutagens and carcinogens. Currently over 50,000 synthetic chemicals are produced and used in significant quantities and close to 1000 new chemicals are introduced each year. Only a small fraction of these were tested for carcinogenicity or mutagenicity before their use. In the past this problem was largely ignored, and even very high-production chemicals with extensive human exposure were produced for decades before adequate carcinogenicity or mutagenicity tests were performed. Such chemicals now known to be both carcinogenic and mutagenic include vinyl chloride (produced at a rate of about 6 billion pounds per year in the United States in 1977) and 1,2-dichloroethane (ethylene dichloride, about 10 billion pounds per year) and a host of high-production pesticides.

The increase in production and use of chemicals has been particularly great since the mid-1950's This flowering of the chemical age may be followed by genetic birth defects and a significant increase in human cancer during the 1980 decade (because of the 20- to 30-year lag) if many of these

environmental factors encompassed by such statements include commonplace causative factors such as background radiation and probably biological processes such as aging, that are largely beyond human control, as well as cigarette smoking, alcohol consumption and dietary factors such as a high fat diet that are the result of lifestyle, not industrial pollutants.³³³

The fraction of cancers and other diseases that could be prevented by reducing or eliminating exposure to man-made toxic chemicals is still in dispute. Studies that have attempted to estimate the fraction of cancers caused by environmental pollution have placed the figure at about six percent, up to as much as fifteen percent.³³⁴ Other exposures and industrial products are thought to add an additional four to five percent, perhaps as much as ten percent.³³⁵

The debate about the role of synthetic chemicals in cancer causation has occurred against a backdrop of increasing cancer rates.³³⁶ The meaning of the data is unclear, however, because most if not all of the increase can be attributed to increases in smoking-related cancers and

chemicals with wide-spread human exposure are indeed powerful mutagens and carcinogens.

Ames, *supra* note 29, at 587-88.

333. Richard Doll and Richard Peto have noted that the phrase "environmental factors" has been "misinterpreted by many people to mean only 'man-made chemicals,' which was certainly not the intent of the WHO committee." Richard Doll & Richard Peto, *The Causes of Cancer: Quantitative Estimates of Avoidable Risk of Cancer in the United States Today*, 66 J. NAT'L CANCER INST. 1192, 1197 (1981). The Doll and Peto article was commissioned as a report to the Office of Technology Assessment of the U.S. Congress. *Id.* at 1193. For a discussion of various avoidable risks, including those of smoking, alcohol use, diet, and other causes, see *id.* at 1220-56.

334. *Id.* at 1256. The six percent figure is the sum of the percentages attributed to occupation, pollution and industrial products. The 15% figure is the sum of the high end of the ranges estimated for each of those sources, which is likely an overestimate because it is the sum of worst case estimates and because the contribution of risk factors is not necessarily additive. See *supra* notes 209-22. This analysis omits medical sources, which include diagnostic X-rays.

335. Doll & Peto, *supra* note 333, at 1256. These figures contrast markedly with a much-cited report filed with the Occupational Safety and Health Administration, which asserted that up to about 40% of all cancers in the U.S. might be occupationally related. NATIONAL CANCER INSTITUTE ET AL., ESTIMATES OF THE FRACTION OF CANCER IN THE UNITED STATES RELATED TO OCCUPATIONAL FACTORS 1 (1978). Doll and Peto point out errors in the methodology of the OSHA report, however, which they believe resulted in overestimation of the proportion of cancers attributable to occupational exposures. Doll & Peto, *supra* note 333, at 1240-41. Causes are not mutually exclusive, however, as the example of asbestos and smoking indicates. See *supra* note 222 and accompanying text. Thus, fractions of total cancer death attributable to various causes could exceed 100%. Doll & Peto, *supra* note 333, at 1219-20. Consequently, attribution of a large fraction of cancers to occupational factors would not necessarily be inconsistent with attributing a similarly large proportion to other factors, such as smoking and alcohol consumption. See ROTHMAN, *supra* note 44, at 14.

336. See, e.g., Earl S. Pollack & John W. Horm, *Trends in Cancer Incidence and Mortality in the United States, 1969-76*, 64 J. NAT'L CANCER INST. 1091 (1980); Smith, *supra* note 33, at 998.

aging of the population.³³⁷ Even if the more pessimistic experts are correct in their conclusion that age adjusted rates are increasing for some cancers,³³⁸ the data do not support the proposition that most cancers are caused by toxic pollution or toxic products (other than cigarettes). Thus, there is no factual basis for a presumption that environmental pollutants cause most cancers.

2. THE BELIEF THAT SCIENCE'S ABILITY TO IDENTIFY CAUSES IS TOO LIMITED.

A second factor, which is related to the belief that toxic substance exposures are causing large amounts of disease, is courts' frustration over the limitations inherent in science's ability to identify causes. Unwilling to accept those limitations, the *Ferebee* court stated:

A cause-effect relationship need not be clearly established by animal or epidemiologic studies before a doctor can testify that, in his opinion, such a relationship exists. As long as the basic methodology employed to reach such a conclusion is sound, such as use of tissue samples, standard tests, and patient examination, products liability law does not preclude recovery until a "statistically significant" number of people have been injured or until science has had the time and resources to complete sophisticated laboratory studies of the chemical. In a courtroom, the test for allowing a plaintiff to recover in a tort suit of this type is not scientific certainty but legal sufficiency; if reasonable jurors could conclude from the expert testimony that paraquat more likely than not caused Ferebee's injury, the fact that another jury might reach the opposite conclusion or that science would require more evidence before conclusively considering the causation question resolved is irrelevant.³³⁹

Not surprisingly, *Ferebee* is widely quoted, particularly by courts that are disposed to admit purported scientific evidence without scrutiny of the underlying reasoning.³⁴⁰ Indeed, the premise of *Ferebee*, namely that the law does not in general require statistical evidence of causation, is hardly subject to dispute. *Ferebee* also appeals to fairness by appearing to

337. Cancer death rates for males from lung cancer have increased dramatically since 1930, while death rates from stomach cancer have steadily declined. See Eliot Marshall, *supra* note 33, at 901. Trends for other common cancers in males are less pronounced. There has been considerable debate about the inferences drawn from the data. A number of statisticians and epidemiologists argue that once the data are adjusted for age and the effects of smoking, the overall incidence of cancer is decreasing. See Smith, *supra* note 33, at 998. Other factors that make interpretation difficult are the effects of increased accuracy of diagnosis and disagreement over the significance of increasing cancer rates among the aged. See Marshall, *supra* note 33, at 901-02.

338. See Marshall, *supra* note 33, at 901.

339. *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529, 1535-36 (D.C. Cir.), cert. denied, 469 U.S. 1062 (1984).

340. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733 (N.J. 1991).

correct the imbalance that disfavors toxic torts plaintiffs, created by the unavailability, high costs, and insensitivity of epidemiologic studies required to link toxic substance exposures to latent injuries.³⁴¹

The inability of epidemiologic studies to detect small increases in risks has been a major concern in the debate over toxic tort causation evidence.³⁴² The power of an epidemiologic study to identify a small increase in risk is a function of the size of the study groups and the background rate of disease, with larger study groups corresponding to greater statistical power.³⁴³ Meta-analysis, in which the data from a number of smaller studies are combined and reanalyzed, can enhance the likelihood of detecting an effect, if one exists.³⁴⁴ Meta-analysis can also provide the opportunity to refine the selection of data included in the analysis to address potential bias in sample selection, as can reanalysis of a single study.³⁴⁵

Systematic error, of which bias is one form, can be introduced into epidemiologic studies in a number of ways, including the failure to control for causal factors other than the factor under study and the failure to accurately delineate exposed and unexposed populations.³⁴⁶ One of the potential sources of bias in the Bendectin studies is recall bias, the possibility that mothers of children born with defects will be more likely to recall drug use during pregnancy than mothers of normal infants. Such recall bias will tend to result, in some kinds of studies, in an overestimation of the effect of the drug.³⁴⁷ Another concern with inaccurate recall is that the "unexposed" group will, in fact, have some individuals who were exposed and who exhibit effects caused by the exposure.³⁴⁸ If there is an effect, part of that effect will be attributed to the

341. Epidemiologic studies require considerable time and money to conduct, often beyond the means of the toxic tort plaintiff. Further, epidemiologic studies are a crude method for detection of small increases in disease that have long latency periods and significant background risks. See Brennan, *supra* note 25, at 54; Doll & Peto, *supra* note 333, at 1219; see also Ames, *supra* note 29, at 587 (advocating the use of short-term assays to identify mutagens and carcinogens). Further, epidemiologic studies are often designed to detect a relative risk of two or more. See M.J. Adams Jr. et al., *The Use of Attributable Fraction in the Design and Interpretation of Epidemiologic Studies*, 42 J. CLINICAL EPIDEMIOLOGY 659, 659 (1989).

342. Green and Brennan have argued that such insensitivity requires that plaintiffs be allowed to resort to other kinds of toxicological evidence to prove their cases. See Brennan, *supra* note 25, at 56; Green, *supra* note 65, at 680-81.

343. See ROTHMAN, *supra* note 44, at 79-80. Power is "the probability of detecting (as 'statistically significant') a postulated level of effect." *Id.* at 79.

344. See Naylor, *supra* note 277, at 892.

345. Both meta-analyses and reanalyses of existing studies have been at issue in the Bendectin litigation. See *supra* notes 65-70 and accompanying text.

346. See ROTHMAN, *supra* note 44, at 82-94.

347. *Id.* at 85. This possibility applies to case control studies. *Id.*

348. See *id.* at 85-87.

unexposed group, tending to diminish the magnitude of the observed effect.³⁴⁹

To counter the negative epidemiologic evidence that predominates in the published literature concerning Bendectin, several plaintiffs have variously offered meta-analyses or reanalyses by one or both of two expert witnesses, Dr. Alan Done, a professor of pediatrics and pharmacology, and Dr. Shanna Helen Swan, an epidemiologist and chief of the a unit of the California Department of Health Services.³⁵⁰ Meta-analyses are subject to questions about the propriety of combining data from studies in which the original criteria for selection of subjects and controls differed.³⁵¹ Both meta-analyses and reanalyses involve selection of data for inclusion and exclusion, which create the opportunity for "data dredging" that may turn up statistically significant correlations that are actually due to chance.³⁵² The methodology by which data were selected for inclusion and exclusion in meta-analyses and reanalyses should therefore be carefully scrutinized.

A number of objections can be made to the reanalyses and meta-analyses offered by various Bendectin plaintiffs. In the case of Dr. Done's reanalysis at issue in *Lynch*, the basis of the data selection seems less than clear,³⁵³ although the *Oxendine* opinion indicates that in the reanalysis offered by Done in that case, some pairs of exposed and unexposed children were eliminated because Done considered the risk of recall bias to be especially high among Canadian subjects who could have purchased Bendectin without a prescription.³⁵⁴ Dr. Shanna Swan's methodology is explained more completely in *Lynch*; it involved the reanalysis of data previously analyzed by four members of the Center for Disease Control.³⁵⁵ All of the subjects in the original group had involved abnormal children. Swan reanalyzed the data, using only children with genetic abnormalities as the control group so that the control group's abnormalities could not have resulted from Bendectin.³⁵⁶ Her reanalysis

349. See *id.*

350. See *Lynch v. Merrell-Nat'l Lab.*, 830 F.2d 1190, 1194-95 (1st Cir. 1987). In *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990), plaintiffs offered a reanalysis by Dr. Jay Glasser. *Id.* at 312.; see *supra* note 321 and accompanying text.

351. See Naylor, *supra* note 277, at 893. Although Naylor is discussing the aggregation of data from clinical trials rather than retrospective exposure cases, the argument still applies.

352. See *id.*, at 894.

353. *Lynch*, 830 F.2d at 1196.

354. *Oxendine v. Merrell Dow Pharmaceuticals, Inc.*, 506 A.2d 1100, 1107-08 (D.C. 1986).

355. See *Lynch*, 830 F.2d at 1195; see also Eliot Marshall, *Supreme Court to Weigh Science*, 259 SCIENCE 588, 590 (1993).

356. *Lynch*, 830 F.2d at 1195.

concluded that Bendectin is associated with an increased risk of birth defects. Swan's reanalysis raises questions because the control group for her reanalysis was acknowledged to have only a 0.57 relative rate (i.e., a 40% lower rate) for certain categories of birth defects. As the First Circuit noted, "Swan made no allowance for the possibility that the very fact of having such a severe genetic deficiency as Down's Syndrome might operate to make other rare deficiencies such as limb reduction less likely," thus skewing the apparent differences between the exposed and control groups.³⁵⁷ The possibility that both Done's and Swan's reanalyses were based on result oriented "data dredging" or other inadvertent introduction of bias cannot be ignored; further, none of the reanalyses or meta-analyses has been published in peer-reviewed scientific journals, although ample time has elapsed for review and publication.³⁵⁸ The failure of either to publish their results leaves courts without any reassurance that concerns about bias are unwarranted.

In any event, the insensitivity of epidemiologic studies in the case of Bendectin is probably an overrated concern. Although the studies cannot be said to eliminate all possibility that Bendectin is teratogenic, they at least indicate that if Bendectin is a teratogen, it is a weak one.³⁵⁹ Moreover, the insensitivity of epidemiologic studies does not improve the probative value of other evidence. Animal studies, mutagenicity testing and structure-activity relationships do not become more persuasive because of the absence of other kinds of proof.

3. THE BELIEF THAT SCIENTISTS REQUIRE TOO MUCH CERTAINTY.

A third argument courts cite for abandoning scientific criteria for proof of causation is the perception that scientists require too great a degree of certainty before they will accept a factual proposition as established. *Rubanick v. Witco Chemical Corp.*³⁶⁰ makes numerous references to the high level of proof required by scientists³⁶¹ and concludes that "the scientific method . . . fails to address or accommodate the needs and goals of the tort system."³⁶² That scientists may require a higher level of certainty than the legal system may in some instances be true. In part, the mismatch between expert testimony and legal requirements is the result of the failure of lawyers and courts to articulate legal requirements

357. *Id.*

358. A meta-analysis offered by Dr. Done in another case also failed to satisfy statistical significance criteria. See *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941, 954-57 (3d Cir. 1990); *Lynch*, 830 F.2d at 1195-96.

359. See *Sanders*, *supra* note 17, at 348.

360. 593 A.2d 733 (N.J. 1991).

361. See *id.* at 737, 739-41.

362. *Id.* at 741.

of proof to scientists. Unless the examining attorneys explore what a scientist-expert witness means by "proof," the risk that the expert understands such terms differently from their legal meaning will always exist.³⁶³

The appropriate level of certainty is particularly an issue when epidemiologic evidence that does not meet epidemiologists' criteria for certainty is offered. Epidemiologists typically are unwilling to conclude that increased disease incidence in an exposed population is associated with a toxic substance exposure unless a statistical analysis of the data shows that the probability of a false positive is 5% or less.³⁶⁴ That requirement represents a 95% confidence level,³⁶⁵ a level that is considerably higher than the more probable than not standard would seem to suggest.³⁶⁶ Moreover, the 5% cutoff for statistical significance is

363. In *Rubanick*, it was clear that the plaintiffs' expert was discussing possibilities and was unable to state that it was more probable than not that PCBs caused the decedents' colon cancers. *Rubanick v. Witco Chem. Corp.*, 576 A.2d 4 , 14-15 (N.J. Super. Ct. App. Div. 1990), modified, 593 A.2d 733 (N.J. 1991); see *supra* note 305 and accompanying text. The mismatch is also due to mechanical application of the *Frye* rule. When the *Frye* general acceptance test is applied to an expert's opinion on whether a toxic substance can cause a particular disease, the test incorporates scientists', rather than the legal system's, standards of proof.

364. See *supra* note 103 and accompanying text.

365. The statistical analysis sometimes focuses on the calculation of a "p-value," which represents the probability that the relative risk produced by the study is due to random variability or chance. See *ROTHMAN, supra* note 44, at 115-19. Often, an upper limit for the p-value is selected as $p = 0.05$; if the p-value of the study falls at or below the cutoff, the results of the study are said to be "statistically significant." A p-value of 0.05 corresponds to a five percent chance that an increase in relative risk is actually a false positive, described as a type I error or alpha-error. *Id.* Alternatively, the statistical analysis may be used to generate confidence intervals, that is, ranges of relative risk that are associated with a specified level of confidence. A 95% confidence interval is the range in which the relative risk would be expected to fall 95% of the time if the study were repeated (hence, a 95% confidence level). *Id.* at 119-20. The confidence level is equal to one minus the probability of type I error; thus, a 95% confidence level corresponds to a statistical significance cutoff value of p equal to 0.05. See *id.* at 119. In *Brock v. Merrell Dow Pharmaceuticals Inc.*, 874 F.2d 307 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990), the court discussed epidemiologic data for which the statistical analysis was expressed in terms of confidence intervals. See *id.* at 312. The court recognized that where a confidence interval includes a relative risk of 1.0 (which represents no effect from the exposure), the study could not be said to demonstrate a statistically significant increased risk of limb defects associated with exposure. *Id.* at 312-13. Confidence intervals are usually calculated for a predetermined confidence level, usually 90% to 95% but occasionally lower. *ROTHMAN, supra* note 44, at 119.

366. This issue was explicitly raised in *DeLuca v. Merrell Dow Pharmaceuticals, Inc.*, 911 F.2d 941 (3d Cir. 1990), in which the plaintiffs sought to present Dr. Done's reanalysis of epidemiologic data as evidence of an increased relative risk associated with *in vitro* Bendectin exposure. Dr. Done's reanalysis did not satisfy statistical significance criteria. *Id.* at 955.

arbitrary, and has been used to meet epidemiologists' perceived needs for certainty.³⁶⁷

Some commentators have questioned whether statistical significance is relevant to the more probable than not standard of proof.³⁶⁸ Green suggests that focus on the relative risk found in a study is more appropriate.³⁶⁹ That approach seems untenable, however, because it fails to distinguish the issue of whether exposure to the toxic substance causes any effect at all, which is the function of statistical significance testing, from the issue of the likelihood that a particular plaintiff's case resulted from the exposure rather than background or other causes, a conclusion that is inferred from the magnitude of the relative risk.

Relative risk greater than 1.0 in an exposed population is sufficient evidence of an association of disease with exposure only if we can be reasonably certain that the unequal distribution of disease in exposed and unexposed populations is not due to chance. Ignoring the possibility that an increased incidence of disease is due to chance leads to the obviously absurd result that a disease cluster, no matter how small, could be argued as sufficient evidence of an association between an exposure and disease, a result that is indefensible. The evaluation of the role of chance in an epidemiologic study is thus an essential part of determining the probative value of the evidence.

The appropriate confidence level is a more difficult question, however. At a minimum, the more probable than not standard of proof would seem to tolerate epidemiologic data on the issue of general causation if there is less than a 50% probability that the result is due to chance, a confidence level far lower than the 95% level typically employed by epidemiologists. Additionally, Green and others have noted that typical statistical significance testing is concerned only with the risk of false positives, that is, the risk that an effect will be inferred when there is actually no effect.³⁷⁰ The legal system is also concerned, however, with the risk of false negatives, namely, in toxic torts the risk

367. ROTHMAN, *supra* note 44, at 118-19. The use of $p = 0.05$ lessens the possibility that an effect will be assumed when, in fact, there is no association between exposure and disease incidence. The use of low p -values, however, increases the probability that no association will be assumed when, in fact, there is an association. *Id.* The considerations that have led epidemiologists to require 95% confidence interval as a cut-off for statistical significance are not necessarily appropriate for tort law. Commentators have been unable to agree on appropriate alternatives, however.

368. See Green, *supra* note 65, at 682, 687; David H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1334 (1986) (decrying the mechanical application of statistical significance criteria without explanation and suggesting confidence interval testing as more useful).

369. See Green, *supra* note 65, at 647.

370. See *id.* at 683.

that no effect will be detected when there actually is an effect.³⁷¹ Decreasing the risk of false positives tends to increase the risk of false negatives, though not in a straightforward way.³⁷² Thus, there is an argument that in some instances, epidemiological studies should be admitted with less stringent significance criteria than are typically applied. Before such a rule, which significantly lowers the standard of acceptability of epidemiologic evidence of increased risk, is adopted, however, it would be well to consider other sources of error. Epidemiologic studies are plagued to a greater or lesser degree with other, nonrandom sources of error. Exposure data can be highly uncertain. There is always the possibility that there are unknown confounding causes that are not randomly distributed between the exposed and control populations. Although statistical testing usually does not address nonrandom error, the possibility of other confounding factors may have a great deal to do with the high confidence levels that epidemiology has typically required to minimize the risk of error due to chance.

It may be instructive to consider Bendectin because it has been the subject of over thirty epidemiologic studies and at least one published meta-analysis of those studies.³⁷³ If statistical significance criteria are indeed too stringent, causing scientists to miss a real effect, one would expect to see relative risks from the various studies falling above 1.0 more often than below that number. In other words, the results should vary around the "true" relative risk even if no single study qualifies as statistically significant.³⁷⁴ In his comprehensive study of the Bendectin litigation, Sanders notes that of twenty-six studies from which he was able to extract a value indicative of relative risk, thirteen reported a value greater than one, twelve reported values less than one, and one study reported a value of exactly one.³⁷⁵ That result is roughly consistent with a published meta-analysis of seventeen prior studies that concluded that

371. David Kaye has analyzed the preponderance of the evidence rule as having the effect of minimizing erroneous verdicts. See David H. Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation*, 1982 AM. B. FOUND. RES. J. 487, 496-503. In the statistical analysis of epidemiologic studies, the assumption that there is no effect where there is, in fact, an effect (i.e., a false negative) is referred to as type II or beta-error. ROTHMAN, *supra* note 44, at 117-18.

372. See ROTHMAN, *supra* note 44, at 117-18.

373. See Sanders, *supra* note 17, at 341 n.182.

374. This is a commonsense application of the rationale behind the meta-analysis of existing studies, in which smaller studies are combined to obtain larger sample and control populations. Meta-analysis runs the risk of comparing populations that differ in nonrandom ways, however, and thus some caution is warranted in drawing conclusions in the casual manner suggested in the text. See also ROTHMAN, *supra* note 44, at 334-36 (discussing trend estimation based on differing exposure levels even where individual studies do not satisfy statistical significance criteria).

375. Sanders, *supra* note 17, at 340-41.

Bendectin is not associated with human birth defects.³⁷⁶ If statistical significance criteria were lowered to a 50% confidence level, one is left to wonder whether both plaintiffs and defendants would be offering "statistically significant" evidence, respectively, that Bendectin causes and prevents birth defects. Thus, it is not clear without further evaluation that scientific confidence level criteria are too stringent where epidemiologic evidence is concerned.

A more basic concern with courts' perceptions that scientists require too much certainty is that such views seem to form the basis for rejection of scientific reasoning altogether. The problem with *Ferebee* and its progeny is that they fail to recognize that in most cases,³⁷⁷ there are no alternative proofs available that amount to anything more than speculation or estimation with a great deal of uncertainty.³⁷⁸ Courts' unwillingness to scrutinize testimony on disease causation leaves the door open to the self-validating experts who can be found to testify to virtually any proposition.³⁷⁹ Even the courts that have deemed such evidence admissible have recognized the hazards of their approach.³⁸⁰ Nonetheless, they are willing to risk that kind of error because scientific evidence is unavailable to satisfy traditional standards of proof.³⁸¹ The irony of that rationale is that it rests on courts' and commentators' acceptance and even distortion of scientific speculation that widespread dissemination of new chemicals might result in increases in cancer, birth defects and other disease. Having accepted scientific speculation, they then reject the cautionary statements of scientists who want greater certainty before they reach conclusions.

C. The Costs of Overcompensation

The position taken herein runs counter to the views of several recent commentators. Troyan Brennan has urged courts to admit and consider all the kinds of evidence that toxicologists bring to bear on the question of whether a substance causes disease, including animal studies, short term

376. See *id.* at 341 & n.182.

377. See *supra* note 57 and accompanying text.

378. See *supra* notes 64-67 and accompanying text (discussing structure-activity relationships, short-term testing, and animal studies).

379. See *In re "Agent Orange" Prod. Liab. Litig.*, 611 F. Supp. 1223, 1242 (E.D.N.Y 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987), *cert. denied*, 487 U.S. Ct. 1234 (1988).

380. See, e.g., *Rubanick v. Witco Chem. Corp.*, 593 A.2d 733, 744 (N.J. 1991) ("There are, assuredly, genuine concerns engendered by a test of reliability of complex scientific theories of causation that does not fully embrace the views of a dominant or of a significant segment of the scientific community.").

381. See, e.g., *id.* at 745 (other courts' demands for "near-scientific certainty are unrealistic" because the level of scientific proof is unavailable).

assays and structure-activity relationships.³⁸² Michael Green goes even further, urging courts to approve of all of the foregoing and even individual case reports as a sufficient evidentiary basis for plaintiffs' verdicts.³⁸³ Moreover, those commentators do not significantly disagree with this author about the uncertainty inherent in those kinds of evidence.³⁸⁴ They do, however, differ on the conclusions reached in the face of those uncertainties.

Brennan's primary suggestion is to propose that questions involving significant scientific uncertainty be resolved by referring those questions to court-appointed experts or science panels.³⁸⁵ There are obviously cases, however, that are not significant enough to warrant science panels, or perhaps even court-appointed experts. Moreover, Brennan does not really come to grips with how evidence with such uncertain probative value can satisfy the more probable than not standard of proof, whether reviewed by a science panel or a lay jury. He recognizes that the acceptance of evidence associated with a high degree of uncertainty is a policy question, but does not provide a rationale for such a radical change in policy.³⁸⁶

Green, on the other hand, recognizes that difficulty. His solution is equally troubling: He states that "plaintiffs should be required to prove causation by a preponderance of the *available evidence*."³⁸⁷ This proposal is at least directly addresses the problem with animal studies and other, even more uncertain kinds of proof. The problem that Green's and Brennan's proposals present, however, is that they create potentially unlimited and ultimately arbitrary liability for cancer, birth defects, and other diseases that lack definitive causal explanations. Rare will be the cancer victim who cannot find some arguably toxic exposure, whether it be the pesticide application on the neighbor's lawn, pumping her own gas at the gas station or other such cause. Rarer still will be the plaintiff who cannot find a treating physician or other expert who is willing to state that based on past experience and review of the literature, that a particular toxic substance exposure is consistent with the plaintiff's disease and that the plaintiff lacked other predisposing factors. Reliance on the available evidence when such evidence suggests only the

382. See Brennan, *supra* note 25, at 21-26.

383. See Green, *supra* note 65, at 646, 674-75.

384. See Brennan, *supra* note 25, at 21-26 (discussing the kinds of uncertainty associated with animal tests, short term assays, and epidemiologic evidence); Green, *supra* note 65, at 680-81 (discussing animal testing, *in vitro* testing, short-term assays, structure-activity analysis, and case studies).

385. See Brennan, *supra* note 25, at 62-71. He suggests that science panels and lists of potential experts be coordinated under a federal science board.

386. See Brennan, *supra* note 15, at 523-32.

387. Green, *supra* note 65, at 680 (emphasis added).

possibility, not the probability, of causation suggests that plaintiffs would do well to proceed to court when the evidence on whether a substance can cause disease is in an unformed stage. Such plaintiffs apparently will not have to contend with the messy questions of distinguishing background risk or other known risks that become issues when epidemiologic evidence is available. Indeed, they would have no basis for making such distinctions.

If there were a way to ease plaintiffs' evidentiary burdens without opening the door to arbitrary and potentially devastating liability for defendants, it would undoubtedly garner considerable support. The zone of uncertainty about the role of toxic chemicals in the causation of many diseases is simply too wide however, to suggest a reasonable way to split the difference.

It must be noted that courts' concerns are not all scientific. Other policy concerns, sometimes unspoken but often implied, seem to underlie courts' willingness to entertain unfounded and poorly reasoned evidence. Those concerns are the indignation and outrage felt by the public in general and plaintiffs in particular over exposure to contaminants or products involving substances suspected of causing harm or whose properties are simply unknown.³⁸⁸ In many of the environmental exposure cases, the exposures or the contamination that could have led to exposure occurred without the plaintiffs' knowledge or consent.³⁸⁹ In the case of potentially toxic products such as breast implants, the exposures have occurred with implicit or explicit assurances that the products were safe.

Traditional tort doctrines, however, do not provide for compensation for egregious conduct without causally related physical injury unless it rises to the level of intentional infliction of emotional distress.³⁹⁰ Commentators have suggested creation of causes of action based on creation of risk,³⁹¹ and a limited number of courts have adopted

388. Studies of risk perception have documented the phenomenon that public acceptance of risk is adversely influenced by the involuntariness of the risk. See Paul Slovic, *Perception of Risk*, 236 SCIENCE 280, 283 (1987).

389. See, e.g., *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 835 (3d Cir. 1990), cert. denied, 111 S. Ct. 1584 (1991); *Sterling v. Velsicol Chem. Corp.*, 855 F.2d 1188 (6th Cir. 1988); *Renaud v. Martin Marietta Corp.*, 749 F. Supp. 1545 (D. Colo. 1990), aff'd, 972 F.2d 304 (10th Cir. 1992).

390. See generally 1 DORE, *supra* note 5, §§ 4.01-05, 7.01-08. Recovery based on negligent infliction of emotional distress has been traditionally limited to cases involving physical impact or injury. 1 *id.* § 7.02[2], at 7-3. The limitations of this doctrine have been mitigated somewhat by courts' relaxation and broadening the notion of physical impact to include exposure or subclinical changes. See 1 *id.* at 7-4 to 7-5.

391. See, e.g., *Robinson, Probabilistic Causation*, *supra* note 50, at 783.

such theories.³⁹² Those theories are implicitly and sometimes explicitly premised on assumptions that some significant level of risk can be proved,³⁹³ assumptions that in many cases would be erroneous.³⁹⁴

In any event, the tort system is probably not the best forum for addressing public concerns over uncertain risk. The inability of toxic tort claimants to prove causation has been one of the more important rationales for environmental regulation.³⁹⁵ Indeed regulation is an area in which risk is explicitly recognized as a basis for restricting the dissemination of a substance in products or in the environment. Regulation does not compensate those who are injured despite regulation or by unregulated risks, but it has an important role to play in minimizing risks.

However desirable it might be to have the tort system fill all the gaps where toxic injury occurs, the current state of knowledge simply does not permit the necessary causal connections to be made. Given that state of affairs, what is at stake is whether the "more probable than not" standard of proof will continue to apply to toxic torts. Whether that burden should be lessened or even shifted to defendants are policy issues of the greatest importance. They should be addressed directly and changes, if any, should be based on their fullest consideration of the implications. To effect a reallocation of burdens of proof under the pretext of admitting reliable evidence which is in fact not probative, is not the appropriate way to bring about a change in such a fundamental principle of tort law.

392. The claims have been variously cast as claims for emotional distress, increased risk of future injury, and medical monitoring costs. *See generally* 1 DORE, *supra* note 5, §§ 7.01-08.

393. *See 1 id.* § 7.07, at 7-16.5, 7-27.6 (citing cases refusing to recognize claims based on unquantified risk of injury).

394. *See supra* notes 140-79 and accompanying text for a discussion of uncertainty in risk estimation from nonepidemiologic evidence.

395. The breast implant controversy, however it is ultimately resolved, represents a holdover from a period in which medical devices did not require approval by the Food and Drug Administration, a situation that does not apply to new devices.

ARTICLE

ANTITRUST AND INTERNATIONAL COMPETITIVENESS: IS ENCOURAGING PRODUCTION JOINT VENTURES WORTH THE COST?

DONALD K. STOCKDALE, JR.[†]

Table of Contents

I.	INTRODUCTION	270
II.	CURRENT PROPOSALS TO CHANGE THE ANTITRUST LAWS	272
III.	THE POTENTIAL BENEFITS AND COSTS OF RESEARCH JOINT VENTURES AND PRODUCTION JOINT VENTURES	274
	A. The Inefficiencies in Market Generated R&D	274
	B. The Social Benefits and Costs of RJs	276
	C. The Lesser Benefits and Greater Costs of PJs	280
	D. Organizational Difficulties Associated with All Joint Ventures	286
IV.	CURRENT LAW AND ENFORCEMENT POLICY CONCERNING PRODUCTION JOINT VENTURES AND THEIR EFFECT ON SUCH VENTURES	289
	A. Courts Judge PJs Under the Rule of Reason	290
	B. Current Antitrust Enforcement Policy Is Hospitable Towards Legitimate PJs	293
	C. The Threat of Private Antitrust Suits Is Exaggerated	294
	D. Current Antitrust Law Has Not Deterred the Formation of an Increasing Number of PJs	296

© 1993 Donald K. Stockdale, Jr.

[†] Assistant Professor, College of Business & Management, University of Maryland. Ph.D. 1989, Yale University; J.D. 1980, Yale Law School; B.A. 1976, King's College, Cambridge University; B.A. 1974, Yale University.

V. ARGUMENTS FAVORING THE PROPOSED SPECIAL TREATMENT OF DOWNSTREAM CONSORTIA ARE FLAWED.....	297
A. The High Costs and Risks of Commercializing and Producing Innovative Products Do Not Exceed the Capacity of Most Individual Firms	298
B. The Cyclical Nature of the Development Process and Shorter Product Lives Does Not Necessitate Cooperation in Both R&D and Production.....	299
C. International Competition Will Not Eliminate the Danger of Collusion Among the Joint Venture Participants	300
D. Use of Joint Ventures in Japan and Europe Does Not Require that the U.S. Encourage Domestic Joint Ventures	302
VI. WEAKNESSES IN THE PROPOSED LEGISLATION.....	309
VII. CONCLUSION	314

I. INTRODUCTION

For more than a decade, slower productivity growth, persistently large trade deficits, and the apparent decline of the international competitiveness of U.S. firms have concerned policy-makers, business leaders and academicians.¹ In analyzing the causes of these ominous trends, many have questioned whether the U.S. antitrust laws have unduly disadvantaged domestic firms relative to their foreign competitors.

In the late 1970s, many commentators began suggesting that cooperative research and development (R&D) warranted special treatment under the antitrust laws.² Congress responded by passing the National Cooperative Research Act of 1984 (NCRA).³

1. See, e.g., *Competitiveness and Antitrust: Hearings Before the Senate Comm. on the Judiciary*, 100th Cong., 1st Sess. (1987) [hereinafter 1987 Senate Hearings]; MARTIN N. BAILEY & ALOK K. CHAKRABARTI, INNOVATION AND THE PRODUCTIVITY CRISIS (1988); WILLIAM J. BAUMOL ET AL., PRODUCTIVITY AND AMERICAN LEADERSHIP: THE LONG VIEW (1989); MICHAEL L. DERTOZOS ET AL., MADE IN AMERICA: REGAINING THE PRODUCTIVE EDGE (1989); INTERNATIONAL TRADE ADMIN., U.S. DEP'T OF COMMERCE, AN ASSESSMENT OF U.S. COMPETITIVENESS IN HIGH TECHNOLOGY INDUSTRIES (1983); NATIONAL RESEARCH COUNCIL, TECHNOLOGY, TRADE, AND THE U.S. ECONOMY (1978); PRESIDENT'S COMM'N ON INDUS. COMPETITIVENESS, GLOBAL COMPETITION: THE NEW REALITY (1985).

2. See, e.g., *Japanese Technological Advances and Possible U.S. Responses Using Research Joint Ventures: Hearings Before the Subcomm. on Investigations and Oversight and the Subcomm. on Science, Research and Technology of the House Comm. on Science and Technology*, 98th Cong., 1st Sess. (1983); *The National Productivity and Innovation Act and Related Legislation: Hearings Before the Senate Comm. on the Judiciary*, 98th Cong., 1st & 2d Sess. (1983 & 1984); INDUSTRIAL RESEARCH INST., INSTITUTIONAL AND LEGAL CONSTRAINTS TO COOPERATIVE

Recently, however, a number of academicians and business leaders have suggested that the NCRA did not go far enough. They argue that in order to improve the international competitiveness of domestic firms, Congress should enact further legislation to encourage joint ventures in downstream activities, such as production and even distribution and marketing.⁴ Responding to such arguments, members of the 101st and 102d Congresses introduced bills which, in various ways, would relax the antitrust laws for production joint ventures (PJs) and, in some cases, for distribution and marketing joint ventures.⁵

This Article argues that such proposals are misguided, and that, if implemented, they would likely undermine American competitiveness and impose significant costs on U.S. consumers. More specifically the Article contends that: (1) the potential social benefits are lower and costs higher for PJs, in comparison with research joint ventures (RJs); (2) the

ENERGY R&D (Technical Advisory Bd., U.S. Commerce Dep't, No. PB-240-929, 1975); NATIONAL RESEARCH COUNCIL, ANTITRUST, UNCERTAINTY, AND TECHNOLOGICAL INNOVATION (1980).

3. The NCRA mandated that research joint ventures (RJs), as defined in the Act, should not be deemed illegal per se, but rather should be evaluated under the rule of reason. It further provided that RJ participants would be liable in private actions for only single, rather than treble, damages if they filed a notification with the Antitrust Division and the Federal Trade Commission. Finally, it enabled RJ participants which had been sued by private plaintiffs to recover attorneys' fees and costs under certain conditions, regardless of whether the RJ had filed a notification. 15 U.S.C. §§ 4301-4305 (1988).

4. See, e.g., *Legislation Concerning Production Joint Ventures: Hearing Before the Subcomm. on Antitrust, Monopolies and Business Rights of the Senate Comm. on the Judiciary*, 101st Cong., 2d Sess. (1990) [hereinafter 1990 Senate Hearing]; *The Government Role in Joint Production Ventures: Hearing Before the Subcomm. on Science, Research and Technology of the House Comm. on Science, Space, and Technology*, 101st Cong., 1st Sess. (1989) [hereinafter 1989a House Hearing]; *Production Joint Ventures Antitrust Legislation: Hearings Before the Subcomm. on Economic and Commercial Law of the House Comm. on the Judiciary*, 101st Cong., 1st Sess. (1989) [hereinafter 1989b House Hearing]; *High Definition Television: Hearing Before the House Comm. on Science, Space, and Technology*, 101st Cong., 1st Sess. (1989) [hereinafter 1989c House Hearing]; Thomas M. Jorde & David J. Teece, *Innovation, Cooperation and Antitrust*, 4 HIGH TECH. L.J. 1 (1989) [hereinafter Jorde & Teece (1989a)]; Thomas M. Jorde & David J. Teece, *Competition and Cooperation: Striking the Right Balance*, 31 CAL MGMT. REV. 25 (1989) [hereinafter Jorde & Teece (1989b)]; Thomas M. Jorde & David J. Teece, *Innovation and Cooperation: Implications for Competition and Antitrust*, 4 J. ECON. PERSP. 75 (1990) [hereinafter Jorde & Teece (1990)].

5. In the 101st Congress the following bills were introduced: S. 952, 101st Cong., 1st Sess. (1989); S. 1006, 101st Cong., 1st Sess. (1989); H.R. 423, 101st Cong., 1st Sess. (1989); H.R. 1024, 101st Cong., 1st Sess. (1989); H.R. 1025, 101st Cong., 1st Sess. (1989); H.R. 2264, 101st Cong., 1st Sess. (1989). The House of Representatives eventually passed H.R. 4611, 101st Cong., 2d Sess. (1990). See also H.R. REP. NO. 516, 101st Cong., 2d Sess. (1990). No bill was passed in the Senate, however.

In the 102d Congress, new bills were introduced that would provide similar relief. See S. 479, 102d Cong., 1st Sess. (1991); H.R. 1604, 102d Cong., 1st Sess. (1991). See also S. REP. NO. 146, 102d Cong., 1st Sess. (1991), reprinted in 61 Antitrust & Trade Reg. Rep. (BNA) 347 (1991); H.R. REP. NO. 972, 102d Cong., 2d Sess. (1992).

antitrust laws currently permit procompetitive PJs, and, in fact, the wide employment of these joint ventures renders further relaxation unnecessary; (3) further relaxing the antitrust laws for downstream joint ventures may encourage the formation of production consortia having substantial market power; and (4) even if antitrust relief were warranted for production consortia in certain strategically important high-technology industries, none of the current legislative proposals is specifically tailored to that goal.

The Article is organized as follows. Part II describes the specific legislation proposed. Part III compares the potential social costs of RJs and downstream JVs and suggests that for PJs the potential benefits are more limited, while the potential costs are much higher. This Part further argues that production consortia in particular tend to impose significant social costs. Part IV examines existing antitrust precedents and antitrust enforcement policy and contends that, with the possible exception of joint ventures possessing substantial market power, current law does not pose an obstacle to joint venture activity. Part V addresses and criticizes certain specific arguments that have been raised in favor of relaxing the antitrust laws for joint ventures in high-technology industries. Part VI evaluates the specific legislative proposals that have been introduced and suggests that they are unlikely to achieve their purported goals.

II. CURRENT PROPOSALS TO CHANGE THE ANTITRUST LAWS

During the 102d Congress,⁶ the Judiciary Committees of the House and the Senate approved and sent to their respective floors bills that would extend the NCRA to cover production joint ventures.⁷ Although the Senate passed a slightly modified version of the bill, the House

6. In the 101st Congress, members introduced several bills that would have amended the antitrust laws to provide various protection for production joint ventures. Basically, the bills adopted one or more of the following four approaches: (1) extending the notification procedures and protections of the NCRA to joint ventures involving production (and in some cases marketing), *see* H.R. 1025, 101st Cong., 1st Sess. (1989); H.R. 2262, 101st Cong., 1st Sess. (1989); S. 1006, 101st Cong., 1st Sess. (1989); (2) codifying in detail the substantive law applicable to innovative joint ventures, *see* H.R. 1024, 101st Cong., 1st Sess. (1989); S. 2322, 101st Cong., 1st Sess. (1989); (3) establishing a safe harbor for PJs whose participants lack market power, *see* H.R. 423, 101st Cong., 1st Sess. (1989); and (4) establishing a certification procedure under which joint ventures, reviewed and approved by the relevant antitrust authorities, would be exempt from any antitrust penalty or damage liability, *see* H.R. 1024, 101st Cong., 1st Sess. (1989); S. 2322, 101st Cong., 1st Sess. (1989). *See generally* H.R. REP. NO. 516, *supra* note 5; Joseph F. Brodley, *Antitrust Law and Innovation Cooperation*, 4 J. ECON. PERSP. 97, 104 (1990). The House ultimately passed H.R. 4611, which adopted the first approach, but no bill reached the floor of the Senate during that Congress.

The bills introduced in the 102d Congress adopted only the first approach.

7. *See* H.R. 1604, 102d Cong., 1st Sess. (1991); S. 479, 102d Cong., 1st Sess. (1991).

adjourned without acting. Nevertheless, the provisions of the bills remain significant because the next Congress will likely introduce similar legislation.

Under both bills, the NCRA's definition of joint venture would expand to include "the production of a product, process or service" in addition to covering research and development activities.⁸ Thus, PJVs that qualify under the bills would receive rule of reason analysis if challenged under the antitrust laws. In addition, qualified production ventures that file a notification⁹ with the antitrust authorities would be liable only for actual, not treble, damages in actions filed by private plaintiffs. Finally, regardless of whether the venture files a notification, it would be able to recover attorneys' fees and costs if it were named a defendant in an antitrust suit and the court finds the claim was "frivolous, unreasonable, without foundation, or in bad faith."¹⁰

Both bills specifically prohibit the joint marketing of any products jointly produced by the venture.¹¹ At the same time, however, neither bill requires that a production joint venture engage in any joint R&D activities to qualify for protection.

Both bills would also add a new section directed specifically at PJVs. The new section in the House bill would exclude a PJV from protection of the Act "if at any time more than 30 percent, in the aggregate, of the beneficial ownership of the voting securities and equity of such joint venture is controlled by foreign entities." The section would also require that any facilities operated by the venture be located in the United States or its territories.¹² The Senate bill establishes two different conditions for a PJV to qualify under the Act: first, the venture must provide "substantial benefits" to the U.S. economy (such as "increased skilled job opportunities," "investments in long-term production facilities," or "participation by United States entities in the venture"); second, the production facilities of the venture must be located in the United States or in a country that accords "national treatment" to American participants in PJVs.¹³

8. H.R. 1604, *supra* note 7, § 2(b)(4); S. 479, *supra* note 7, § 2(2)(c).

9. The information required to be provided in a notification is limited. For example, the House bill only requires that the joint venture provide the identities of the participants and a brief description of the nature and objectives of the venture. *See, e.g.*, H.R. REP. NO. 516, *supra* note 273, at 19.

10. *See* S. REP. NO. 146, *supra* note 5, at 23.

11. H.R. 1604, *supra* note 7, § 2(c)(3)(B); S. 479, *supra* note 7, § 2(2)(G).

12. H.R. 1604, *supra* note 7, § 2(f). According to the Committee Report, the section is intended to "stimulate more collaborative activity by American-owned firms." H.R. REP. NO. 516, *supra* note 5, at 15.

13. S. 479, *supra* note 5, § 2(10). According to the Senate Report, the requirements are intended to ensure that the act benefits American workers. S. REP. NO. 146, *supra* note 5, at 7.

The Senate bill contains two additional provisions not found in the House version. First, the Senate bill requires that, if a joint venture uses existing facilities, those facilities must produce a "new product or technology."¹⁴ Second, the Senate bill imposes new reporting requirements on the Federal Trade Commission and the Department of Commerce.¹⁵

III. THE POTENTIAL BENEFITS AND COSTS OF RESEARCH JOINT VENTURES AND PRODUCTION JOINT VENTURES

The rationale for giving special treatment to cooperative research stems principally from certain market failures associated with market generated R&D. These market failures can create inefficiencies in the level of R&D investment, the allocation of R&D expenditures, and the dissemination of the R&D results.¹⁶ Before comparing the potential benefits and costs of RJs relative to downstream joint ventures, it is useful to review these market failures associated with R&D.

A. The Inefficiencies in Market Generated R&D

The special problems connected with R&D activities result principally because the product of R&D activities is *information* or *knowledge*. Information resembles a public good, in that (1) the acquisition of the information by one party need not reduce its availability to others, and (2) the cost of transferring the information to others is often, though not always, low. The public good nature of information creates problems both for private firms engaging in R&D and for society as a whole.

The most widely recognized inefficiency of privately funded R&D is the generation of positive externalities: that is, the benefits of the R&D

14. S. 479, *supra* note 5, § 2(2)(G).

15. Specifically, the bill requires the FTC to prepare an annual report listing the joint ventures that had filed under the Act and any enforcement actions that had been brought by the Department of Justice against ventures filing under the Act. The bill requires the Department of Commerce to prepare triennial reports which describe the "technologies most commonly pursued by joint ventures" (and assess the competitiveness of U.S. industry in those technologies), describe the areas of production most commonly engaged in by PJs, and review foreign laws concerning joint R&D and production. See S. 479, *supra* note 5, § 2(10).

16. See generally Gene M. Grossman & Carl Shapiro, *Research Joint Ventures: An Antitrust Analysis*, 2 J.L. ECON. & ORGANIZATION 315 (1986); Michael L. Katz, *An Analysis of Cooperative Research and Development*, 17 RAND J. ECON. 527 (1986); Michael L. Katz & Janusz A. Ordover, *R&D Cooperation and Competition*, 1990 BROOKINGS PAPERS ON ECON. ACTIVITY: MICROECONOMICS 139; Janusz Ordover & William Baumol, *Antitrust Policy and High-Technology Industries*, 4 OXFORD REV. ECON. POL'Y 13 (1988); Janusz A. Ordover & Robert D. Willig, *Antitrust for High-Technology Industries: Assessing Research Joint Ventures and Mergers*, 28 J.L. & ECON. 311 (1985).

frequently spill over from the researching firm to others. Because firms cannot appropriate the full rewards or benefits of their investment in R&D, they will tend to invest less than the socially optimal amount.¹⁷ The severity of the inappropriability and underinvestment problems increases with more basic research.¹⁸ In addition, the lumpiness of R&D inputs and economies of scale and scope in R&D may exacerbate this underinvestment.¹⁹

The patent system and trade secrecy laws are intended to alleviate this appropriability problem by assigning and enforcing property rights in the information produced by R&D.²⁰ Unfortunately, these mechanisms for increasing appropriability create other problems.

First, these mechanisms result in an inefficient *ex post* dissemination of the knowledge produced by R&D. That knowledge or information can be used simultaneously by others at little or no extra cost suggests that society should encourage its widest possible dissemination. By utilizing exclusion to increase the appropriability of knowledge, society creates inefficiencies in its *ex post* dissemination.²¹

17. In other words, the private return on investment in R&D will be less than the social return. See Kenneth J. Arrow, *Economic Welfare and the Allocation of Resources for Invention*, in THE RATE AND DIRECTION OF INVENTIVE ACTIVITY 609, 619-622 (R. Nelson ed., 1962); Richard R. Nelson, *The Simple Economics of Basic Scientific Research*, 67 J. POL. ECON. 297, 302 (1959). In addition, because a firm can gain from the R&D of others, it reduces the competitive risk of failing to conduct independent R&D. Katz & Ordover, *supra* note 16, at 39.

18. See Partha Dasgupta, *The Welfare Economics of Knowledge Production*, 4 OXFORD REV. ECON. POL'Y 1, 4 (1988); Katz, *supra* note 16, at 537; Nelson, *supra* note 17, at 302-04.

19. R&D inputs are said to be lumpy because large minimum expenditures are often required before any R&D can be performed or before such R&D can yield any useful results. See WILLIAM D. NORDHAUS, *INVENTION, GROWTH, AND WELFARE* 36 (1969); JEAN TIROLE, *THE THEORY OF INDUSTRIAL ORGANIZATION* 414 (1988); see also Partha Dasgupta, *The Economic Theory of Technology Policy: An Introduction*, in *ECONOMIC POLICY AND TECHNOLOGICAL PERFORMANCE* 9 (Partha Dasgupta & Paul Stoneman eds., 1987). In addition, there is evidence that R&D frequently exhibits significant economies of scale and scope. See NORDHAUS, *supra*, at 414; Grossman & Shapiro, *supra* note 16. Finally, imperfections in the capital markets may limit the availability of firms to obtain outside funding for R&D investment. See Paul Stoneman & John Vickers, *The Assessment: The Economics of Technology Policy*, 4 OXFORD REV. ECON. POL'Y i, viii (1988). These facts suggest that the market may not yield an efficient investment in R&D, and more particularly, that the level of investment necessary for the efficient performance of certain types of R&D may exceed the financial resources of smaller firms..

20. As many have shown, however, the patent system and trade secrecy laws in general fail to eliminate all spillovers. See, e.g., Richard C. Levin et al., *Appropriating the Returns from Industrial Research and Development*, 1987 BROOKINGS PAPERS ON ECON. ACTIVITY 783 (1987); Edwin Mansfield et al., *Social and Private Rates of Return from Industrial Innovation*, 91 Q.J. ECON. 221 (1977).

21. If the information generated is valuable enough, it may confer market power on the innovating firm, which can lead to higher prices and reduced output. This will generate the static, deadweight loss associated with monopoly. See Arnold C. Harberger, *Monopoly and Resource Allocation*, 44 AMER. ECON. REV. 77, 78 (1954); Richard R. Nelson & Sidney Winter, *The Schumpeterian Tradeoff Revisited*, 72 AMER. ECON. REV. 114, 116 (1982).

In addition, to the extent that the patent laws enable an innovator to capture a significant proportion of the social benefits in the form of profits, a race to be first may result in too many firms engaging in duplicative R&D. As a result, the patent laws may create inefficient and possibly excessive investments in R&D.²²

B. The Social Benefits and Costs of RJs

The three above-mentioned inefficient aspects of market generated R&D in turn suggest the three most significant potential benefits of cooperative research. First, RJs can help internalize the externality caused by the inappropriability of R&D and can thus increase R&D investment incentives. This internalization occurs because the RJV compels the participants to commit to sharing costs before the research is conducted and hence before any spillovers can occur. This benefit is likely to be greatest when the RJV is directed at basic research²³ or at research involving areas of limited commercial importance, such as that directed to environmental, health and safety problems, because the

In addition, restricting dissemination of information concerning the most efficient technology can raise the average production cost in the industry over that which would result with widespread use of the new technology. *Id.* Finally, denying competitors the right to use new technological information can induce them to spend research funds on inventing around the patent or on developing new technologies that are less efficient than the current best, but inaccessible technology. See Donald K. Stockdale, Jr., *Three Essays on Antitrust and Innovation* 66 (1989) (unpublished Ph.D. dissertation, Yale University). See generally Katz & Ordover, *supra* note 16, at 145.

22. This excessive investment in a particular area of research has been termed the "common pool" problem. See Partha Dasgupta & Joseph Stiglitz, *Industrial Structure and the Nature of Innovative Activity*, 90 ECON. J. 266, 279 (1980); Partha Dasgupta & Joseph Stiglitz, *Uncertainty, Industrial Structure, and the Speed of R&D*, 11 BELL J. ECON. 1, 3 (1980); J. Hirshleifer & John G. Riley, *The Analytics of Uncertainty and Information*, 17 J. ECON. LIT. 1375, 1404 (1979); Pankaj Tandon, *Rivalry and the Excessive Allocation of Resources to Research*, 14 BELL J. ECON. 152 (1983); see also Partha Dasgupta & Eric Maskin, *The Simple Economics of Research Portfolios*, 97 ECON. J. 581 (1987).

23. The National Science Foundation has defined the various categories of research and development as follows:

Basic Research—Original investigations for the advancement of scientific knowledge not having specific immediate commercial objectives, although such investigations may be in fields of present or potential interest to the . . . company.

Applied Research—Investigations directed to the discovery of new scientific knowledge having specific commercial objectives with respect to products or processes . . .

Development—Technical activities of a nonroutine nature concerned with translating research findings or other scientific knowledge into products or processes . . .

NATIONAL SCIENCE FOUND., RESEARCH AND DEVELOPMENT IN INDUSTRY: 1987, at 2 (1989).

results of such research are the least appropriable.²⁴ In addition, benefits increase when a high percentage of firms in an industry participate in basic or externalities research.²⁵

Second, by providing access to all participants, an RJV may improve the *ex post* dissemination of the information produced by the RJV.²⁶

Third, by replacing a number of independent and competing research centers with a joint facility, an RJV may reduce excessive R&D expenditures associated with a race to be first. More importantly, the RJV may eliminate wasteful duplication in research and hence use research expenditures more efficiently.²⁷

In addition to the benefits that arise from the special characteristics of R&D, RJVs also encompass the more traditional benefits of joint ventures. Like other joint ventures, RJVs permit the participants to share risks and costs, combine complementary skills and resources, and take advantage of economies of scale and scope.²⁸ These advantages are likely to prove especially appealing for smaller firms that lack the skills or resources to conduct R&D on their own.

RJVs may also impose social costs, however, by adversely affecting R&D activity and by reducing other forms of competition. These potential costs may be grouped into three types.

First, if the RJV participants are competitors in the downstream product market, the RJV may reduce the expected return to each participant and, hence, total investment in R&D. Because all cooperating firms have equal and simultaneous access to the results of the R&D, no firm will enjoy a temporary monopoly return resulting from the innovation; rather, price competition among the participants will dissipate any excess profits from the innovation, with the surplus going to consumers. Accordingly, the RJV participants may cut back on R&D investment.²⁹ This reduction is less likely to occur, however, if: (1) the

24. See Grossman & Shapiro, *supra* note 16, at 332-33; Katz, *supra* note 16, at 537; Nelson, *supra* note 17, at 302-04; see also P.S. JOHNSON, CO-OPERATIVE RESEARCH IN INDUSTRY (1973).

25. Increasing the percentage of firms participating in the RJV will reduce the externality by reducing the number of free-riders and by committing the beneficiaries of the research-generated information to share its costs *ex ante*. See Grossman & Shapiro, *supra* note 16, at 321; Stockdale, *supra* note 21, at 67. But this result will not necessarily hold for RJVs directed to applied research and development. See *infra* text accompanying notes 29-32.

26. See, e.g., Grossman & Shapiro, *supra* note 16, at 323.

27. See *id.* at 322; Katz, *supra* note 16, at 528; Ordover & Baumol, *supra* note 16, at 27.

28. 1989a House Hearing, *supra* note 4, at 68-69 (statement of Claude E. Barfield, American Enterprise Institute); DAVID C. MOWERY & NATHAN ROSENBERG, TECHNOLOGY AND THE PURSUIT OF ECONOMIC GROWTH 239 (1989); Grossman & Shapiro, *supra* note 16, at 321-22; Katz, *supra* note 16, at 528-29.

29. See Katz & Ordover, *supra* note 16, at 152, 156; Katz, *supra* note 16. The RJV may also be used to suppress innovation, where implementation of the innovation would

research has no immediate commercial objective (such as basic research), (2) the participants operate in separate downstream product markets,³⁰ (3) there are strong nonparticipants performing competing research, or (4) the RJV agreement permits participants to continue their independent R&D efforts.³¹ Finally, besides possibly reducing investment in R&D, an RJV may also reduce the productivity of the R&D performed by limiting the diversity of approaches to a research problem. This would tend to offset the gains described above.³²

Second, an RJV may reduce competition in the downstream product market(s), which will generate social costs when firms limit output and raise prices to consumers. Participants have a clear incentive to maximize joint returns to any innovation generated by the joint venture, either by cooperating in production³³ or by employing ancillary restraints, such as field of use or geographic restrictions, to restrain product market competition.³⁴ Participants may use the RJV as a forum for exchanging price and cost data in order to collude in the downstream product markets.³⁵ In addition, the RJV may serve as a means for extending cooperation or collusion into other product areas.³⁶

Third, by denying access to the RJV or to its research results, participants may disadvantage, and possibly drive from the market, actual and potential competitors. Although this is most likely to occur

impose significant costs on the participants or destabilize the industry, *see Johnson, supra* note 24, at 84, or where the RJV is used to delay meeting government environmental or safety regulations. *See United States v. Automobile Mfrs. Ass'n*, 1969 Trade Cas. (CCH) ¶ 72,907 (C.D. Cal. 1969) (consent decree), *modified*, 1982-1983 Trade Cas. (CCH) ¶ 65,088 (C.D. Cal. 1982); *LAWRENCE A. SULLIVAN, HANDBOOK OF THE LAW OF ANTITRUST* 301-03 (1977).

30. *See Katz, supra* note 16, at 529.

31. *See id.* at 542; *Grossman & Shapiro, supra* note 16, at 324.

32. CARMELA S. HAKLISCH ET AL., *TRENDS IN COLLECTIVE INDUSTRIAL RESEARCH* 153 (2d ed. 1986); MOWERY & ROSENBERG, *supra* note 28, at 240; Ordover & Baumol, *supra* note 16, at 27; *see also* John T. Scott, *Diversification Versus Cooperation in R&D Investment*, 9 *MANAGERIAL & DECISION ECON.* 173 (1988) (suggesting that NCRA may have reduced diversity and productivity of R&D effort); *cf. JOHN JEWKES ET AL., THE SOURCES OF INVENTION* 221 (1960); RICHARD R. NELSON & SIDNEY G. WINTER, *AN EVOLUTIONARY THEORY OF ECONOMIC CHANGE* 366 (1982) (discussing possible insufficient diversification of research efforts under monopoly).

33. *See infra* text accompanying note 58.

34. *See Grossman & Shapiro, supra* note 16, at 325; Katz & Ordover, *supra* note 16, at 156; *cf. F.M. SCHERER & DAVID ROSS, INDUSTRIAL MARKET STRUCTURE AND ECONOMIC PERFORMANCE* 625 (3d ed. 1990) (discussing use of cross-licensing agreements and patent pools to facilitate collusion and exclude competitors); George L. Priest, *Cartels and Patent License Agreements*, 20 *J.L. & ECON.* 309, 356-77 (1978).

35. *See Katz & Ordover, supra* note 16, at 156. The NCRA attempts to alleviate this danger by specifically excluding the exchange of cost and price data from the protection of the Act. 15 U.S.C. § 4302(b) (1988); *see also* S. REP. NO. 427, 98th Cong., 2d Sess. 15-16 (1984) (explaining reasons for exclusions from protection).

36. *See Katz & Ordover, supra* note 16, at 156; Stockdale, *supra* note 21, at 71.

where horizontal competitors make up the RJV, it may also occur where dominant firms in different markets cooperate to produce a new product or process.³⁷ This exclusionary behavior will only create significant social costs when both upstream and downstream markets are concentrated with high barriers to entry and reentry,³⁸ and when the participants can successfully collude with respect to price, R&D, and other dimensions of competition. This is frequently difficult to achieve.³⁹

Thus, RJVs may be used for anticompetitive purposes and may impose net social costs under certain circumstances. These anticompetitive effects are most likely to occur where: (1) the cooperation extends downstream to areas of competitive concern, (2) the relevant markets are concentrated and exhibit barriers to entry, (3) the combined market power of the participants is significant, and (4) collateral restraints in the agreement restrict competition among the participants.

Therefore, RJVs may have a positive or negative effect on social welfare; assessing the net welfare effect of a particular RJV requires an examination of the particular facts. Nevertheless, certain types of RJVs most likely to yield a net benefit to society are identifiable.

For example, an RJV directed at basic or precommercial research is likely to generate significant benefits without imposing substantial social costs. Such an RJV will likely increase industry expenditures on basic research by permitting the sharing of costs and risks and by internalizing

37. See, e.g., *Berkey Photo, Inc. v. Eastman Kodak Co.*, 603 F.2d 263, 299-304 (2d Cir. 1979), cert. denied, 444 U.S. 1093 (1980) (Kodak's joint development with flash manufacturer of new camera flash held to violate section 1 of the Sherman Act); cf. Janusz A. Ordover & Robert D. Willig, *An Economic Definition of Predation: Pricing and Product Innovation*, 91 YALE L.J. 8 (1981) (analyzing introduction by single firm of new, but incompatible product system, as a form of predation).

38. See Grossman & Shapiro, *supra* note 16, at 317; Stockdale, *supra* note 21, at 71-72; cf. Paul L. Joskow & Alvin K. Klevorick, *A Framework for Analyzing Predatory Pricing Policy*, 89 YALE L.J. 213, 227-33 (1979) (private incentives for, and social costs resulting from, predatory conduct will be significant only in the presence of substantial market power and barriers to entry).

39. The following factors, among others, have been identified as limiting the effectiveness of or possibilities for oligopolistic collusion: (1) a large number of competitors, (2) a large variance in the size of competitors, (3) relatively free entry conditions, (4) differentiated products, (5) differential cost structures among competitors, (6) relatively elastic demand, (7) growing demand, and (8) significant non-price competition. See SCHERER & ROSS, *supra* note 34, at 277-315; Peter Asch & Joseph J. Seneca, *Characteristics of Collusive Firms*, 23 J. INDUS. ECON. 223 (1975); George J. Stigler, *A Theory of Oligopoly*, 72 J. POL. ECON. 55 (1964), reprinted in GEORGE J. STIGLER, THE ORGANIZATION OF INDUSTRY 39-63 (1968). See generally Alexis Jacquemin & Margaret E. Slade, *Cartels, Collusion, and Horizontal Merger*, in 1 HANDBOOK OF INDUSTRIAL ORGANIZATION 415 (Richard Schmalensee & Robert D. Willig eds., 1989). In addition, it is recognized that successful collusion is especially difficult in industries subject to rapid technological change. *Id.* at 420.

the externalities associated with such research.⁴⁰ It may also increase the efficiency of the R&D by reducing duplication.⁴¹ Moreover, because basic research and precommercial R&D are distanced from the competitive concerns of the market, the RJV will not likely spur collusion. Finally, a research-directed JV will unlikely injure nonparticipants because of the significant research spillovers and because the generally long lag between generation of the idea and its commercial application gives nonparticipants time to catch up.⁴²

Likewise, if the relevant research and product markets are unconcentrated with relatively free entry, it appears improbable that an RJV made up of a nonmajority of firms in those markets or of noncompetitors can impose significant social costs, since the participants would still face stiff competition from nonparticipants in both upstream and downstream markets. Such a venture could yield substantial benefits in the form of increased R&D expenditures and increased efficiency in the performance of R&D, however.⁴³ Moreover, in evaluating the conditions of the research market, it is generally accepted that the relevant geographic market is global in scope.⁴⁴ Accordingly, even if the RJV consists of a majority of U.S. competitors, this may not result in anticompetitive effects if there are foreign competitors with sufficient research capabilities.

In summary, the NCRA justly encourages the above-described RJVs in general because the social benefits outweigh the social costs.

C. The Lesser Benefits and Greater Costs of PJVs

In contrast to RJVs, joint ventures in production and distribution offer fewer social benefits and pose greater social costs. Although the market failures associated with R&D suggest that it often makes sense to include as many participants as possible in an industry-wide research consortia, no similar arguments justify industry-wide production

40. Based on survey and interview data, Wolek found that U.S. research consortia "are significantly more committed to basic research than are competitive, industrial programs." He further found that, on average, consortia devoted 23.4% of their budget to basic research in 1974. FRANCIS W. WOLEK, THE ROLE OF CONSORTIA IN THE NATIONAL R&D EFFORT (National Science Found., NTIS No. PB-277-366, 1977). Similarly Haklisch, Fusfeld, and Levenson found that 89% of the RJVs that they surveyed performed fundamental research, and that fundamental research represented 32% of the RJVs overall activities. HAKLISCH ET AL., *supra* note 32, at 18.

41. Grossman & Shapiro, *supra* note 16, at 333.

42. *Id.*; Katz, *supra* note 16, at 537.

43. See Grossman & Shapiro, *supra* note 16, at 326; Katz, *supra* note 16, at 540.

44. See, e.g., Antitrust Div., U.S. Dep't of Justice, Antitrust Guidelines for International Operations-1988, 4 Trade Reg. Rep. (CCH) ¶ 13,109.10, at 20589-3 (1989) [hereinafter International Guidelines]; William F. Baxter, *Antitrust Law and Technological Innovation*, 1 ISSUES SCI. & TECH. 80, 85 (1985); Ordover & Baumol, *supra* note 16, at 30.

consortia. To illustrate these differences, this Section will focus on production and distribution joint ventures that do not involve cooperation in research.⁴⁵

Because the production and distribution of goods and services do not suffer from the same market failures affecting R&D, the major justifications for research cooperation—internalizing the externalities associated with R&D, improving the *ex post* dissemination of research results, and eliminating wasteful duplication of research efforts—do not apply to PJVs. Rather the potential advantages of domestic PJVs⁴⁶ are considerably more narrow.

First, a PJV may permit the realization of economies of scale or scope, where the minimum efficient scale of a plant is beyond the capacity of individual companies or is large relative to total demand.⁴⁷ Empirical studies generally agree, however, that the minimum efficient scale of plant is small relative to market size in the vast majority of industries, and this ratio of scale to market size has been declining over time in many industries.⁴⁸ Furthermore, even in the most scale-intensive industries, numerous competing production facilities can coexist.⁴⁹ This suggests that, although economies of scale may justify production joint ventures between two or three smaller firms,⁵⁰ they do not justify

45. This limitation is chosen not only for expositional simplicity, but also because most of the bills currently being considered by Congress would not require PJVs to perform cooperative research to qualify for protection.

46. Joint ventures involving United States and foreign firms may be based on additional motivations, most importantly, the desire to gain access to foreign markets. See MICHAEL E. PORTER, THE COMPETITIVE ADVANTAGE OF NATIONS 66 (1990); David C. Mowery, *Collaborative Ventures Between U.S. and Foreign Manufacturing Firms: An Overview*, in INTERNATIONAL COLLABORATIVE VENTURES IN U.S. MANUFACTURING 12-15 (David C. Mowery ed., 1988); MOWERY & ROSENBERG, *supra* note 28, at 248-50.

47. See, e.g., J. PETER KILLING, STRATEGIES FOR JOINT VENTURE SUCCESS 7-8 (1983); PORTER, *supra* note 46, at 66; Robert Pitofsky, *Joint Ventures under the Antitrust Laws: Some Reflections on the Significance of Penn-Olin*, 82 HARV. L. REV. 1007, 1015 (1969); Carl Shapiro & Robert D. Willig, *On the Antitrust Treatment of Production Joint Ventures*, 4 J. ECON. PERSP. 113, 114 (1990).

48. See, e.g., C.F. PRATTEN, ECONOMIES OF SCALE IN MANUFACTURING INDUSTRY (1971) (compilation of studies for 25 industries); Leonard W. Weiss, *Optimal Plant Size and the Extent of Suboptimal Capacity*, in ESSAYS ON INDUSTRIAL ORGANIZATION IN HONOR OF JOE S. BAIN (Robert T. Masson & P. David Qualls eds., 1975). See generally SCHERER & ROSS, *supra* note 34, at 111-20 (reviewing empirical studies of minimum efficient scale relative to market size).

49. For example, in Japan there are nine competing automobile producers, six competing manufacturers of mainframe computers, and 34 competing producers of semiconductor chips. PORTER, *supra* note 46, at 412; cf. 1989b House Hearing, *supra* note 4, at 379 (statement of George Gilder) (in industries in which the Japanese surpassed the United States, "they had at least four times as many competitors in the marketplace").

50. Even for joint ventures among small numbers of firms there is reason to question the importance of scale economies as a motivating factor. For example, Mariti and Smiley found that only 11 of 70 cooperative agreements studied indicated that achieving economies of scale in production was a major motivating factor, and of those 11, six were

production consortia involving many or most of the firms in an industry.⁵¹

Second, PJVs permit the sharing of costs and risks, especially in cases involving uncertain demand or a new technology.⁵² The risks involved in producing a new product are generally significantly less, however, than the risk that basic or fundamental applied research will yield a reasonable return.⁵³

Third, PJVs may generate synergies resulting from the sharing of complementary assets and skills of the participants.⁵⁴ Again, however, the synergies resulting from joint production should not exceed those from joint research. Nor does it appear, in general, that a joint venture requires large numbers of cooperating firms to achieve such synergies.⁵⁵

Finally, RJV participants may benefit by extending cooperation from research into production. As previously noted, firms cooperating in R&D may dissipate any returns from the R&D by competing among themselves in the downstream product market.⁵⁶ To avoid such dissipation, firms may agree to cooperate in producing and/or marketing the results of the R&D or to limit downstream competition through the use of collateral restraints, such as geographic or field-of-use restrictions. Although such strategies are facially anticompetitive, they may be necessary to secure cooperation among the participants. Moreover, they

in the automobile industry. P. Mariti & R.H. Smiley, *Cooperative Agreements and the Organization of Industry*, 31 J. INDUS. ECON. 437, 445 (1983); cf. Jeffrey Pfeffer & Phillip Nowak, *Patterns of Joint Venture Activity: Implications for Antitrust Policy*, 21 ANTITRUST BULL. 315, 328 (1976) (in a survey of 163 joint ventures, the median level of assets and sales of participants exceeded \$500 million, suggesting that firms of this size did not require joint ventures to achieve economies of scale).

51. See 1987 Senate Hearings, *supra* note 1, at 128 (statement of Richard C. Levin).

52. See 1990 Senate Hearing, *supra* note 4, at 23 (statement of Assistant Attorney General James F. Rill); Shapiro & Willig, *supra* note 47, at 114.

53. In the former case, the risks concern whether the product will prove commercially successful. In the latter case, however, additional uncertainties exist concerning whether the research will yield any information that could lead to a new product or process, in addition to generally longer lag times before these uncertainties are resolved. See EDWIN MANSFIELD ET AL., THE PRODUCTION AND APPLICATION OF NEW INDUSTRIAL TECHNOLOGY 22-32 (1977); MOWERY & ROSENBERG, *supra* note 28, at 214; Dasgupta, *supra* note 18, at 6; see also Arrow, *supra* note 17, at 616 (discussing uncertainty connected with basic research); Nelson, *supra* note 17, at 298-300.

54. 1990 Senate Hearing, *supra* note 4, at 24 (statement of Assistant Attorney General James F. Rill); Shapiro & Willig, *supra* note 47, at 114.

55. Empirical studies of PJVs suggest that most involve a small number of firms. See, e.g., SANFORD V. BERG ET AL., JOINT VENTURE STRATEGIES AND CORPORATE INNOVATION 35 (1982) (in survey of chemical joint ventures, 90% had only two parents); Albert N. Link & Gregory Tassey, *Editors' Introduction to COOPERATIVE RESEARCH AND DEVELOPMENT: THE INDUSTRY-UNIVERSITY-GOVERNMENT RELATIONSHIP* vii-viii (Albert N. Link & Gregory Tassey eds., 1989) (two-firm joint ventures are most common for applied R&D).

56. See *supra* text accompanying notes 29-32.

are unlikely to impose significant social costs if the participants are few in number and collectively lack market power.⁵⁷

While the potential benefits of PJVs appear less than those of RJs, the potential anticompetitive effects are far greater. Most importantly, the PJV may have anticompetitive effects in the relevant product market. Where the participants are horizontal competitors and the joint venture controls a major portion of the production assets in the market, the participants will have a clear incentive to maximize their joint profits by reducing output and increasing price.⁵⁸

The PJV also increases the likelihood of either tacit or explicit collusion among the participants in other downstream product markets⁵⁹ and in upstream research markets.⁶⁰ Discussions concerning the appropriate prices for the joint venture's products may lead to discussions and collusion concerning the prices charged for products the participants

57. See Baxter, *supra* note 44, at 89-91; Grossman & Shapiro, *supra* note 16, at 332; Ordover & Baumol, *supra* note 16, at 30.

58. See, e.g., Katz & Ordover, *supra* note 16, at 156; Katz, *supra* note 16, at 541; Shapiro & Willig, *supra* note 47, at 114-15; cf. Joseph F. Brodley, *Joint Ventures and Antitrust Policy*, 95 HARV. L. REV. 1521, 1552 (1982) ("Of all joint ventures, the horizontal is inherently the most anticompetitive . . . [T]he parents, through their representatives in the joint venture, will necessarily agree on prices and output in the very market in which they themselves operate.").

Even if the participants distribute the joint venture's product independently, they can accomplish the same socially costly goal by raising the price at which the joint venture transfers its product to the participants.

Moreover, even if the participants did not control the pricing of the joint venture's product and do not coordinate their actions, their common ownership interests result in the internalization of a competitive externality which can lead to an increase in price-cost margins. Robert J. Reynolds & Bruce R. Snapp, *The Competitive Effects of Partial Equity Interests and Joint Ventures*, 4 INT'L J. INDUS. ORGANIZATION 141, 142 (1986). See generally Timothy F. Bresnahan & Steven C. Salop, *Quantifying the Competitive Effects of Production Joint Ventures*, 4 INT'L J. INDUS. ORGANIZATION 155 (1986) (examining effect on competitive incentives of non-cooperating oligopolists participating in joint ventures under alternative control arrangements).

59. See, e.g., Daniel R. Fusfeld, *Joint Subsidiaries in the Iron and Steel Industry*, 48 AMER. ECON. REV. 578, 585 (1958) (hypothesizing that joint ventures could be a mechanism through which emerging industries could be dominated by existing large firms in related industries); Walter J. Mead, *The Competitive Significance of Joint Ventures*, 12 ANTITRUST BULL. 819, 820-21 (1967) (finding that joint ventures formed to bid on government-owned property resulted in restrained bidding on subsequent bids).

60. If participants in a PJV collectively account for a large percentage of the competitors in the relevant product market, and if the participants do not independently manufacture goods that compete with the joint venture's products, then this may result in a significant reduction in research effort, since the participants need not worry that they will be preempted by new products resulting from other participants' research. See PORTER, *supra* note 46, at 621; cf. MOWERY & ROSENBERG, *supra* note 28, at 99 (weak British antitrust policy between the wars led to price and market-sharing agreements and "undercut the incentives for the pursuit of competitive advantage through innovation").

manufacture independently.⁶¹ Further, the joint venture will reduce the likelihood that individual participants would attempt to cheat on any collusive agreement because the ongoing relationship creates disincentives.⁶² The likelihood of collusion, moreover, is generally recognized as significantly greater in the case of production and distribution joint ventures involving direct competitors than with RJs involving direct competitors, especially RJs directed to basic or precompetitive research.⁶³ Such collusion will also be more likely when the combined market power of the participants is greater and the barriers to entry in the affected markets are higher.

Finally, the PJV may injure competition by excluding non-participants from an essential input. This "essential facilities" problem will most likely occur where competitors possessing market power organize a vertical joint venture to supply a particular, relatively unavailable, input. It may also occur, however, in the case of a horizontal joint venture, where participants deny competitors access to new technology or to a more efficient marketing facility.⁶⁴

61. Brodley, *supra* note 58, at 1530-31; Jacquemin & Slade, *supra* note 39, at 438-39; Pitofsky, *supra* note 47, at 1030. Econometric analyses of a large sample of U.S. joint ventures in a number of industries further suggest that where the participants are horizontal competitors, a potential for market-power augmentation exists. Sanford V. Berg & Philip Friedman, *Impacts of Domestic Joint Ventures on Industrial Rates of Return: A Pooled Cross-Section Analysis, 1964-1975*, 63 REV. ECON. & STAT. 293, 295 (1981); Jerome L. Duncan, Jr., *Impacts of New Entry and Horizontal Joint Ventures on Industrial Rates of Return*, 64 REV. ECON. & STAT. 339 (1982).

62. Brodley, *supra* note 58, at 1530-31; Reynolds & Snapp, *supra* note 58, at 148-49; see also Richard N. Clarke, *Collusion and the Incentives for Information Sharing*, 14 BELL J. ECON. 383, 384 (1983) (pooling of information "makes cheating more difficult and collusive quantity restriction more effective by improving the accuracy of every firm's market estimates"); cf. Walter Adams & James W. Brock, *The "New Learning" and the Euthanasia of Antitrust*, 74 CAL. L. REV. 1515, 1527-37 (1986) (discussing use of transnational joint ventures to solidify cartels and enforce oligopolistic collusion).

63. See, e.g., KATHRYN R. HARRIGAN, *STRATEGIES FOR JOINT VENTURES* 380 (1985); Grossman & Shapiro, *supra* note 16, at 334; David C. Mowery, *Collaborative Research and High-Temperature Superconductivity*, in *COOPERATIVE RESEARCH AND DEVELOPMENT*, *supra* note 55, at 151; MOWERY & ROSENBERG, *supra* note 28, at 241; Ordover & Baumol, *supra* note 16, at 30; Section of Antitrust Law, A.B.A., *Recommendations and Report on Production Joint Venture Legislation* 6 (Sept. 1, 1989) (unpublished manuscript, on file with the author) [hereinafter ABA Production Joint Venture Report].

64. See Brodley, *supra* note 58, at 1532; Jacquemin & Slade, *supra* note 39, at 439; Lawrence A. Sullivan, *The Viability of the Current Law on Horizontal Restraints*, 75 CAL. L. REV. 835, 868 (1987); see also 1989b House Hearing, *supra* note 4, at 199, 359, 374 (statements of Dr. T.J. Rogers, President and Chief Executive Officer of Cypress Semiconductor Corporation; Mr. D.R. Coelho, Chairman of Vantage Analysis Systems, Inc.; and Dr. L.R. Tomasetta, President and Chief Executive Officer of Vitesse Semiconductor Corporation) (detailing disadvantages of entrepreneurial firms when research consortia begin performing competing research). See generally PHILLIP E. AREEDA & HERBERT HOVENKAMP, *ANTITRUST LAW: 1990 SUPPLEMENT* ¶¶ 736.1-2 (1990) (discussing case law and applications of essential facilities doctrine).

As in the case of RJs, it is difficult to accurately predict whether a particular PJV is socially beneficial or socially costly without examining the specific characteristics of the participants, the markets involved, and the joint venture agreement itself. Nevertheless, certain generalizations can be made.

For example, it is generally recognized that anticompetitive effects are more probable where the participants are horizontal competitors,⁶⁵ although such effects are not limited to such ventures. Anticompetitive effects are also more likely where the relevant market is concentrated and exhibits entry barriers and where the participants collectively account for a significant portion of the market.⁶⁶ This suggests that production consortia involving a majority of the firms in an industry pose a special antitrust risk.

Therefore, the nature and structure of the joint venture and possible collateral restraints in the joint venture agreement can affect the likelihood that it will impose a net social cost. For example, a distribution or marketing JV is more likely to have anticompetitive effects than a production JV, since it prevents the participants from competing in marketing their products.⁶⁷ In addition, collateral restraints in the joint venture agreement may limit competition among the participants. For example, the joint venture agreement may contain field of use or geographic restrictions in intellectual property licenses.⁶⁸ Alternatively, the agreement may simultaneously prohibit the participants from independently manufacturing products that compete with those produced by the venture while limiting the amount of the venture's product that is distributed to each participant.⁶⁹ Also, collateral restraints that restrict distribution of the venture's product to the participants may disadvantage nonparticipants.⁷⁰ Thus, while PJs offer smaller potential

65. See Brodley, *supra* note 58, at 1552; Pitofsky, *supra* note 47, at 1031; cf. Fusfeld, *supra* note 59; Mead, *supra* note 59 (discussing possible anticompetitive effects of horizontal joint ventures in the iron and steel industry and in the bidding for oil and gas leases).

66. See International Guidelines, *supra* note 44, at 20,600; Brodley, *supra* note 58, at 1541-42.

67. See, e.g., 1990 Senate Hearing, *supra* note 4, at 63 (letter from James F. Rill, Assistant Attorney General, Antitrust Div., to Sen. Metzenbaum); 1989b House Hearing, *supra* note 4, at 129 (statement of Edward Rock); Brodley, *supra* note 58, at 1555-56; ABA Production Joint Venture Report, *supra* note 63, at 34. But see *supra* note 58.

68. See Grossman & Shapiro, *supra* note 16, at 329; Stockdale, *supra* note 21, at 80.

69. Cf. Brodley, *supra* note 58, at 1560-61.

If the parent must procure the input from the joint venture, regulation of the joint venture's output effectively controls the output of the parents. Moreover, in establishing the production level of the joint venture, the parents necessarily reveal their own output plans and thus diminish the uncertainty necessary for effective competition

Id.

70. See *id.* at 1563-65; ABA Production Joint Venture Report, *supra* note 63, at 14.

social benefits than RJs, they pose significantly higher social costs. This especially holds for production consortia consisting of a significant number of horizontal competitors.

D. Organizational Difficulties Associated with All Joint Ventures

In evaluating public policies that may encourage joint ventures, one must also consider the organizational difficulties and transaction costs associated with this form of business organization. These organizational difficulties will influence not only the types of joint ventures formed, but also the likely balance of social benefits and costs that will result. In addition, these difficulties will likely limit both the number of PJs formed and their likely success.

In attempting to overcome these organizational difficulties, joint venture participants will frequently attempt to limit competition among themselves. While this should not impose social costs for ventures involving small numbers of firms that collectively lack market power, it can pose dangers for joint ventures involving a large number of firms, especially where the firms are cooperating in production. In such cases, the major purpose of the venture may be to eliminate competition rather than to achieve efficiencies.

It is widely recognized that the presence of multiple participants makes management of joint ventures extremely difficult,⁷¹ and these difficulties tend to increase as the number of participants increases.⁷² As a result, decisionmaking tends to become slower and more cumbersome than in other forms of organization.⁷³

71. Joint venture participants often disagree on such fundamental matters as the goals of the venture, likely developments in technology or the market, and the relative contributions of the parents. *See, e.g.*, KILLING, *supra* note 47, at 8; PORTER, *supra* note 46, at 66; MOWERY & ROSENBERG, *supra* note 28, at 247; Michael E. Porter & Mark B. Fuller, *C coalitions and Global Strategy in COMPETITION IN GLOBAL INDUSTRIES* 326 (Michael E. Porter ed., 1986).

72. In the case of research consortia involving large numbers of participating firms, the Department of Commerce has estimated that one year is "the minimum time required to reach agreement on the research agenda and other management issues." Link & Tassey, *supra* note 55, at xix; *see also* 1989a House Hearing, *supra* note 4, at 74 (statement of Claude E. Barfield, American Enterprise Institute) ("Organizational difficulties will be [sic] tend to vary inversely with the number of firms involve [sic]: the more firms involved, the more illusive [sic] a consensus on agenda, increased potential for conflict in business cultures, and increased likelihood that the purpose of [the] venture will be defeated."); George R. Heaton, Jr., *The Truth About Japan's Cooperative R&D*, 5 ISSUES SCI. & TECH. 32, 37 (1988) (among Japanese RJs, "[a]s the membership increases, the difficulties in agreeing on a technical agenda rise proportionately; the larger the group, the less ambitious and more basic the research aims tend to be").

73. Based on a study of 37 joint ventures, J. Peter Killing found that management problems occurred not only at the board level of the parent firms but also at the management level of the joint venture itself. This latter problem occurred because the

That the venture itself can pose a competitive threat to its parents further complicates the management of the venture. The venture may itself become a competitor to one or more of the parents or may increase the competitive strength of one parent relative to the others.⁷⁴ More importantly, although the joint venture may depend on technological transfer, the participants are frequently reluctant to share strategic technological information.⁷⁵ In addition, firms often attempt to free-ride by contributing their less able personnel or by withholding their most advanced technology.⁷⁶ In other cases, participants will vigorously attempt to prevent the disclosure of proprietary information to their partners.⁷⁷

These problems in turn influence the structure of joint ventures. For example, in order to minimize competitive threats to participants, research consortia have tended to focus on basic research, pre-competitive research or non-competitive research and have eschewed applied

management staff of the joint venture tended to be drawn from the various parent organizations, and the working relationship among managers from different parents tended to be strained and inefficient. KILLING, *supra* note 47, at 9-10. A subsequent study of over 400 joint ventures found that decisionmaking was more cumbersome in joint ventures compared with a wholly-owned subsidiary and that it was "more difficult to get something done quickly." HARRIGAN, *supra* note 63, at 373. It further found that joint ventures having a 50-50 ownership split were disfavored by some parent managers, because this further slowed and complicated decisionmaking. *Id.* at 368.

74. See PORTER, *supra* note 46, at 66; Porter & Fuller, *supra* note 71, at 326. In some cases, a product developed independently by one of the participants may compete with jointly developed products, leading to the demise of the venture. See MOWERY & ROSENBERG, *supra* note 28, at 247; see also Mariti & Smiley, *supra* note 50, at 446 (giving examples of joint venture participants that were injured by the joint venture itself or by their partners).

75. See, e.g., HARRIGAN, *supra* note 63, at 344-47; Stockdale, *supra* note 21, at 252-53.

76. See, e.g., MOWERY & ROSENBERG, *supra* note 28, at 225; Shapiro & Willig, *supra* note 47, at 114; Stockdale, *supra* note 21, at 252. Microelectronics and Computer Technology Corporation (MCC) presents a clear example of this problem. Initially, MCC was designed to be operated with a staff drawn from its member companies. The members, however, sent their less able researchers. After MCC rejected 90% of the researchers sent by the member firms, it staffed its laboratories primarily with outside personnel. See HARRIGAN, *supra* note 63, at 231; MOWERY & ROSENBERG, *supra* note 28, at 270; cf. ALBERT N. LINK & LAURA L. BAUER, COOPERATIVE RESEARCH IN U.S. MANUFACTURING: ASSESSING INITIATIVES AND CORPORATE STRATEGIES 95 (1989) (discussing complaints of chemical firm researchers participating in RJsVs, who believed that none of the participants were sending their best scientists). Moreover, these same problems appear to have plagued the much-touted Japanese research joint ventures. See *id.* at 225; PORTER, *supra* note 46, at 635.

77. Frequently, participants will insist on confidentiality agreements that prevent a joint venture from disclosing information concerning one participant to another. See HARRIGAN, *supra* note 63, at 344-45; Stockdale, *supra* note 21, at 251. In other cases, such as the joint development of the International Aero Engines V2500 jet engine, the partners will each be assigned separate development of particular components in order to minimize technology transfer, even if this causes inefficiency. See DAVID C. MOWERY, ALLIANCE POLITICS AND ECONOMICS: MULTINATIONAL JOINT VENTURES IN COMMERCIAL AIRCRAFT 94-95 (1987).

research and development activities.⁷⁸ Firms appear to view applied RJs as a second or third best alternative and participate only when they cannot accomplish a task on their own.⁷⁹ Even then, they will seek to limit the number of partners to those absolutely necessary to accomplish the objective—usually two or three.⁸⁰ The predominance of two- and three-firm joint ventures appears to reflect both an attempt to minimize coordination problems and to reduce the likelihood that a partner would lose any competitive advantage to other partners.

As previously indicated, where two or three firms participate in a joint venture, they frequently will seek to protect their strategic technological knowledge and capabilities from their partners.⁸¹ In addition, participants often will try to minimize internal competition that can dissipate joint profits. Thus, they may extend cooperation from R&D through production and even distribution and marketing.⁸² Participants also frequently choose as partners firms that are not direct competitors,

78. See HAKLISCH ET AL., *supra* note 32, at 2; Link & Tassey, *supra* note 55, at viii, xii. Despite this focus, many firms have refused to participate, either because they believe they will be able to gain access to the research results without participating or because they fear disclosing sensitive proprietary information. WOLEK, *supra* note 40, at 130-33, 151-59; Johnson, *supra* note 24, at 80-81. And those who do join appear willing to contribute only modest amounts to the cooperative effort. See Stockdale, *supra* note 21, at 252; cf. HAKLISCH ET AL., *supra* note 32, at 16 (cooperative research at research consortia accounted for only 1.2% of total national R&D expenditures by industry in 1982).

79. Based on a questionnaire survey and interviews of firms participating in joint ventures, Professor Berg and colleagues concluded:

The ranking of alternatives to JVs was similar across firms. A wholly controlled internal project was deemed most preferable, everything else being equal. Where feasible, a merger ran second as a way to enter new markets. Since joint ventures provide an equity position, they were preferred to licensing by some firms; others asserted that the coordination problems of a joint venture render that interfirm linkage undesirable.

BERG ET AL., *supra* note 55, at 45 (1982); cf. HARRIGAN, *supra* note 63, at 56-57 (based on a study of over 400 joint ventures, the author concluded that the "most likely candidates for joint ventures are firms that lack the capabilities, strengths, or resources needed to exploit business opportunities alone," and that joint ventures "will not occur unless firms need to diversify, acquire new skills and resources, consolidate their positions, or attain objectives that they cannot reach alone"); PORTER, *supra* note 46, at 67 ("Alliances [of which joint ventures are a type] appear to be most common among second-tier competitors or companies trying to catch up.").

80. See *supra* note 55; cf. LINK & BAUER, *supra* note 76, at 28 (survey of early filings under NCRA found that degree of appropriability and average number of joint venture participants are inversely related).

There are three major exceptions to this generalization: Bell Communications Research, Inc. (Bellcore), the Electric Power Research Institute, and the Gas Research Institute. These RJs conduct not only significant basic research but also considerable applied research. It is plausible that the participants agreed to cooperate in applied research in these cases, because they are all subject to regulation and hence do not view themselves as competitors. See HAKLISCH ET AL., *supra* note 32, at 200.

81. See *supra* note 77.

82. Katz & Ordover, *supra* note 16, at 156; Stockdale, *supra* note 21, at 80.

such as firms that operate in related industries, that focus on different market niches, or that are located in different geographic markets.⁸³ Finally, as noted above, participants, in order to limit competition, may include collateral restraints in the joint venture agreement, such as field of use or geographic restrictions.⁸⁴

Such attempts to limit dissipation of profits should not pose significant anticompetitive risks where the market is competitive and the participants collectively lack market power. Significant antitrust concerns can arise, however, where the participants individually possess market power or where, as in the case of production consortia, a significant proportion of firms in the industry cooperate. These concerns are heightened by the significant possibility that, in such cases, one of the main purposes of the joint venture may be to facilitate or enforce collusion.

Thus, the organizational difficulties associated with joint ventures not only reduce their potential for achieving significant economies, but also frequently induce the participants to limit competition among themselves. This, in turn, raises the possibility of significant anti-competitive effects when the joint venture participants individually or collectively possess market power.

In summary, while PJVs offer much more limited benefits than RJs, they also create significantly greater dangers to competition. Therefore, serious doubt exists as to whether PJVs should receive the same favorable treatment accorded RJs.

IV. CURRENT LAW AND ENFORCEMENT POLICY CONCERNING PRODUCTION JOINT VENTURES AND THEIR EFFECT ON SUCH VENTURES

Proponents of the proposed PJV legislation contend that uncertainty over the legality of PJVs and the possibility that courts will condemn a JV as per se illegal deters potential PJV formation.⁸⁵ This concern appears exaggerated.⁸⁶ Current antitrust law and enforcement policy clearly

83. *Id.* at 253; cf. William G. Ouchi & Michele K. Bolton, *The Logic of Joint Research and Development*, 30 CAL. MGMT. REV. 9, 27 (1988) ("Most of the inter-corporate R&D which has occurred consists of contractual joint development of a new, applied technology by two companies at different stages in the vertical stream of an industry.").

84. See *supra* text accompanying notes 56-57.

85. See, e.g., 1990 Senate Hearing, *supra* note 4, at 35 (statement of Robert A. Mosbacher, Secretary of Commerce); 1989b House Hearing, *supra* note 4, at 57, 185-87 (statements of Rep. Thomas Campbell & Gordon E. Moore); Jorde & Teece (1989a), *supra* note 4, at 38-41.

86. In contrast to PJVs today, evidence exists that, in the early 1980s, uncertainty concerning the legality of *research* joint ventures may have deterred their formation and justified the special treatment accorded them under the NCRA. There were several factors that contributed to this uncertainty on the part of potential RJV participants. First, in one of the few decided cases involving RJs, the Government had successfully challenged a

indicate that legitimate PJVs will be judged under the rule of reason and that procompetitive PJVs will not be condemned. In addition, substantive and procedural changes in the law have reduced the threat of private antitrust challenges. Finally, the large numbers of PJVs that have been formed in recent years belies the need for any special treatment.⁸⁷

A. Courts Judge PJVs Under the Rule of Reason

Recent Supreme Court decisions clearly indicate that bona fide joint ventures—*i.e.*, joint ventures that are not merely shams to cover anticompetitive collusion—will be evaluated under the rule of reason.⁸⁸

RJV, made up of the three major automobile manufacturers, which had sought to develop automobile pollution reduction technology. *See United States v. Automobile Mfrs. Ass'n*, 1969 Trade Cas. (CCH) ¶ 72,907 (C.D. Cal. 1969); *see also SULLIVAN, supra* note 29, at 301-03. Second, in several business review proceedings, the Antitrust Division either indicated an intention to challenge proposed RJVs should the participants pursue their plans, *see ANTITRUST DIV., U.S. DEP'T OF JUSTICE, ANTITRUST GUIDE CONCERNING RESEARCH JOINT VENTURES* app. B., at 5, 9 (1980) [hereinafter ANTITRUST RJV GUIDE], made burdensome requests for information and unreasonably delayed in giving consent, *see Stockdale, supra* note 21, at 201-02, 228, 254; ABA Production Joint Venture Report, *supra* note 67, at 22 n.4, or refused to provide a definite answer, ANTITRUST RJV GUIDE, *supra*, app. B, at 13-15. Third, the Antitrust Division's 1980 *Antitrust Guide Concerning Research Joint Ventures*, while ostensibly intended to reduce legal uncertainty and encourage cooperative research, may have had the opposite effect. For example, the *Antitrust Guide's* general opposition to industry-wide research consortia, *id.* at 11, its statement that certain collateral restraints are per se illegal, *id.* at 14-15, and its suggestion that denying competitors access to a RJV, either *ex ante* or *ex post*, might constitute a per se violation of section 1 of the Sherman Act, *id.* at 22, could easily have discouraged firms from participating in RJVs. *See Stockdale, supra* note 21, at 114-22; *see also 1990 Senate Hearing, supra* note 4, at 93 (statement of Joseph Brodley). Finally, the courts had yet to issue certain decisions clarifying the appropriate method for analyzing joint ventures. *See infra* Section IV.B.

Further support for this view is found in the fact that 145 RJVs were notified in the first five years after passage of the NCRA, while only 21 RJVs were formed in the three years prior to the NCRA. *See Brodley, supra* note 6, at 100.

87. *See infra* Section IV.D.

88. *See Northwest Wholesale Stationers, Inc. v. Pacific Stationery & Printing Co.*, 472 U.S. 284 (1985); *National Collegiate Athletic Ass'n v. Board of Regents of Univ. of Okla.*, 468 U.S. 85 (1984); *Broadcast Music, Inc. v. Columbia Broadcasting Sys.*, 441 U.S. 1 (1979).

Even before these decisions were rendered, the Supreme Court had generally applied the rule of reason in evaluating joint ventures. *See, e.g.*, *United States v. Penn-Olin Chem. Co.*, 378 U.S. 158 (1964); *see also E. THOMAS SULLIVAN & JEFFREY L. HARRIS, UNDERSTANDING ANTITRUST AND ITS ECONOMIC IMPLICATIONS* 102 (1988) (Supreme Court sanctioned "broad rule of reason . . . for joint ventures" in *Chicago Bd. of Trade v. United States*, 246 U.S. 231 (1918)); *Brodley, supra* note 58, at 1534-35 ("Although joint ventures may be challenged under each of the three major antitrust statutes . . . [t]he guiding legal principle is the Rule of Reason, except in cases of flagrant cartel practices . . .").

Nevertheless, in some earlier cases, the Supreme Court held ancillary restraints to joint marketing efforts to be per se illegal. *See United States v. Topco Assoc., Inc.*, 405 U.S. 596 (1972); *United States v. Sealy, Inc.*, 388 U.S. 350 (1967); *Timken Roller Bearing Co. v. United States*, 341 U.S. 593 (1951). Although no decision has explicitly overruled these cases, a number of scholars have suggested that subsequent Supreme Court decisions have implicitly overruled them. *See, e.g.*, *Rothery Storage & Van Co. v. Atlas Van Lines*,

In *Broadcast Music, Inc. v. Columbia Broadcasting System, Inc. (BMI)*,⁸⁹ the Supreme Court reviewed a court of appeals decision holding that the blanket licenses issued by defendants BMI and the American Society of Composers, Authors and Publishers (ASCAP) to the television networks constituted a form of price fixing that was illegal per se under the Sherman Act. Rejecting the Court of Appeals' "literal" approach to price fixing, the Court stated:

Literalness is overly simplistic and often overbroad. When two partners set the price of their goods or services they are literally "price fixing," but they are not per se in violation of the Sherman Act.⁹⁰

Emphasizing that agreements that may increase economic efficiency or competition should not be held per se illegal, the Court held that the blanket license should be evaluated under the rule of reason, because the license "is not a 'naked restrain[t] of trade with no purpose except stifling of competition,' . . . but rather accompanies the integration of sales, monitoring, and enforcement against unauthorized copyright use."⁹¹ The Court further noted that, absent a blanket license, copyright holders might find it too expensive to enter into individual sales contracts or individually to monitor and enforce their copyrights.⁹²

The Supreme Court's decision in *National Collegiate Athletic Association v. Board of Regents of University of Oklahoma* (NCAA)⁹³ reaffirmed the analytical approach adopted in *BMI*. In *NCAA*, certain colleges challenged the NCAA's policy which limited the total number, and number-per-college, of televised intercollegiate football games and which prohibited any member college from selling television rights independently. Although it found that the plan was a horizontal restraint involving both prices and output,⁹⁴ the Supreme Court nevertheless rejected the per se approach adopted by the court of appeals. Instead, it applied a rule of reason analysis, because the "case involve[d] an industry in which horizontal restraints on competition are essential if the product is to be available at all."⁹⁵ The Court also recognized that intercollegiate sports required a "myriad of rules" to function and that the NCAA

792 F.2d 210, 226-27 (D.C. Cir. 1986) (Bork, J.), cert. denied, 479 U.S. 1033 (1987); Martin B. Louis, *Restraints Ancillary to Joint Venture and Licensing Agreements: Do Sealy and Topco Logically Survive Sylvania and Broadcast Music?*, 66 VA. L. REV. 879, 880 (1980).

89. 441 U.S. 1 (1979).

90. *Id.* at 9.

91. 441 U.S. at 20 (quoting *White Motor Co. v. United States*, 372 U.S. 253, 263 (1963)).

92. *Id.* at 10.

93. 468 U.S. 85 (1984).

94. *Id.* at 97.

95. *Id.* at 101.

played a vital role in enforcing the rules and preserving the character of the game.⁹⁶

Finally, in *Northwest Wholesale Stationers, Inc. v. Pacific Stationery & Printing Co.*⁹⁷, the Supreme Court refused to apply a per se rule to a concerted refusal to deal by a wholesale purchasing cooperative. Although it stated that a per se rule was appropriate where a concerted refusal to deal involved a "joint effort . . . to disadvantage competitors" and was unlikely to "enhance overall efficiency and make markets more competitive,"⁹⁸ the Court, following *BMI* and *NCAA*, nevertheless held that a rule of reason approach was warranted in the case of wholesale purchasing cooperatives, because such cooperatives "must establish and enforce reasonable rules in order to function effectively."⁹⁹

As expected, subsequent lower court decisions have followed the Supreme Court's rule of reason approach to joint ventures. For example, in *Rothery Storage Van Co. v. Atlas Van Lines, Inc.*,¹⁰⁰ the D.C. Circuit, in a decision by Judge Bork, refused to apply a per se rule to an alleged "group boycott" of the plaintiff by the defendants, Atlas Van Lines and several affiliated carrier agents. Instead, the court, applying the rule of reason, found that the challenged restraints were ancillary to the joint venture, that they "preserved the efficiencies of the nationwide van line by eliminating the problem of the free ride, and accordingly, that Atlas' decision to terminate plaintiff's agency contract did not violate the Sherman Act."¹⁰¹ Similarly, in *Polk Brothers, Inc. v. Forest City Enterprises, Inc.*,¹⁰² the Court of Appeals for the Seventh Circuit reviewed a district court's decision that a noncompetition agreement was per se illegal. Finding that the covenant played an essential role in inducing the firms to cooperate, the Court held that the restraint was ancillary and that it therefore should be evaluated under the rule of reason.¹⁰³

These cases clearly indicate that joint ventures, and collateral restraints in joint venture agreements, will be evaluated under the rule of reason if they have the potential for creating new products, increasing efficiency or promoting competition in the market. Only where a joint

96. *Id.* at 101-02.

97. 472 U.S. 284 (1985).

98. *Id.* at 294.

99. *Id.* at 296.

100. 792 F.2d 210 (D.C. Cir. 1986), *cert. denied*, 479 U.S. 1033 (1987).

101. *Id.* at 229.

102. 776 F.2d 185 (7th Cir. 1985).

103. *Id.* at 188-91; see also *National Bancard Corp. v. VISA U.S.A.*, 779 F.2d 592 (11th Cir. 1986) (interchange fee charged by VISA not a naked restraint of competition and therefore not per se price fixing), *cert. denied*, 479 U.S. 923 (1986); *Berkey Photo, Inc. v. Eastman Kodak Co.*, 603 F.2d 263, 299-302 (2d Cir. 1979) (joint R&D between monopolist in one market and major firm in complementary market not a per se violation of section 1), *cert. denied*, 444 U.S. 1093 (1980).

venture is a sham "with no purpose except stifling competition,"¹⁰⁴ will it be subject to a per se rule.

B. Current Antitrust Enforcement Policy Is Hospitable Towards Legitimate PJVs

In their published guidelines and enforcement actions, the Antitrust Division of the Department of Justice and Federal Trade Commission have likewise adopted hospitable joint venture policies based on the rule of reason.

For Example, in its *Antitrust Enforcement Guidelines for International Operations*,¹⁰⁵ the Antitrust Division emphasizes that it will apply a rule of reason analysis in evaluating joint ventures that involve "some form of economic integration that goes beyond the mere coordination of the parties' decisions on price or output and that in general may generate procompetitive efficiencies."¹⁰⁶ The Guidelines state that the Division will first consider whether the joint venture is likely to have any anticompetitive effects in the market in which it operates¹⁰⁷ or in any "spillover markets."¹⁰⁸ The Division performs a similar rule-of-reason analysis for any vertical non-price restraints associated with the joint venture to determine if the restraints could facilitate collusion or exclude competitors.¹⁰⁹ If the Antitrust Division concludes that anticompetitive

104. Broadcast Music, Inc. v. Columbia Broadcasting Sys., 441 U.S. 1, 20 (1979) (quoting White Motor Co. v. United States, 372 U.S. 253, 263 (1963)).

105. International Guidelines, *supra* note 44, at 20,600.

106. *Id.* In a footnote, the Antitrust Division explains that, in determining whether to apply the rule of reason, it does not consider whether the "economic integration involved in the particular transaction actually would generate efficiencies. It is enough if the form of integration involved in general generates efficiencies." *Id.* at 20,594 n.47. However, if a purported joint venture involves no economic integration, but rather is "simply a device to restrict output or raise price," then it will not hesitate to challenge it. *Id.* at 20,600.

107. If, under its Merger Guidelines, the Division would not challenge the merger of the joint venture participants, then it will conclude that the joint venture and any associated restraints are unlikely to have any anticompetitive effects in the joint venture market. *Id.* Moreover, even if a merger of the participants would raise concern, the Division recognizes that a joint venture "may have a less restrictive effect on the independent decision-making of the joint venture participants with respect to output and price than would an outright merger," and accordingly may treat the joint venture more leniently. *Id.* at 20,601.

108. The Antitrust Division acknowledges that a "joint venture may . . . include operational or procedural safeguards that substantially eliminate any risk of anticompetitive spill-over effects," *id.*, and states that the presence of such safeguards may render an elaborate structural analysis of the spill-over market(s) unnecessary, *id.* at 20,602. In the absence of such safeguards, the Division will perform a market-power analysis using the same factors it uses in merger analysis. *Id.*

109. *Id.* The Division emphasizes that selectivity in choosing partners may be important to the success of a joint venture and that, accordingly, it will be concerned with the exclusion of rivals only if "(i) an excluded firm cannot compete in a related market or markets . . . without having access to the joint venture and (ii) there is no reasonable basis

effects are unlikely, it will not challenge the venture regardless of whether it generates any efficiencies. If, however, the Division concludes that significant anticompetitive effects are likely, it then considers whether "those anticompetitive effects are outweighed by the procompetitive efficiency benefits" generated by the joint venture.¹¹⁰

The enforcement actions of the Antitrust Division and Federal Trade Commission have been consistent with the 1988 Guidelines. In recent years, for example, the enforcement agencies have allowed firms with significant market shares to enter into PJVs where those ventures involved genuine economic integration.¹¹¹ More importantly, between 1984 and 1990, the Antitrust Division challenged only three PJVs, none of which involved joint R&D,¹¹² while the Federal Trade Commission challenged only four PJVs.¹¹³ In those cases, moreover, the challenged joint ventures involved marketing collaboration along with severe market concentration.¹¹⁴

Thus, existing policies of the antitrust enforcement agencies impose no unreasonable obstacle to the formation of PJVs involving genuine economic integration.

C. The Threat of Private Antitrust Suits Is Exaggerated

Proponents of the proposed legislation also argue that the threat of private suits may deter procompetitive PJVs. This concern appears exaggerated for two reasons.

First, the economic incentives for private plaintiffs to bring an antitrust challenge have decreased. The courts' use of the rule of reason in evaluating joint ventures and willingness to consider the potential benefits as well as possible costs of the venture have raised the expected costs of bringing an antitrust challenge against a joint venture and reduced the likelihood that a plaintiff will prevail in challenging a joint

related to the efficient operation of the joint venture for excluding other firms." *Id.* (footnote omitted).

110. *Id.* The Division notes, however, that it "will not recognize claimed efficiencies if it is clear that equivalent efficiencies can be achieved by means that involve no anticompetitive effect." *Id.*

111. See, e.g., General Motors Corp., 103 F.T.C. 374 (1984) (General Motors and Toyota, the world's first and third largest automobile manufacturers, allowed to enter into PJV, partially because it would permit the diffusion of existing production techniques and know-how from Toyota to G.M., despite producing an existing Toyota model rather than a new product).

112. 1990 Senate Hearing, *supra* note 4, at 39 (letter from Bruce C. Navarro, Acting Assistant Attorney General).

113. *Id.* at 229-30 (letter from Janet D. Steiger, Chairman, FTC).

114. See, e.g., United States v. Ivaco, 704 F. Supp. 1409 (W.D. Mich. 1989) (enjoined joint venture which would combine operations of two of three remaining producers of automatic tampers to create firm holding 70% of the relevant market). See generally H.R. REP. NO. 516, *supra* note 5, at 8; Brodley, *supra* note 6, at 101.

venture. This should reduce the number of plaintiffs willing to sue, especially where the suit is primarily intended to harass or extort a settlement from the defendants.¹¹⁵

Second, in recent years, the Supreme Court and lower federal courts have developed various procedural barriers which make it more difficult for plaintiffs to maintain private antitrust actions. For example, elaboration of the concepts of antitrust standing and antitrust injury has limited the number and types of parties permitted to bring antitrust challenges.¹¹⁶ Similarly, the Supreme Court's decision in *Illinois Brick Co. v. Illinois*¹¹⁷ made it much more difficult for indirect purchasers to bring a private antitrust action. Finally, the Georgetown antitrust project¹¹⁸ found that antitrust defendants frequently succeeded in bringing pretrial motions for summary judgment and motions to dismiss against private antitrust plaintiffs.¹¹⁹

At least in part as a result of these legal developments, the number of private antitrust actions filed in federal court has declined over the past 15 years. According to the Administrative Office of the United States

115. See, e.g., HOUSE COMM. ON THE JUDICIARY, 98TH CONG., 2D SESS., STUDY OF ANTITRUST TREBLE DAMAGE REMEDY 31 (Comm. Print 1984) (prepared by G. Garvey) (reduction in number of private suits filed in recent years may be due to new judicial commitment to rule of reason); Steven C. Salop & Lawrence J. White, *Economic Analysis of Private Antitrust Litigation*, 74 GEO. L.J. 1001, 1019 (1986) ("A plaintiff is more likely to sue when his perceived probability of success is greater, [and] when his litigation costs are lower . . .").

116. See, e.g., *Cargill, Inc. v. Monfort of Colo., Inc.*, 479 U.S. 104 (1986); *Brunswick Corp. v. Pueblo Bowl-O-Mat, Inc.*, 429 U.S. 477 (1977), *cert. denied*, 429 U.S. 1090 (1977).

117. 431 U.S. 720 (1977) (holding that, in general, indirect purchasers are barred from recovering for overcharges allegedly passed down to the plaintiff purchaser through a chain of distribution).

118. The Georgetown antitrust project collected data on all private antitrust actions filed between 1973 and 1983 in five federal districts. Of those, usable data was obtained on 2357 cases. See Steven C. Salop & Lawrence J. White, *Private Antitrust Litigation: An Introduction and Framework*, in *PRIVATE ANTITRUST LITIGATION: NEW EVIDENCE, NEW LEARNING* 3-4 (Lawrence J. White ed., 1988).

119. See Stephen Calkins, *Equilibrating Tendencies in the Antitrust System, with Special Attention to Summary Judgment and to Motions to Dismiss*, in *PRIVATE ANTITRUST LITIGATION*, *supra* note 118, at 185, 200, 207; Stephen Calkins, *Summary Judgment, Motions to Dismiss, and Other Examples of Equilibrating Tendencies in the Antitrust System*, 74 GEO. L.J. 1065, 1127 (1986) [hereinafter Calkins (1986)].

In addition, the success of such motions may well increase following the Supreme Court's decision in *Matsushita Electric Industrial Co. v. Zenith Radio Corp.*, 475 U.S. 574 (1986). See, e.g., *International Distrib. Ctrs., Inc. v. Walsh Trucking Co.*, 812 F.2d 786 (2d Cir.), *cert. denied*, 482 U.S. 915 (1987) (summary judgment for defendants in section 2 monopolization case); *In re Apollo Air Passenger Computer Reservation Sys.*, 720 F. Supp. 1068 (S.D.N.Y. 1989) (summary judgment for defendant in antitrust challenge to computer reservation system); *Florida Fuels v. Belcher Oil*, 717 F. Supp. 1528 (S.D. Fla. 1989) (summary judgment for defendants in section 2 essential facilities claim). See generally Calkins (1986), *supra*, at 1127; John T. Soma & Andrew P. McCallin, *Summary Judgment and Discovery Strategies in Antitrust and RICO Actions after Matsushita v. Zenith*, 36 ANTITRUST BULL. 325 (1991); ABA Production Joint Venture Report, *supra* note 63, at 34.

Courts, the number of private antitrust suits filed per year in the federal courts peaked at 1611 in 1977. This represented 1.2% of all civil cases filed that year. By 1980 the number of private actions filed had dropped to 1457, or 0.8%, of all civil cases filed. In 1990, only 452 private antitrust actions were filed, which represented only 0.2% of all civil actions filed in federal courts.¹²⁰ The Georgetown antitrust study further indicates that of the 2357 private antitrust suits studied, only 5.8% were challenges to mergers or joint ventures.¹²¹

The changes in the law and the sharp decrease in the number of private actions together strongly suggest that: (1) joint venture participants are unlikely to be sued by private antitrust plaintiffs, and (2) if these participants are sued, they will have considerably greater protection against frivolous or harassment-motivated claims than they had in the past through the use of pre-trial motions.

D. Current Antitrust Law Has Not Deterred the Formation of an Increasing Number of PJs

The increasing number of domestic and international joint ventures that have been formed in recent years further suggests that the antitrust laws pose no obstacle to legitimate joint ventures. Although data on recent joint venture activity in the United States is inadequate for a comprehensive conclusion, the empirical studies that have been conducted all agree that the number of PJs formed each year since the middle to late 1970s has been both significant and growing. For example, an informal survey of joint venture announcements in the *Wall Street Journal* by the Antitrust Division's Office of Economic Policy found 130 joint venture announcements during a two and one-half year period in the late 1980s.¹²² Another, more in-depth, survey of domestic joint ventures formed between 1960 and 1984 found that joint venture activity had "blossomed" since 1978, and that in some industries, the number of joint ventures formed in 1983 alone exceeded all previously announced joint ventures in that industry.¹²³ Moreover, the growth in joint venture

120. ANNUAL REPORT OF THE DIRECTOR OF THE ADMINISTRATIVE OFFICE OF THE UNITED STATES COURTS Table C-2 (1977), (1980), & (1990).

121. Salop & White, *supra* note 118, at 6.

122. 1989a House Hearing, *supra* note 4, at 45 (statement of James Rill, Assistant Attorney General, Antitrust Division).

123. HARRIGAN, *supra* note 63, at 7. See also Link & Tassey, *supra* note 55, at vii (joint ventures involving two or three firms increased from less than 200 per year in the 1970s to more than 400 per year by the mid-1980s); Mowery, *supra* note 46, at 3 (in recent years, the number of domestic and international collaborations involving U.S. firms has increased considerably).

activity has been especially rapid in certain high-technology industries, such as semiconductors.¹²⁴

During this same period, the number of international joint ventures involving U.S. firms also increased significantly. According to a study by Hladik of international joint ventures formed between 1974 and 1982 involving at least one American firm, the number of such ventures formed during the latter half of the period roughly doubled that of the first half.¹²⁵ This increased rate of growth appears to have continued beyond 1984, the termination date of the study.¹²⁶

Moreover, many of the joint ventures have involved large companies with substantial market shares and which are direct competitors. For example, joint ventures have been formed between General Motors and Toyota, General Motors and Chrysler, Merck and Johnson & Johnson, Dow and Eli Lilly, IBM and Microsoft, and IBM and Apple.¹²⁷

This evidence of widespread joint venture activity clearly suggests that the antitrust laws do not pose a significant obstacle to the formation of PJVs. Further support for this view lies in the lack of hard evidence that antitrust concerns deterred any planned joint ventures. For example, the Antitrust Section of the American Bar Association reported that it found only one instance in which domestic firms declined to pursue an integrative PJV for reasons that would be remedied by the proposed legislation.¹²⁸

V. ARGUMENTS FAVORING THE PROPOSED SPECIAL TREATMENT OF DOWNSTREAM CONSORTIA ARE FLAWED

Proponents of antitrust reform for PJVs have advanced several supporting arguments, most of which relate to special needs of high-

124. See, e.g., Katz & Ordover, *supra* note 16, at 170 (between January 1985 and July 1989, U.S. firms formed over 140 joint ventures in the semiconductor industry); Shapiro & Willig, *supra* note 47, at 117 (there has been a "decade-long trend in the distribution of joint venture formations" from energy, chemical, and metals industries "towards computer, electronic components, communications systems, pharmaceuticals, medical equipment, and financial services industries").

125. KAREN J. HLADIK, INTERNATIONAL JOINT VENTURES: AN ECONOMIC ANALYSIS OF U.S.-FOREIGN BUSINESS PARTNERSHIPS 39 (1985).

126. MOWERY & ROSENBERG, *supra* note 28, at 243 n.5.

127. See Brodley, *supra* note 6, at 101; Shapiro & Willig, *supra* note 47, at 117; Richard Brandt et al., *IBM and Microsoft: They're Still Talking, But . . .*, Bus. Wk., Oct. 1, 1990, at 164; Deidre A. Depke & Kathy Rebello, *IBM-Apple Could Be Fearsome*, Bus. Wk., Oct. 7, 1991, at 28.

128. ABA Production Joint Venture Report, *supra* note 63, at 20; cf. MOWERY & ROSENBERG, *supra* note 28, at 253 ("There is little evidence to support the argument that U.S. antitrust policy is a central factor in the decisions of American firms to collaborate with foreign [rather than with domestic] enterprises.").

technology industries. This Part reviews these arguments and shows that they either do not justify the broad proposed antitrust relief or that they are of questionable empirical importance or validity.

A. The High Costs and Risks of Commercializing and Producing Innovative Products Do Not Exceed the Capacity of Most Individual Firms

A frequently cited justification for encouraging PJVs is that the costs and risks of developing and manufacturing a new product have increased beyond the resources of many individual firms.¹²⁹ Citing such examples as dynamic random access memory chips (DRAMs), high-definition television (HDTV), and high-temperature superconductors,¹³⁰ proponents argue that not only have costs of R&D risen, but so too have the costs of plants for manufacturing any products of R&D.

Accepting the validity of these cost increases in certain high-technology industries, it does not follow that we should encourage industry-wide production consortia. First, to the extent that high basic or fundamental applied research costs deter firms from developing new products, such research could be conducted in a research consortia, such as Sematech or Microelectronics and Computer Technology Corporation, with the results then transmitted to the member companies. Extending cooperation into development and production is not a necessary requirement.

Second, with respect to the cost of plants and equipment necessary to produce new products, empirical studies suggest that, in the vast majority of industries, the minimum efficient scale of a plant is small relative to market demand.¹³¹ This suggests that production consortia are seldom necessary to achieve efficient-scale plants, and that PJVs involving two or three firms would solve the problem. This appears true even in the industries cited as requiring industry-wide production consortia. For example, thirteen Japanese companies manufacturing semiconductor memory chips in thirty separate plants¹³² rebuts the argument that a single industry-wide DRAM PJV is necessary. Similarly, with HDTV,

129. See, e.g., 1990 Senate Hearing, *supra* note 4, at 23 (statement of James F. Rill, Assistant Attorney General, Antitrust Division); 1989b House Hearing, *supra* note 4, at 55-56 (statement of Rep. Tom Campbell); H. R. REP. NO. 516, *supra* note 5, at 1; Jorde & Teece (1990), *supra* note 4, at 81.

130. See, e.g., 1989a House Hearing, *supra* note 4, at 12-13 (statement of Rep. Thomas Campbell); 1989b House Hearing, *supra* note 4, at 182-84 (statement of G. Moore, Chairman, Intel Corp.).

131. See *supra* text accompanying note 48.

132. 1990 Senate Hearing, *supra* note 4, at 78 (statement of Michael Porter).

three joint ventures involving U.S. firms currently are developing competing HDTV systems.¹³³

Thus, neither research costs nor plant economies justify further relaxation of the antitrust laws.

B. The Cyclical Nature of the Development Process and Shorter Product Lives Does Not Necessitate Cooperation in Both R&D and Production

Professors Jorde and Teece find support for cooperative production in what they call a "cyclical" view of the innovation process.¹³⁴ Rejecting the traditional view of innovation as a sequential process proceeding from basic research through applied research, product development, and finally to production, Jorde and Teece argue instead that product development involves "tight linkages and feedback mechanisms" between the various levels of activity and frequent "mid-course corrections to design, and redesign." This "cyclical" innovation process together with shorter product lives, Professors Jorde and Teece contend, necessitate close linkages between those performing the research and those actually developing the product.¹³⁵ Furthermore, since smaller firms may have to go outside to obtain necessary complementary assets, Professors Jorde and Teece conclude that, in order to achieve rapid commercialization of innovation, joint ventures must operate from the research level through at least the production level.¹³⁶

Although this argument may have some validity in the case of applied research joint ventures involving two or three firms, there is reason to doubt its validity as applied to industry-wide research consortia. First, in stressing the need for communication between manufacturing and research groups, Professors Jorde and Teece fail to distinguish among basic research, applied research, and developmental activities. While those developing a new product will have to, and do, talk to those who will manufacture the product in order to ensure that the product can be manufactured efficiently and inexpensively, there appears to be no similar need for those involved in basic or fundamental applied research to be in frequent communication with the plant floor. Thus, the

133. See Andrew Kupfer, *The U.S. Wins One in High-Tech TV*, FORTUNE, Apr. 8, 1993, at 60, 63; Lucy Reilly, *Making HDTV All-Digital Delays FCC Selection*, WASH. TECH., Apr. 9, 1992, at 6.

134. Jorde & Teece (1989a), *supra* note 4, at 14-15; see also MOWERY & ROSENBERG, *supra* note 28, at 8 (arguing that many primary sources of innovation are located downstream and operate independently of frontier scientific research); Jorde & Teece (1990), *supra* note 4, at 77 (referring to the "simultaneous model of innovation").

135. Jorde & Teece (1989a), *supra* note 4, at 15.

136. Jorde & Teece (1990), *supra* note 4, at 82-84.

need for communication with and "feedbacks" from those in production clearly varies with the kind of research or development activity.

Consortia, however, have generally shunned applied or "competitive" research or developmental activities where "feed backs" might be important,¹³⁷ because of the participants' fears that they may be forced to share strategic proprietary technology. To the extent that this competitive threat deters cooperation in "competitive" research, this casts doubt on the likelihood that firms in a consortium would agree to cooperate from basic research through production, at least in the absence of collateral restraints that would effectively limit competition among the participants. On the other hand, companies may be willing to cooperate in consortia if their primary purpose is to facilitate collusion.

Finally, even if the participants could be persuaded to cooperate in a research and production consortia, it is doubtful that this would speed the commercialization of any resulting innovations, because decision-making in joint ventures is slow and cumbersome, and becomes more so as the number of participating firms increases.¹³⁸ This suggests that any gains from combining the complementary skills and resources of the participants would be far outweighed by the more inefficient decision-making processes involved in consortia. Thus, participants interested in production consortia should be viewed with suspicion.

In summary, the cyclical nature argument lends little support to the argument for further relaxing the antitrust laws.

C. International Competition Will Not Eliminate the Danger of Collusion Among the Joint Venture Participants

Proponents of PJVs argue that such ventures will likely not have anticompetitive effects because of the presence of international competition.¹³⁹ However, a number of factors may limit the effectiveness of foreign competition in restraining domestic collusion.

In assessing *research* joint ventures, it is recognized that the relevant geographic market for research should usually be worldwide.¹⁴⁰ This results because information, the product of the research, is easily communicated. Thus, a competing foreign firm with commercially

137. See *supra* text accompanying note 78.

138. See *supra* text accompanying note 73.

139. See, e.g., Jorde & Teece (1989a), *supra* note 4, at 4; Jorde & Teece (1990), *supra* note 4, at 91; see also 1989a House Hearing, *supra* note 4, at 32 (statement of James F. Rill, Assistant Attorney General, Antitrust Division) ("[I]ncreasing globalization of markets, particularly those incorporating advanced technologies, has dramatically reduced the risk that cooperative production efforts among some competitors would result in higher prices to American consumers.").

140. See H.R. REP. NO. 1044, 98th Cong., 2d Sess. 10 (1984); International Guidelines, *supra* note 44, at 20,625; Ordover & Baumol, *supra* note 16, at 30.

valuable technology can either incorporate it in products which it exports to the United States or license the technology to U.S. firms that are not participants in the RJV.¹⁴¹ In either case, the foreign technology will place downward pressure on prices and thus act to prevent collusive behavior on the part of RJV participants. In addition, because of appropriability problems, research results may spill over to nonparticipants, further limiting the ability of participants to collude. Thus, the presence of foreign competition reduces the need for strict antitrust enforcement of RJs.

Foreign competition will not necessarily play such an effective policing role in the case of PJVs, however. First, exchange rate fluctuations could raise the price of imported goods, rendering them less competitive. If foreign firms attempt to remain competitive on price, they make themselves vulnerable to charges of dumping. Fear of sanctions restricts the ability of foreign firms to compete effectively and prevent monopoly profits by a U.S. consortia.

More importantly, foreign competitors may be subject to trade restraints, such as tariffs, quotas, or other quantitative restrictions. Such restraints can significantly limit the ability of foreign competitors to respond to and take advantage of anticompetitive behavior by domestic firms. Quotas and other quantitative restrictions, such as voluntary restraint agreements (VRAs) or orderly marketing agreements (OMAs) are especially pernicious because they prevent the foreign firms from increasing exports to meet demand should the joint venture participants attempt to restrict output and raise price.¹⁴² Further, the foreign importers have no incentive to oppose such restrictions, since the restrictions generate scarcity rents for them as well.¹⁴³ A final reason that these trade restraints present a competitive danger is that the participants in a PJV may seek such trade protection at any time after notification to the antitrust authorities.¹⁴⁴ Thus, even if there exists strong competition

141. Baxter, *supra* note 44, at 89.

142. See Diane P. Wood, *Commentary: Antitrust and International Competitiveness in the 1990s*, 58 ANTITRUST L.J. 591, 600-01 (1989); see also ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, COMPETITION AND TRADE POLICIES: THEIR INTERACTION 49 (1984) ("VERs . . . ultimately involve actions by exporting firms to limit their volume of exports, and in some instances, to raise prices. This, in turn, can promote collusive behavior among firms and weaken competitive forces in both markets.").

143. See, e.g., GARY C. HUFBAUER & HOWARD F. ROSEN, TRADE POLICY FOR TROUBLED INDUSTRIES 15 (1986); PAUL R. KRUGMAN & MAURICE OBSFELD, INTERNATIONAL ECONOMICS: THEORY AND POLICY 191 (1988).

144. The Antitrust Division recognizes that trade restrictions can limit the competitive significance of foreign competitors and considers the existence of such restrictions in performing market analysis. See, e.g., Merger Guidelines § 3.23, 49 Fed. Reg. 26,823 (Antitrust Div., Dep't of Justice 1984), reprinted in 4 Trade Reg. Rep. (CCH) ¶ 13,103 at 20,551, 20,561 (1984); International Guidelines, *supra* note 44, at 20,598. It can not, however, anticipate the effect of trade restraints that have yet to be imposed.

from foreign firms at the time of the PJV's formation, there is no assurance that such competition will remain effective.

D. Use of Joint Ventures in Japan and Europe Does Not Require that the U.S. Encourage Domestic Joint Ventures

Joint venture proponents also rely upon the alleged widespread use of joint ventures in Europe and, especially, in Japan to support their arguments.¹⁴⁵ They argue that, in order to remain competitive, the United States must follow the examples of Japan and Europe and encourage more PJVs. But these proponents fail to understand the nature of the joint ventures that have been created in those jurisdictions.

1. JAPAN

There is little, if any, cooperation in production or marketing among Japanese companies.¹⁴⁶ Rather, cooperation is generally limited to R&D.

Even in the area of cooperative research, cooperation is more limited than generally thought. For example, although the Japanese Fair Trade Commission found that fifty-five percent of 250 major Japanese firms surveyed participated in cooperative research, ninety percent of such cooperative agreements were private contractual arrangements between two companies, most of which were already affiliated.¹⁴⁷

In addition, a widespread misunderstanding exists concerning the industry-wide research consortia, sponsored by Japan's Ministry of Trade and Industry (MITI). In the 1970s, MITI began to use existing engineering research associations¹⁴⁸ to launch a few large cooperative projects in the

145. See, e.g., 1990 Senate Hearing, *supra* note 4, at 131-32 (statement of David J. Teece); 1989b House Hearings, *supra* note 4, at 278-79 (statement of J.D. Kuehler, President, IBM Corporation); Jorde & Teece (1989a), *supra* note 4, at 27-33, 55-61.

146. 1990 Senate Hearing, *supra* note 4, at 79 (statement of Michael Porter). However, Japanese companies are increasingly entering into PJVs with non-Japanese firms. Shapiro & Willig, *supra* note 47, at 122.

147. Heaton, *supra* note 72, at 34. Moreover, only six percent of these 250 companies belonged to technology research associations, and those that belonged performed only a small fraction of their research in the associations. *Id.* at 34-35.

148. The engineering research associations (ERAs), were created by MITI in the 1960s and copied after the British Research Associations. See David B. Audretsch, *Joint R&D and Industrial Policy in Japan*, in COOPERATIVE RESEARCH AND DEVELOPMENT, *supra* note 55, at 106-07 (discussing differences between Japanese and British research associations); see also JOHNSON, *supra* note 24 (discussing development and operation of British Research Associations). The ERAs' principal function was to coordinate various member research projects and facilitate the exchange and diffusion of technological information. Audretsch, *supra*, at 107; Heaton, *supra* note 72, at 33. Most of the ERAs were quite modest in scope, and, because they lacked physical premises, research was performed in the laboratories of the member companies. See MOWERY & ROSENBERG, *supra* note 28, at 223; Heaton, *supra* note 72, at 33.

electronics, computer and aircraft industries.¹⁴⁹ The primary purpose of these projects was to close the technological gap between Japanese firms and more advanced foreign firms by adapting, disseminating and using existing technological knowledge.¹⁵⁰ Given this goal, the projects focused on generic or pre-competitive research and avoided applied research and product development.¹⁵¹ Despite this focus, participants frequently attempted to free-ride by contributing second-rate personnel or equipment or by withholding proprietary technological know-how,¹⁵² and they continued to spend far greater sums on independent research efforts.¹⁵³ Evidence also exists that as Japanese firms have caught up technologically to advanced firms of other countries, they have become increasingly reluctant to participate in cooperative research projects.¹⁵⁴ Consequently, cooperative research will probably play a less significant role in the future.¹⁵⁵

In summary, the Japanese experience in cooperative research suggests that it can be an effective mechanism for diffusing technological knowledge among firms that lag in technological sophistication. However, it hardly provides a justification for industry-wide consortia in production.

149. MOWERY & ROSENBERG, *supra* note 28, at 223; Audretsch, *supra* note 148, at 110-11. Among the more famous of these projects were the Very Large Scale Integration (VLSI) project of 1976-1979, the project that attempted to develop a fourth-generation computer, *see, e.g.*, Audretsch, *supra* note 148, at 112; Merton J. Peck, *Joint R&D: The Case of Microelectronics and Computer Technology Corporation*, 15 RES. POL'Y 219, 222 (1986), the 3.75 Computer project of 1972-76, *see* Daniel I. Okimoto, *Regime Characteristics of Japanese Industrial Policy*, in JAPAN'S HIGH-TECHNOLOGY INDUSTRIES: LESSONS AND LIMITATIONS OF INDUSTRIAL POLICY 35, 54 (Hugh Patrick ed., 1987), and the Fifth Generation Computer project, *id.* at 52.

150. *Id.* at 54; MOWERY & ROSENBERG, *supra* note 28, at 223, 225; Heaton, *supra* note 72, at 33, 37. Professor Porter sees an additional purpose for these consortia. He views them as a "signaling device to indicate important areas for long-term research attention, and as a stimulus to proprietary company research." PORTER, *supra* note 46, at 398; *see also* Okimoto, *supra* note 149, at 52.

151. George C. Eads & Richard R. Nelson, *Japanese High Technology Policy: What Lessons for the United States?*, in JAPAN'S HIGH-TECHNOLOGY INDUSTRIES, *supra* note 149, at 254; Heaton, *supra* note 72, at 35-37; *see also* Okimoto, *supra* note 149, at 52 (Japanese consortia tend to focus on the "development of precommercial prototype models").

152. MOWERY & ROSENBERG, *supra* note 28, at 225; PORTER, *supra* note 46, at 398; *see also* Okimoto, *supra* note 149, at 54 ("for the first several years, mutual suspicion and concerns about the leakage of proprietary information impeded the free exchange of information" in the VLSI project).

153. *Id.* at 53.

154. *See* MOWERY & ROSENBERG, *supra* note 28, at 226.

155. *See* Heaton, *supra* note 72, at 37, 39.

2. EUROPE

Like Japan, the European Community (EC) has attempted to strengthen the international competitiveness of its high-technology industries by encouraging and subsidizing cooperative research. The majority of these subsidized cooperative research programs, however, have been limited to *precompetitive research*¹⁵⁶ and have not extended to joint production or marketing. Furthermore, with respect to private unsubsidized PJs, EC competition policy appears to present a greater obstacle to cooperation than do the U.S. antitrust laws.

The European Strategic Programme for Research and Development in Information Technology (ESPRIT), created in 1984, was the first major EC-subsidized collaborative research project. It focused on pre-competitive research in information technology and was intended to revitalize the EC's information technology and electronics industries.¹⁵⁷ Several other major, and minor, subsidized collaborative research projects followed, including RACE (Research in Advanced Communications for Europe), BRITE (Basic Research in Industrial Technologies for Europe), EURAM (European Research in Advanced Materials), and several biotechnology projects.¹⁵⁸ These various programs, which were brought together under the FRAMEWORK Programme, all focus on precompetitive research and share the same goals: raising the research capabilities of European firms in certain high-technology industries and encouraging transnational research cooperation among EC firms and between those firms and universities.¹⁵⁹

In 1985, another collaborative project, EUREKA (the European Research Cooperation Agency), was launched.¹⁶⁰ Like the FRAMEWORK Programme, EUREKA seeks to improve the competitiveness of members' high-technology firms and to foster intercompany cooperation across borders. Although EUREKA projects are not restricted to precompetitive

156. See NATIONAL SCIENCE BD., SCIENCE AND TECHNOLOGY INTEGRATION IN EUROPE AND INFLUENCES ON U.S.-EUROPEAN COOPERATION 1 (1990) [hereinafter NSB REPORT]; OFFICE OF TECHNOLOGY ASSESSMENT, U.S. CONGRESS, COMPETING ECONOMIES: AMERICA, EUROPE, AND THE PACIFIC RIM 29 (1991) [hereinafter OTA REPORT]; MARGARET SHARP & CLAIRE SHEARMAN, EUROPEAN TECHNOLOGICAL COLLABORATION 42-74 (1987); Roy Rothwell, *Public Innovation Policies: Some International Trends and Comparisons*, 12 PAPERS IN SCI. TECH. & PUB. POL'Y 13-14 (1986).

157. OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 47-53.

158. NSB REPORT, *supra* note 156, at 1; OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 56-62.

159. NSB REPORT, *supra* note 156, at 2; OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 63-65.

160. OTA REPORT, *supra* note 156, at 29. EUREKA is not an EC program, though all the EC member countries and the EC Commission itself are members. *Id.*

research, and can extend to more commercial research,¹⁶¹ they receive significantly less public funding.¹⁶²

Although the EC and its member states have contributed significant funds to these programs,¹⁶³ the programs have received some criticism.¹⁶⁴ More importantly, both subsidized programs are limited to cooperative research, and in the case of the FRAMEWORK Programme, limited to solely precompetitive research. Neither program permits cooperation to extend to coproduction or to joint distribution.

If EC firms wish to enter into private R&D ventures or to extend cooperation beyond R&D to production or marketing, they must deal with the EC competition laws. Complying with these laws often proves more burdensome and uncertain than complying with U.S. antitrust law.

In the EC, joint ventures must be primarily concerned with article 85 of the Treaty of Rome (the "Treaty"),¹⁶⁵ which basically prohibits agreements that restrict competition.¹⁶⁶ Under article 85(2), any agreement that

161. OTA REPORT, *supra* note 156, at 29; SHARP & SHEARMAN, *supra* note 156, at 71.

162. The U.S. International Trade Commission estimated that less than 10% of EUREKA's funding comes from public sources. U.S. INT'L TRADE COMM'N, 1992: THE EFFECTS OF GREATER ECONOMIC INTEGRATION WITHIN THE EUROPEAN COMMUNITY ON THE UNITED STATES: SECOND FOLLOWUP REPORT 16-10 (1990) [hereinafter ITC REPORT].

163. According to a 1990 estimate by the U.S. International Trade Commission, public funding of the FRAMEWORK program will amount to about ECU 5.7 billion (\$6.7 billion) from 1990-94. *Id.* at 16-6.

164. Numerous and varied criticisms have been leveled against these programs. First, some have criticized the way project participants are selected, suggesting both that small- and medium-sized firms do not have sufficient access, *see, e.g.*, Leigh Bruce, *EUREKA Has Found It*, INT'L MGMT., Dec. 1987, at 38, 41, and that projects are being forced to accept firms that merely seek a free ride. *See* OTA REPORT, *supra* note 156, at 29. A number of projects involving a large number of participants have also been criticized for experiencing management problems. DIRECTORATE GEN. XIII, TELECOMMUNICATIONS, INFO. INDUS. & INNOVATION, THE REVIEW OF ESPRIT, 1984-1988: THE REPORT OF THE ESPRIT REVIEW BOARD 9 (1989) [hereinafter ERB REPORT]. Others have criticized the procedures for selecting research projects. *See, e.g.*, SHARP & SHEARMAN, *supra* note 156, at 74. The research projects selected have also been criticized for being on the periphery of firms' true concerns, ERB REPORT, *supra*, at 8, and for producing little commercially useful technology. Guy de Jonquieres, *ESPRIT, JESSI Come Under Attack*, NEW TECH. WK., Nov. 5, 1990; *see also* ERB REPORT, *supra*, at 8. Finally, and most importantly, questions have been repeatedly raised whether the programs are improving the competitiveness of EC firms or making them more dependent on government subsidies. *See, e.g.*, OTA REPORT, *supra* note 156, at 29; Bruce, *supra*, at 41; Jonquieres, *supra*.

165. Article 86 of the Treaty may also prove applicable if the parties individually or collectively enjoy a dominant position before the formation of the venture. *See* 2 BARRY E. HAWK, UNITED STATES, COMMON MARKET AND INTERNATIONAL ANTITRUST: A COMPARATIVE GUIDE 308 & n.65 (2d ed. 1990); Frank L. Fine, EEC Antitrust Aspects of Production Joint Ventures, 26 INT'L LAW. 89, 90 (1992).

166. Article 85(1) provides:

1. The following shall be prohibited as incompatible with the common market: all agreements between undertakings; decisions by associations of undertakings; and concerted practices which may affect trade between Member States and which have as their objective or effect the prevention,

falls within article 85(1) is automatically deemed null and void. In addition, the Commission may impose substantial fines on parties to an agreement that violates article 85(1).¹⁶⁷

The possibility that the agreement will be nullified or fined creates a significant incentive for joint venture participants to notify the agreement to the Commission and to request a negative clearance¹⁶⁸ or a special exemption under article 85(3).¹⁶⁹ Because the Commission has found the majority of joint ventures to fall within article 85(1), at least where the participants are actual or potential competitors,¹⁷⁰ the joint venture

restriction, or distortion of competition within the common market, and in particular those which:

- (a) directly or indirectly fix purchase or selling prices or any other trading conditions;
- (b) limit or control production, markets, technical development, or investment;
- (c) share markets or sources of supply;
- (d) apply dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage;
- (e) make the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts.

TREATY ESTABLISHING THE EUROPEAN ECONOMIC COMMUNITY [EEC TREATY] art. 85(1).

167. Under Regulation 17, the Commission may fine the parties up to 1 million ECU or 10% of the parties' preceding year's turnover, whichever is greater. Council Regulation 17/62 of 21 February 1962, *Premier règlement d'application des articles 85 et 86 du traité*, art. 15(2), 1962 J.O. (13) 204. See generally 2 HAWK, *supra* note 165, at 20; Sara G. Zwart, *Innovate, Integrate, and Cooperate: Antitrust Changes and Challenges in the United States and the European Economic Community*, 1989 UTAH L. REV. 63, 80 & n.74.

168. See Council Regulation 17/62, *supra* note 167, art. 2. A negative clearance basically is a Commission determination that an agreement does not violate article 85(1). See Zwart, *supra* note 167, at 80 n.75.

169. Article 85(3) provides:

The provisions of paragraph 1 may, however, be declared inapplicable in the case of:

- any agreement or category of agreements between undertakings;
 - any decision or category of decisions by associations of undertakings;
 - any concerted practice or category of concerted practices;
- which contributes to improving the production or distribution of goods or to promoting technical or economic progress, while allowing consumers a fair share of the resulting benefit, and which does not:
- (a) impose on the undertakings concerned restrictions which are not indispensable to the attainment of these objectives;
 - (b) afford such undertakings the possibility of eliminating competition in respect of a substantial part of the products in question.

See Commission Notice Concerning Assessment of Cooperative Joint Ventures Pursuant to Article 85 of the EEC Treaty, 1993 O.J. (C 43) 2 (describing Commission's procedures for evaluating cooperative joint ventures under article 85) [hereinafter Notice Concerning Cooperative Joint Ventures].

170. See 2 HAWK, *supra* note 165, at 310, 314; Fine, *supra* note 165, at 93-94, 98-101.

participants must usually prove that the venture qualifies for a special exemption under article 85(3).

This notification procedure can impose significant burdens and create substantial uncertainty for joint venture participants, however. First, the Commission is not required to render a decision within any specified time limit and, in practice, frequently waits years after the notification is filed before granting or refusing an exemption request.¹⁷¹ Second, in conducting its investigation of a notification, the Commission will frequently demand significant amounts of highly sensitive information about the participants and the venture.¹⁷² Third, the Commission is required to grant an exemption for a specified period of time, and, in many cases, this period is shorter than the proposed duration of the venture.¹⁷³ Fourth, the Commission frequently conditions the grant of an exemption upon the parties' agreeing to modify or eliminate provisions deemed unnecessarily restrictive¹⁷⁴ or to comply with certain continuing reporting and operating obligations.¹⁷⁵ Finally, all notices applying for an exemption are published in the Official Journal of the European Communities with an invitation for third parties to make comments.¹⁷⁶

A quicker and easier approval process is available for joint ventures that qualify under various block exemptions or the recent merger control regulation. However, many PJVs will not qualify.

Under a 1984 block exemption for certain R&D joint ventures,¹⁷⁷ any RJV agreement that satisfies the conditions of the block exemption is

171. Fine, *supra* note 165, at 105; Zwart, *supra* note 167, at 85-86.

172. In this regard, the Commission holds broader discovery powers than do the U.S. antitrust authorities, including the power to make on-site inspections without prior notice to those involved. See 2 HAWK, *supra* note 165, at 26-28; Fine, *supra* note 165, at 105.

173. Fine, *supra* note 165, at 105-06.

174. HAWK, *supra* note 165, at 138-39.

175. The Commission may require the parties to make periodic reports on marketing and pricing policy, licensing information, or market share data. In some cases, the Commission also has required the parties to give the Commission advance notice of planned changes in the agreement. *Id.* at 325-26.

176. Fine, *supra* note 165, at 106.

177. Commission Regulation 418/85 of 19 December 1984 on the Application of Article 85(3) of the Treaty to Categories of Research and Development Agreements, 1985 O.J. (L 53) 5 [hereinafter R&D Regulation]. The block exemption was intended to encourage the formation of such ventures by reducing any antitrust uncertainty concerning their legality under article 85(1). Even before the issuance of the 1984 block exemption, however, the EC had adopted a favorable stance towards cooperative research. For example, Regulation 17, issued in 1962, provided that agreements that had as their sole objective joint research to improve techniques did not have to file a notification requesting an individual exemption. Council Regulation 17/62, *supra* note 167. Similarly, in its 1968 Notice on Cooperation, the Commission stated that agreements, whose sole purpose is the joint implementation of R&D projects, the placing of R&D contracts, or the sharing of R&D projects among participants, were not restrictions of competition within the meaning

deemed to fall outside of article 85(1). Unlike the NCRA, the block exemption permits joint exploitation of the results of the cooperative R&D through joint production or joint licensing. Nevertheless, it contains a number of restrictions that disqualify many joint ventures. For example, where the agreement provides only for joint R&D, the parties must be free to license or otherwise exploit the results of the joint R&D independently.¹⁷⁸ Second, where the agreement provides for joint exploitation, the exploitation must relate only to research results which are protected by intellectual property rights or constitute know-how which contributes substantially to technical or economic progress, and the results must be decisive for the manufacture of the contract products.¹⁷⁹ Third, where the agreement provides for joint production, the venture may only supply the products to the participants; it cannot also engage in joint distribution or marketing.¹⁸⁰ Finally, and most importantly, the Regulation imposes strict twenty percent market share limitations on R&D joint ventures that involve joint exploitation.¹⁸¹

The EC's 1989 Merger Control Regulation (MCR),¹⁸² on the other hand, requires that "concentrations"¹⁸³ that have a "Community dimension," including certain "concentrative" joint ventures, be notified

of Article 85(1), and hence did not need to be notified. Commission Notice, O.C. 75/3 (July 19, 1968). *See generally* Notice Concerning Cooperative Joint Ventures, *supra* note 169, at 9-10; 2 HAWK, *supra* note 165, at 341.

178. R&D Regulation, *supra* note 177, art. 2(c). This means, however, that the parties may compete away, through licensing or otherwise, any short-term rents resulting from the R&D.

179. *Id.* art. 2(d). Although this provision is clearly intended to exclude joint ventures that are primarily joint production and/or marketing ventures, it may create uncertainty for joint ventures for which joint R&D is only a component. HAWK, *supra* note 165, at 348.

180. R&D Regulation, *supra* note 177, art. 2(e).

181. Specifically, where the participants are competing manufacturers, the exemption will only apply if, at the time in which the agreement is entered, the combined market shares of the participants with respect to products "capable of being improved or replaced by the contract products" does not exceed 20% of the market for such products in the Common Market or a substantial part thereof. *Id.* art. 3(2). In addition, the 20% market share limitation must be satisfied for the duration of agreements involving joint exploitation. *Id.* art. 3(3). Finally, where the joint production involves components used by the participants in the manufacture of other products, the 20% limitation applies to the latter products for which the components represent a significant part. *Id.*

182. Corrigendum to Council Regulation 4064/89 of 21 December 1989 on the Control of Concentrations Between Undertakings, 1990 O.J. (L 257) 13 [hereinafter MCR]. The MCR became effective in September 1990.

183. Concentrations are deemed to exist where:

- (a) two or more previously independent undertakings merge, or
- (b) one or more persons controlling at least one undertaking, or one or more undertakings, acquire, whether by purchase of securities or assets, by contract or by any other means, direct or indirect control of the whole or parts of one or more other undertakings.

Id. art. 3(1).

to the Commission and cleared by it before they can be implemented. Although the MCR offers certain advantages to "concentrative" joint ventures,¹⁸⁴ it also creates some problems for these ventures. First, although the Commission has issued guidelines to assist in distinguishing "cooperative" joint ventures (still subject to article 85) from "concentrative" joint ventures (subject to the MCR),¹⁸⁵ there may remain considerable uncertainty as to whether a particular joint venture qualifies as a "concentrative" joint venture.¹⁸⁶ In addition, even where a joint venture qualifies as a "concentration" subject to the MCR, complying with the notification is likely to prove burdensome.¹⁸⁷

In summary, the EC, like Japan, has subsidized and encouraged cooperative research of a precompetitive nature as a means of helping European firms catch up with more advanced American and Japanese rivals. This subsidized collaboration, however, does not extend to joint production. As shown above, private PJs in the EC face more burdensome and time-consuming clearance procedures than their U.S. counterparts. Therefore, neither the Japanese nor European policies concerning antitrust and cooperative research provide a precedent for the proposed joint venture legislation.

VI. WEAKNESSES IN THE PROPOSED LEGISLATION

The most obvious weakness with the current bills is that proponents have failed to demonstrate any compelling need for the legislation. Although they claim that fear of antitrust liability deters procompetitive joint ventures, they have produced no clear and convincing evidence that this results. On the contrary, the available evidence suggests that the antitrust laws do not constitute a significant deterrent to the formation of joint ventures.¹⁸⁸ The current law quite clearly indicates that bona fide

184. The major advantages of the MCR are that: first, agreements notified under the MCR do not have to be separately notified under article 85, see, e.g., Fine, *supra* note 165, at 101; Barry E. Hawk, *The EEC Merger Regulation: The First Step Toward One-Stop Merger Control*, 59 ANTITRUST L.J. 195, 202 (1990) [hereinafter Hawk (1990)] and second, the Commission is required to render a decision within strict time limits, basically one month to four months. MCR, *supra* note 182, art. 10; see also Barry E. Hawk, *European Economic Community Merger Regulation*, 59 ANTITRUST L.J. 457, 459 (1991) [hereinafter Hawk (1991)]; Patrick Thieffry et al., *The Notification of Mergers Under the New EEC Merger Control Regulation*, 25 INT'L LAW. 615, 618-19 (1991).

185. Commission Notice Regarding the Concentrative and Cooperative Operations under Council Regulation (EEC) 4064/89 of 21 December 1989 on the Control of Concentrations between Undertakings, 1990 O.J. (C 203) 10.

186. See Fine, *supra* note 165, at 101-02; Hawk (1990), *supra* note 184, at 202-06; Hawk (1991), *supra* note 184, at 460-61; Thieffry et al., *supra* note 184, at 621-24.

187. Professor Hawk has described the MCR notification form as requiring "something like a Hart-Scott second request combined with a white paper." Hawk (1991), *supra* note 184, at 462; see also Thieffry et al., *supra* note 184, at 628-34.

188. See *supra* Section IV.D.

PJVs will be evaluated under the rule of reason and that efficiencies resulting from economic integration will be weighed against any restrictions on competition. In addition, there is no evidence of any alarming pattern of public or private antitrust litigation brought against joint ventures; rather, the available evidence suggests that the total number of public and private antitrust suits of all kinds has declined dramatically within the last ten years.

The best argument made by proponents of the legislation is that businessmen may be under a "misperception" that the legality of PJVs remains uncertain under antitrust laws. Not only is this an unusual justification—that the law should be changed because the business community misunderstands it¹⁸⁹—but there also appears little evidence to support it. As previously noted, the American Bar Association's Section of Antitrust Law was unable to confirm the existence of such a widespread misperception.¹⁹⁰ Moreover, the large and increasing number of U.S. firms that are participating in joint ventures, despite the difficulties in collective management, suggests that the antitrust laws hardly constitute a significant deterrent.¹⁹¹

A second major weakness of the legislation is that the pending bills do not appear targeted to achieve their ostensible goals. Legislative proponents claim that the bills, by encouraging cooperation in production, will help U.S. firms become more innovative and competitive. Unfortunately, neither the House nor Senate bills requires a PJV to engage in cooperative R&D as well as joint production. Nor do the bills require qualified joint ventures to produce new or innovative products.¹⁹² Finally, neither bill contains any requirement that the joint venture be involved in a high technology sector. Accordingly, there is little reason to expect the legislation will achieve its apparent purpose of encouraging the formation of innovative joint ventures in high-technology industries.

More importantly, the proposed legislation may well have significant adverse effects on the U.S. economy and on U.S. competitiveness. First, the proposed legislation may encourage greater collusion and cartelization by U.S. firms. As indicated above, coop-

189. See 1989b House Hearings, *supra* note 4, at 126-27 (statement of Edward Rock).

190. See *supra* note 128.

191. See *supra* Section IV.D.

192. The Senate bill apparently attempts to deal with this problem by requiring joint ventures that use existing facilities produce or process a "new process or technology." Unfortunately, the language is insufficient to solve the problem. First, there is no explanation of what constitutes a "new product or technology." Accordingly, it appears that simply introducing a slightly modified version of an existing product where no new technology is involved—such as a sterling silver garlic press—would satisfy the requirement. Second, where the joint venture builds a new facility, there is no extra requirement that the venture produce a new product or use a new technology.

eration at the production or marketing levels results in significantly more collusion than cooperation at the R&D level.¹⁹³ But the proposed legislation contains no provisions that effectively reduce this risk. For example, there are no limits on the allowed market shares, either individual or combined, of the participants. However, the anticompetitive risks of such joint ventures rise substantially as the combined market shares and market power of the participants increase. At the same time, it appears unlikely that economies of scale or other justifications for collaborative production warrant participation by firms holding a majority share of the relevant market. Accordingly, the bills should include some requirement that participants collectively holding market power demonstrate some need for such broad participation.

In addition, although the bills prohibit participants from engaging in joint marketing and from exchanging information concerning "costs, sales, profitability, prices . . . that is not reasonably required to carry out the purposes of the venture,"¹⁹⁴ they do not prevent participants from combining their competing production facilities. Thus, theoretically, an entire domestic industry could cooperate in production, jointly determine total output, and hence, indirectly agree on price.¹⁹⁵ Moreover, despite the statutory prohibition, it is not unlikely that production cooperation may lead to cooperation or collusion in pricing or other dimensions of competition.

Finally, even if production cooperation does not lead to explicit collusion, it may reduce rivalry among the participants. Again, this danger increases as the proportion of industry firms participating increases. The reduction in rivalry in turn could lead to higher prices and reduced innovation, making the participants less able to respond to foreign competition.¹⁹⁶

The danger of collusion and reduced rivalry suggests that increasing the number of PJVs will heighten the need for vigilant antitrust enforcement. Unfortunately, the bills weaken, if not eviscerate, antitrust enforcement for PJVs. First, although the bills require joint venture participants to file notifications with the antitrust authorities if they wish immunity from treble damages, the information required under the notification is not sufficiently detailed¹⁹⁷ to enable the antitrust authorities to perform an accurate evaluation of the competitive effects of the

193. See *supra* text accompanying notes 59-62.

194. See, e.g., S. REP. NO. 146, *supra* note 5, at 21, reprinted in 61 Antitrust & Trade Reg. Rep. at 353.

195. See *supra* note 58.

196. See, e.g., 1989b House Hearing, *supra* note 4, at 142 (written response of Edward Rock); PORTER, *supra* note 46, at 117, 169-70, 530.

197. See *supra* note 174.

venture.¹⁹⁸ In addition, since the proposed legislation would provide no new resources to the relevant regulators, there is a significant risk that they will be overwhelmed if the legislation results in a flood of filings.¹⁹⁹ Finally, the proposed legislation contains no provision for periodic monitoring by the antitrust enforcement agencies. Accordingly, the authorities may receive no notice should the joint venture subsequently engage in anticompetitive practices or seek trade protection to eliminate foreign competition.

The effect of the bill on private antitrust enforcement will be even greater. Although the issue of the appropriate measure of antitrust damages is beyond the scope of this article,²⁰⁰ it appears reasonable to infer that a reduction in the damage multiplier from three to one will reduce the incentive for private plaintiffs to bring suit.²⁰¹ Similarly, the possibility that a plaintiff will be ordered to pay the defendants' attorneys' fees may deter even meritorious suits by plaintiffs with less resources.²⁰² At the same time, by reducing the likelihood of private suits, these changes will increase the incentives for joint venture participants to engage in collusion or other anticompetitive behavior.²⁰³

The relevant question then becomes whether a reasonable justification exists for singling out PJVs for special treatment as opposed

198. The limited information required and the ministerial review provided under the 1984 NCRA was arguably sufficient for RJs, given their limited possible anticompetitive effects. This does not mean, however, that such information and review are sufficient where the competitive dangers are much more significant.

199. See 1989b House Hearing, *supra* note 4, at 131-32 (statement of Edward Rock); *id.* at 433 (statement of Arthur Kaplan).

200. The literature on antitrust damages stems largely from the writings of Gary Becker on the economic theory of deterrence. See Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169 (1968). Breit and Elzinga were among the first to apply Becker's approach to antitrust damages. See KENNETH G. ELZINGA & WILLIAM BREIT, *THE ANTITRUST PENALTIES: A STUDY IN LAW AND ECONOMICS* (1976). For more recent analyses, see, e.g., SECTION OF ANTITRUST LAW, A.B.A., MONOGRAPH NO. 13, *TREBLE-DAMAGES REMEDY* (1986); WARREN F. SCHWARTZ, *PRIVATE ENFORCEMENT OF THE ANTITRUST LAWS: AN ECONOMIC CRITIQUE* (1981); William Breit & Kenneth G. Elzinga, *Private Antitrust Enforcement: The New Learning*, 28 J.L. & ECON. 405 (1985); Frank H. Easterbrook, *Detrebling Antitrust Damages*, 28 J.L. & ECON. 445 (1985); William M. Landes, *Optimal Sanctions for Antitrust Violations*, 50 U. CHI. L. REV. 652 (1983); A. Mitchell Polinsky, *Detrebling versus Decoupling Antitrust Damages: Lessons from the Theory of Enforcement*, in *PRIVATE ANTITRUST LITIGATION*, *supra* note 118, at 7.

201. See, e.g., 1989b House Hearing, *supra* note 4, at 257 (statement of Joseph Alioto); *id.* at 457 (written response of Arthur Kaplan); Peter W. Rodino, *Let's Fix Only What's Broken: Some Thoughts on Proposed Reform of Private Antitrust Litigation*, in *PRIVATE ANTITRUST LITIGATION*, *supra* note 118, at 421.

202. See Shapiro & Willig, *supra* note 47, at 128.

203. See 1989b House Hearing, *supra* note 4, at 457 (written response of Arthur Kaplan). See generally Edward D. Cavanaugh, *Detrebling Antitrust Damages: An Idea Whose Time Has Come?*, 61 TUL. L. REV. 777, 786-87 (discussing incentive effects of treble damages on potential violators); Salop & White, *supra* note 115, at 1017-21.

to other antitrust concerns.²⁰⁴ In the case of RJs, the justification for the 1984 NCRA was that R&D activities suffered from certain unique market failures and that RJs could help correct those market failures without imposing any significant anticompetitive costs.²⁰⁵ No similar market failures afflict production activities, however. Moreover, the potential benefits offered by PJs appear considerably smaller, while the potential anticompetitive costs appear significantly larger. The reasoning underlying the special treatment of R&D joint ventures thus does not appear to extend to PJs.

A final weakness of the House bill involves the attempt to benefit U.S. companies and workers by requiring that plant facilities be located in the United States and by limiting the companies eligible to qualify for protection.²⁰⁶ These provisions appear both protectionist and misguided. Foreign firms clearly offer access to foreign markets and, to an increasing extent, possess technological information that would be valuable to U.S. companies.²⁰⁷ These protectionist provisions could pose a barrier to cooperation with technologically advanced foreign firms and thereby substantially undermine the legislative purpose.²⁰⁸ In addition, such

204. One such justification is suggested by Shapiro and Willig. They note that the economic theory of deterrence suggests that multiple damages are more appropriate for violations that are less likely to be detected. Given this, they argue that the reduction in damages under the Act may be a suitable quid pro quo for notification. Shapiro & Willig, *supra* note 47, at 126; *see also* Easterbrook, *supra* note 200, at 456-57.

This argument appears unpersuasive. The examples usually cited of antitrust violations that are likely to go undetected include such acts as price fixing or market division, where the actions are intentionally concealed. *Id.* at 456. In the case of the formation of a PJ, however, the fact of the agreement is not generally hidden, but rather publicized in the press or at least generally known by competitors in the industry. Considering the limited information required by the notification, it would appear to provide no information that is not publicly available or available to those acquainted with the industry.

205. *See supra* Section III.B. *But see* Easterbrook, *supra* note 200, at 456 (arguing that single damages for violations by RJs is too low).

206. The Senate Bill, as it emerged from the Senate Judiciary Committee, contained provisions similar to those in the House version. Prior to passage, however, the Senate amended these provisions to reduce their protectionist tone. Unfortunately, the substitute language is vague and ambiguous.

207. Empirical studies of joint ventures indicate that technological transfer and market access are the two main reasons a firm may enter into a joint venture with a foreign partner. *See, e.g.*, MOWERY & ROSENBERG, *supra* note 28, at 248; Mowery, *supra* note 46, at 13-15.

208. Mowery and Rosenberg come to a similar conclusion:

Restrictions or controls on international collaborative ventures involving U.S. firms do not appear to be an effective means to improve U.S. international competitiveness and in fact might impair competitiveness. The complexity of international collaborative ventures, the fact that the pattern and impact of these ventures vary considerably across industries, and the historical evidence that restrictions on technology transfer are either

restrictions may well result in other countries retaliating by restricting participation by U.S. companies in foreign joint ventures.²⁰⁹

VII. CONCLUSION

One of the great virtues of the U.S. antitrust laws is that they are sufficiently broad and general so that the courts have been able to adapt their language to changing market and technological conditions so as to ensure that U.S. markets remain competitive and efficient. Congress has generally recognized this virtue and has accordingly shown considerable reluctance to make special exceptions and exclusions.²¹⁰ Before creating a special exemption or immunity, Congress has generally required a "convincing prior showing of public interest or compelling economic need."²¹¹ Congress should demonstrate similar restraint here. It should also demand substantial and convincing empirical evidence that justifies the extension of the NCRA's special protection to PJVs before it enacts any of the proposed legislation. Such evidence has not yet been produced.

There is obviously considerable political appeal to passing legislation intended to spur U.S. competitiveness, especially where the legislation will not require substantial federal expenditures. In the case of the proposed PJV legislation, however, this political allure should be resisted. If passed, the legislation at best will have little effect on the economy; at worst, it will foster collusion, undermine U.S. competitiveness, and impose significant costs on U.S. consumers.

ineffective or perverse in their impacts . . . all argue against controls on collaborations involving non-defense technologies.

MOWERY & ROSENBERG, *supra* note 28, at 252; see also 1990 Senate Hearing, *supra* note 4, at 102 (statement of Joseph Brodley); 1989b House Hearing, *supra* note 4, at 140 (statement of George Heaton); *id.* at 261 (statement of Thomas Jorde).

209. See, e.g., 1990 Senate Hearing, *supra* note 4, at 102 (statement of Joseph Brodley); 1989 House Hearing, *supra* note 4, at 140 (written response of George Heaton); *id.* at 328 (statement of J.D. Huehler, President, IBM Corporation).

210. See 1989b House Hearing, *supra* note 4, at 95 (statement of George Heaton); Hamilton Fish, Jr., *Antitrust Relief and the House Judiciary Committee*, 35 ANTITRUST BULL. 219 (1990); Rodino, *supra* note 201, at 421-23.

211. Fish, *supra* note 210, at 222; see also NATIONAL COMM'N FOR THE REVIEW OF ANTITRUST LAWS & PROCEDURES, REPORT TO THE PRESIDENT AND THE ATTORNEY GENERAL 186 (1979) ("Each existing or proposed exemption should be justified in terms of empirically demonstrated characteristics of the specific industry that make competition unworkable. The defects in the market place necessary to justify an antitrust exemption must be substantial and clear.").

ARTICLE

SOFTWARE LITIGATION IN THE YEAR 2000: THE EFFECT OF OBJECT-ORIENTED DESIGN METHODOLOGIES ON TRADITIONAL SOFTWARE JURISPRUDENCE

DAVID M. BARKAN[†]

Table of Contents

I.	INTRODUCTION	315
II.	TRADITIONAL SOFTWARE DESIGN METHODS	317
III.	THE OBJECT-ORIENTED MODEL	320
	A. Overview	320
	B. The Basic Concepts of the Object-Oriented Model	321
	C. The Process of Designing Software Under the Object-Oriented Model	335
IV.	COPYRIGHT PROTECTION FOR OBJECT-ORIENTED SOFTWARE	341
	A. <i>Whelan</i> and Its Progeny	343
	B. The Filtering Approach	346
	C. Economic Balancing Approach	352
	D. Copyright Doctrine Properly Applied Provides Little Protection for Object-Oriented Software	354
V.	PATENT PROTECTION FOR OBJECT-ORIENTED SOFTWARE	358
	A. Patentable Subject Matter	358
	B. Non-Obviousness and the Relevant Prior Art	363
VI.	IMPLICATIONS FOR INNOVATION POLICY	365

I. INTRODUCTION

Ever since *Whelan Associates v. Jaslow Dental Laboratory*,¹ courts dealing with software have attempted to understand the process of

© 1993 David M. Barkan.

[†] J.D. 1992, School of Law (Boalt Hall), University of California, Berkeley; A.B. 1987, Harvard University. The author wishes to thank Peter Menell and Jeremy Barkan for their helpful comments and criticism.

1. 797 F.2d 1222 (3d Cir. 1986), cert. denied, 479 U.S. 1031 (1987).

software design. In doing so, they have focused almost exclusively on traditional methods of procedural programming and "top-down" design.² Moreover, most commentators on software protection present this traditional model of software design before articulating their proposed level of protection.³ While this model accurately reflects software design in the 1980s, the traditional approach is not well-suited to the exponential increase in size and complexity that will characterize software projects in the 1990s.⁴ As one critic of traditional design noted: "If builders built buildings the way programmers wrote programs, then the first woodpecker that came along would destroy civilization."⁵

In order to address the problem of complexity, programmers are likely to turn to object-oriented design and analysis because it allows programmers to adopt an entirely different approach toward problem solving and strongly encourages the development of libraries of reusable software "components." The next generation of software cases is likely to involve alleged infringement of complete programs designed according to the object-oriented model or alleged infringement of reusable software libraries.⁶ This article attempts to explain object-oriented design to judges

2. See *id.* at 1229-31 (describing the process as identifying the problem, outlining the solution, and then creating a flowchart, modules, and subroutines); Computer Assocs. Int'l v. Altai, Inc., 775 F. Supp. 544, 559 (E.D.N.Y. 1991) (citing expert testimony describing a computer program as "made up of sub-programs and sub-sub-programs, and so on"), *aff'd in relevant part*, 982 F.2d 693 (2d Cir. 1992); E.F. Johnson v. Uniden Corp. of Am., 623 F. Supp. 1485, 1501-02 n.17 (D. Minn. 1985); SAS Inst., Inc. v. S & H Computer Sys., 605 F. Supp. 816, 825 (M.D. Tenn. 1985) ("Beginning with a broad and general statement of the overall purpose of the program, the author must decide how to break the assigned task into smaller tasks, each of which must in turn be broken down into successively smaller and more detailed tasks."). Other cases have implicitly adopted this model in determining similarities between the plaintiff's and defendant's programs. See, e.g., Plains Cotton Co-op v. Goodpasture Computer Serv., 807 F.2d 1256, 1260 (5th Cir. 1987) (analyzing evidence of organizational copying), *cert. denied*, 484 U.S. 821 (1987). Pearl Systems v. Competition Electronics, 8 U.S.P.Q.2d (BNA) 1520, 1523-24 (S.D. Fla. 1988) (accepting expert testimony on similarities at the "subroutine" and "module" level); Q-Co Indus., Inc. v. Hoffman, 625 F. Supp. 608, 614-15 (S.D.N.Y. 1985) (comparing similarities in program "modules").

3. See 3 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT, § 13.03 [F] at 13-78.30 to .32 (1991); Peter S. Menell, *An Analysis of the Scope of Copyright Protection for Application Programs*, 41 STAN. L. REV. 1045, 1055-56 (1989); David Nimmer et al., *A Structured Approach to Analyzing the Substantial Similarity of Computer Software in Copyright Infringement Cases*, 20 ARIZ. ST. L.J. 625, 637-38 (1988); Gary L. Reback & David L. Hayes, *The Plains Truth: Program Structure, Input Formats and Other Functional Works*, COMPUTER LAW., Mar. 1987, at 1, 5.

4. For a general discussion of the inherent complexity of modern software see GRADY BOOCHE, OBJECT-ORIENTED DESIGN WITH APPLICATIONS 2-8 (1991) [hereinafter BOOCHE, OBJECT-ORIENTED DESIGN].

5. Booch, *Reuse of Software Components Could Reduce Costs*, GOV'T COMPUTER NEWS, Sept. 25, 1987, at 86.

6. The object-oriented model is most likely to arise first in cases involving graphical user interfaces. Apple Computer distributes an object library called MacApp which provides programmers with many tools necessary to implement the Macintosh user

and lawyers and to raise the problems that this new method of writing software will pose for courts applying traditional copyright and patent doctrines. The article presumes no technical background, but it does assume some familiarity with basic copyright and patent concepts. Part II of this article briefly reviews the traditional model of software development and notes some of the limitations that are encouraging the switch to object-oriented design. Part III presents the basic concepts in the object-oriented model and explains the process of designing software under this model. Part IV then evaluates the scope of copyright protection both for complete programs written using the object-oriented model and for reusable object libraries. While some existing case law and commentary could be used to support broad protection for such programs, this Section shows that pure copyright doctrine should not provide any protection for the aspects of the program that make it object-oriented. Part V analyzes patent protection for object-oriented software and concludes that sufficiently innovative programs and libraries could qualify for patent protection. Finally, Part VI considers the potential effects that copyright and patent protection would have on the growth of object-oriented technology.

II. TRADITIONAL SOFTWARE DESIGN METHODS

Traditional programming languages are based on the concept of a procedure, which allows programmers to write a small section of code which performs one small task.⁷ Well-written programs begin with a top-down approach to software design that has been described by Nimmer:

In practice, a programmer usually will start with a general description of the function that the program is to perform. Then, a specific outline of the approach to this problem is developed, usually by studying the needs of the end user. Next, the programmer begins to develop the outlines of the program itself, and the data structures and algorithms to be used. At this stage, flowcharts, pseudo-code, and other symbolic representations often are used to help the programmer organize the program's structure. The programmer will then break down the problem into modules or subroutines, each of which addresses a particular element of the overall programming

interface. Similarly, Symantec Corp. distributes a similar object-oriented library with its Pascal and C compilers. For a description of major software vendors working on object-oriented development, see Richard K. Aeh, *OOPS Are Picking Up Speed*, J. SYS. MGMT., Feb. 1991, at 19; Rick Whiting, *The Quest for a Better Way to Develop Software*, ELECTRONIC BUS., July 10, 1989, at 16. For a general discussion of the some of the problems that may slow widespread adoption of the object-oriented model, see Jeff Moad, *Cultural Barriers Slow Reusability*, DATAMATION, Nov. 15, 1989, at 87.

7. See generally ELLIOT B. KOFFMAN, PROBLEM SOLVING AND STRUCTURED PROGRAMMING IN PASCAL 52-59 (1985). The terms "procedure," "module," and "subroutine" are used interchangeably to denote a small number of programming instructions which perform a single task.

problem, and which itself may be broken down into further modules and subroutines. Finally, the programmer writes specific source code to perform the function of each module or subroutine, as well as to coordinate the interaction between modules or subroutines.⁸

This process has also been described as "functional decomposition," since the "primary question addressed by the systems analysis and design is WHAT does the system do [or] what is its *function*?"⁹ The complex function identified at this stage must be further decomposed into smaller functions, a process which is repeated until the problem can be "expressed as some combination of many small, solvable problems."¹⁰

Several important implications flow from this model that affect the way programmers are taught to approach problem solving. First, the traditional model forces programmers to focus on specific tasks that must be achieved, resulting in strong analysis of the procedures and functions that must be used, but little emphasis on data structures.¹¹ Data structures are usually conceived only after the procedures have been generally defined.¹² While diligent programmers may go back and rethink some of their procedures after analyzing their data structures, as a general rule, "[i]n a typical procedural programming language such as C or Pascal, programmers approach data and algorithms as separate entities."¹³ Second, data to be used by several procedures are usually

8. NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.31 to .32.

9. Brian Henderson-Sellers & Julian M. Edwards, *The Object-oriented Systems Life Cycle*, COMM. ACM, Sept. 1990, at 142, 145 (emphasis in original).

10. Menell, *supra* note 3, at 1055 (citation omitted).

11. See Tim Korson & John D. McGregor, *Understanding Object-Oriented: A Unifying Paradigm*, COMM. ACM, Sept. 1990, at 40, 46.

The term "data structure" denotes the symbolic representation of a particular area of memory where specific data will be stored. For example, if we wished to store data about this article, we could create a data structure in Pascal called a "record" that includes various fields for storing different pieces of information. The record could be defined as follows:

```
TheArticle = record
    Author: str;
    NumberOfPages: int;
    Issue: int;
    StartingPage: int;
end;
```

In this data structure, the name of the entire record is "TheArticle." The name of the author is stored in the field "Author," the length of the article is stored in the "NumberOfPages" field, the issue number is stored in the "Issue" field, and the starting page in the issue is stored in the "StartingPage" field.

12. An obvious exception would be a database program since the programmer is usually given highly specific information on what type of information must be stored in the database, thus leading to an initial focus on data structures.

13. Randy Leonard, *OOP: The Future for Macintosh Development*, MACTECH Q., Spring 1989, at 22.

defined in one place and can be accessed by any module or subroutine.¹⁴ While this approach may seem logical and efficient, it creates severe problems as soon as a complex program needs to be updated or revised since it "leads to the stack of dominoes effect familiar to anyone working in program maintenance whereby changes to one part of a software system often cause a problem in an apparently dissociated program area."¹⁵ As a result, each time a data structure needs to be changed or refined, one would have to identify all the procedures that rely on the old definition of the data structure and change them accordingly. Finally, since the top-down approach forces programmers to think in terms of a series of single functions, programmers are less likely to incorporate evolutionary changes in the data structures into the big picture of the overall system.¹⁶

In general, the traditional model provides few easy ways to reuse existing pieces of software, thus making software development less efficient than other engineering disciplines that are accustomed to reusing existing components.¹⁷ The close dependence between procedures and the specific definition of data structures makes it extremely difficult to reuse selected procedures in a different project. While programmers certainly copy solutions to certain problems from earlier projects and from public domain sources, rarely can programmers lift an existing procedure from another project and incorporate it into their program without serious modification.

The traditional model leads to several additional problems from a project management perspective. While the top-down model has worked well until recently, software projects in the 1990s will involve new levels of complexity and will require better systems for managing multiple programmers working on different parts of one system. As software will increasingly be required to model real world behavior in meaningful

14. Ray Duncan, *Redefining the Programming Paradigm*, PC MAG., Nov. 13, 1990, at 526 ("[Y]ou visualize the data to be worked on as sitting in one place, while various routines call upon each other to do things to the data."); Henderson-Sellers & Edwards, *supra* note 9, at 146; Chris Terry, *Objects Facilitate Modular, Reusable Code*, EDN, Nov. 9, 1989, at 85.

15. Henderson-Sellers & Edwards, *supra* note 9, at 146.

16. *Id.*

17. Peter Coffee, *Honing the Software Equivalent of the Transistor*, PC Wk., Sept. 25, 1989, at 38 (comparing object-oriented development to electrical engineering which emphasizes the reuse of standard building block components); Chris Terry, *Reusable Software Requires Building Blocks*, EDN, Jan. 3, 1991, at 59 ("Power-supply designers don't have to manufacture their own capacitors and resistors, let alone their connectors, nuts and bolts. They use parts that conform to universal (or at least widely accepted) standards. Yet software-systems designers have to write code for almost every function in their system except the services the operating system provides.").

ways, the human capacity for managing complexity will be severely tested.¹⁸

While traditional methods of functional decomposition provide one way of managing complexity, they are constrained in ways that make it difficult to manage projects that require large numbers of programmers. Since changes made by one programmer can effect other disparate parts of the program in a ripple fashion, traditional design eventually breaks down in extremely complex projects. Of course, the traditional model can be improved by forcing each programmer to adhere to detailed specifications that dictate how each programmer's portion interacts with the program as a whole.¹⁹ While this modification of the traditional method was sufficient for "programming-in-the-large," it may not be sufficient for the "programming-in-the-colossal" that will be required in the 1990s.²⁰

III. THE OBJECT-ORIENTED MODEL

A. Overview

While object-oriented principles were originally developed in the 1960s,²¹ the current object-oriented model is an attempt to address the modern problem of "programming-in-the-colossal." Generally speaking, the object-oriented model emphasizes use of small, discrete components which can be used without any knowledge of how they work internally, thus breaking the tight dependency between data structures and procedures that constrained the traditional model. Several positive benefits flow directly from this one assumption:

Object-oriented decomposition yields smaller systems through the reuse of common mechanisms, thus providing an important economy of expression. Object-oriented systems are also more

18. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 14 ("As we first begin to analyze a complex software system, we find many parts that must interact in a multitude of intricate ways, with little perceptible commonality among either the parts or their interactions: this is an example of disorganized complexity. As we work to bring organization to this complexity through the process of design, we must think about many things at once. For example, in an air traffic control system, we must deal with the state of many different aircraft at once, involving such properties as their location, speed, and heading. . . . Unfortunately, it is absolutely impossible for a single person to keep track of all of these details at once.").

19. *Id.* at 29-30 (discussing the development of modular programming techniques in late third-generation programming languages).

20. *Id.* at 32.

21. For a good history of the development of object-oriented languages, see Duncan, *supra* note 14 (noting that object-oriented programming languages were developed first in 1967 and then implemented in the 1970s at Xerox's Palo Alto Research Center, but were not practical for true commercial programming projects until the development of fast and efficient C++ compilers).

resilient to change and thus better able to evolve over time, because their design is based upon stable intermediate forms. Indeed, object-oriented decomposition greatly reduces the risk of building complex software systems, because they are designed to evolve incrementally from smaller systems in which we already have confidence.²²

Object-oriented programming is better suited for complex projects, because it more naturally models complex behavior in the real world. Through the concept of *inheritance*,²³ object-oriented programming focuses on hierarchical relationships between different components in a particular real-world system. For example, rather than considering a rectangle and a square to be two separate objects, the object-oriented model views the square as a particular kind of rectangle with special properties. This view is far more than a simple conceptual device. On a practical level, a programmer who wishes to simulate the behavior and properties of a square does not have to start from scratch; rather, the programmer starts with a model of the rectangle which may already exist in a library and then simply adds those properties that are unique to the square. Moreover, this model closely fits the way humans approach physical systems in the real world:

For example, with just a few minutes of orientation, an experienced pilot can step into a multi-engine jet aircraft he or she has never flown before, and safely fly the vehicle. Having recognized the properties common to all such aircraft, such as the functioning of the rudder, ailerons, and throttle, the pilot primarily needs to learn what properties are unique to that particular aircraft. If the pilot already knows how to fly a given aircraft, it is far easier to know how to fly a similar one.²⁴

As a result, the programmer can approach a new software project as an incremental learning process that closely models real behavior. A programmer seeking to simulate the behavior of an F-15 fighter jet would first start with programs that simulate the behavior of a standard commercial jet. If those earlier programs were written according to object-oriented principles, the programmer can simply start with the old program and then incrementally add those behaviors and processes that make the F-15 different from the commercial jet.

B. The Basic Concepts of the Object-Oriented Model

Before further exploring this approach to software design, it is necessary to define some basic concepts that are essential to the object-oriented model. Rather than present these concepts in a vacuum, it is useful to illustrate them in the context of an actual software project.

22. BOOCHE, OBJECT-ORIENTED DESIGN, *supra* note 4, at 16.

23. See *infra* Section III.B.2.

24. BOOCHE, OBJECT-ORIENTED DESIGN, *supra* note 4, at 12.

Assume that you are given the following software project which we will call "QuadWorld."²⁵ You are told that QuadWorld must allow the user to draw and manipulate various types of quadrilaterals,²⁶ such as squares, rectangles, parallelograms, and rhombi. Specifically, the user must be able to choose a particular shape, draw the shape, rotate the shape, move the shape in any direction and erase the shape. The user should also be able to have the program calculate the area of any selected shape. We will use these requirements to illustrate the basic principles of the object-oriented model.

1. OBJECTS, ENCAPSULATION, AND MESSAGES

An *object* is the basic programming unit in the model and essentially combines the traditional concepts of data structures and procedures into a single entity. For example, an object designed to represent a rectangle would include two pieces of data: length and width. It would also contain a complete set of procedures to draw, rotate, move, and otherwise manipulate the rectangle. In this way, the object captures both the state (data regarding length and width) of the rectangle as well as its behavior (the set of procedures for manipulating the rectangle).²⁷ Once the object is defined, any other part of QuadWorld can use that rectangle by simply sending it a *message*, which is a command telling the object to perform one of its defined behaviors. The definition of such a rectangle object might look like this:²⁸

25. This example is taken directly from KURT J. SCHMUCKER, OBJECT-ORIENTED PROGRAMMING FOR THE MACINTOSH 32-35 (1986). Where figures or drawings are adapted directly from this text, they will be appropriately cited.

26. A quadrilateral is any geometric figure with four sides.

27. For a more formal definition of an object, see Korson & McGregor, *supra* note 11, at 42:

Objects are the basic run-time entities in an object-oriented system. Objects take up space in memory and have an associated address like a record in Pascal or a structure in C.

The arrangement of bits in an object's allocated memory space determines that object's state at any given moment. Associated with every object is a set of procedures and functions that define the meaningful operations on that object. Thus, an object encapsulates both state and behavior.

28. Each object-oriented programming language uses slightly different terms to define the basic concepts of object-oriented programming. The definition given here is not meant to mimic any particular language but rather to provide an easily understandable illustration of the basic concepts.

Object Definition of a Rectangle:

Internal Data:

length

width

current position of the top left corner of the rectangle on the screen

Messages that the object is able to perform:

Create, Draw, Move, Stretch, Rotate, Calculate Area

Internal implementation of those messages:

Calculate Area:

Area = length X width

Create: Code for implementing the create message

Draw: Code for implementing the draw message

Move: Code for implementing the move message

(Code for remaining messages as above)

Several observations must be drawn from this definition. First, note that the data fields are called "*internal*" data. In the object-oriented model, data structures are completely private to the object and cannot be used directly by any other part of the program. Similarly, the internal implementation of each message, such as the actual source code that would draw the rectangle, is also private to the object. When another part of the program wishes to draw a rectangle, it simply sends the "Draw" message to the object. Other parts of the program do not care how the object implements that message or what data structures the object uses to perform the draw message. In essence, the other part of the program tells the object "draw yourself, and I don't care how you do it."

This process of "hiding" the internal data structures and the implementation of messages within the object illustrates the principle of *encapsulation*. Since "no part of a complex system should depend on the internal details of any other part,"²⁹ encapsulation is critical to successful "programming-in-the-colossal." Since the object's internal structure is private to the object, changes can be made to the internal structure without having any effect on the rest of the program. For example, the definition of the object could be changed as follows (changes are shown in *italics*):

29. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 45. Booch defines encapsulation as the "process of hiding all of the details of an object that do not contribute to its essential characteristics." *Id.* at 46.

Alternative Object Definition of a Rectangle:

Internal Data:

Position of the top, left corner

Position of the bottom, right corner

(Note that the length and width pieces of data have been deleted)

Messages that the object is able to perform:

Create, Draw, Move, Stretch, Rotate, Calculate Area

Internal implementation of those messages:

Calculate Area:

Area = (first coordinate of bottom, right corner – first coordinate of the top left corner) X (second coordinate of the top left corner – second coordinate of the bottom right corner)

Create: Code for implementing the create message

Draw: Code for implementing the draw message

Move: Code for implementing the move message

(Code for remaining messages as above)

The beauty of the object-oriented model is that these internal changes can be made with absolutely no effect on any other part of the program. Unlike traditional programming, where changes in one procedure ripple through many other parts of the program, the internal definitions of objects can be changed many times with only minimal effect on the overall program. Moreover, tasks can be divided among numerous programmers without requiring the extensive coordination that is necessary when writing code with traditional methods. Each programmer need only be told the desired characteristics of a particular object; how the programmer chooses to write the internal code for that object doesn't matter to the other programmers working on the project. In fact, an object with the desired characteristics may already exist in a company-owned library of objects. As a result, many tasks can be completed without requiring any programmer to write new code.

In fairness to the traditional model, good programming practice can produce some of this modularity within procedural techniques. However, such modularity was usually achieved in a sporadic, informal, and inconsistent fashion.³⁰ In contrast, the object-oriented model requires

30. While late third-generation programming languages allow for separately compiled modules, this concept was used primarily to allow several programmers to work on the same project rather than as an independent tool for abstraction:

Modules were rarely recognized as an important abstraction mechanism; in practice they were used simply to group logically related subprograms. Most languages of this generation, while supporting some sort of modular structure, had few rules that required semantic consistency among module interfaces . . . Unfortunately, because most of these languages had dismal

programmers to incorporate the concept of encapsulation into their basic approach to building software.³¹

2. CLASSES AND INHERITANCE

The concept of a *class* extends the object-oriented model by providing a way to describe a group of objects that have similar properties and behavior. For example, if more than one rectangle is needed at a time, we need not define each rectangle individually. Instead, we define a *class* of rectangles, which serves as a template for creating rectangle objects. Each rectangle object is called an *instance* of the rectangle class. A class definition includes the data that describe each object, called *instance variables*, the messages that the classes must accept,³² and the code that implements each message, known as *methods*.³³ A class can also be thought of as a "factory," which is like a

cookie cutter that stamps out new instances of its class, new objects, whenever necessary. All instances of a given class have the same structure although the actual data stored in any one object may be different from the data stored in another object of the same type, just as all cookies stamped out with the same cookie cutter have the same shape even if they are made of different types of dough and decorated differently.³⁴

In other words, if the user wants to draw two different rectangles of different sizes, our program will use the class definition for the rectangle to create two rectangle objects which possess the same types of behavior but have different values stored in their width and length variables.

The concept of a class does not become truly powerful, however, until combined with the concept of *inheritance*.³⁵ Inheritance is the primary organizing principle behind object-oriented design and the major reason why this model naturally follows the hierarchical structure of real-

support for data abstraction and strong typing, such [semantic] errors could be detected only during execution of the program.

Id. at 30.

31. See the discussion of the object-oriented design process *infra* Section III.C.

32. The complete set of messages accepted by a particular class is sometimes called the class's protocol. SCHMUCKER, *supra* note 25, at 17.

33. Again, the precise terminology varies somewhat depending on the particular object-oriented programming language. Instance variables, messages, and methods are the terms used by Object Pascal and Symantec Corporation's object extensions to the C programming language. *See, e.g.*, SCHMUCKER, *supra* note 25, at 17; SYMANTEC CORP., THINK C OBJECT-ORIENTED PROGRAMMING MANUAL 20 (1991) [hereinafter THINK C OBJECT-ORIENTED PROGRAMMING MANUAL]. The programming language C++ instead uses the terms data member, message, and member function, respectively. *See generally* BJARNE STROUSTRUP, THE C++ PROGRAMMING LANGUAGE (1st ed. 1986).

34. SCHMUCKER, *supra* note 25, at 18.

35. Inheritance also distinguishes the class concept from the idea of user-defined types that is found in Pascal and C. *Id.* at 21.

world entities.³⁶ Given any particular class, we can define a *sub-class*³⁷ or *immediate descendant* which automatically inherits all of the behavior and properties of the original class, (now called the *super-class*³⁸ or *immediate ancestor*); we then take this sub-class and add any additional behavior (through new messages and methods) and any additional properties (through new instance variables) that make this sub-class different from its super-class. For example, once we have defined the "rectangle" class, we might realize that a square is simply a special type of rectangle. We could then simply define the "square" class to be a sub-class of the "rectangle" class. The "square" class would inherit all the instance variables, messages, and methods of the "rectangle" class. The "square" class would already "know" how to respond to the messages for draw, move, and rotate. At this point, we would also add the features that make a square different from a rectangle.

Object-oriented languages provide two different ways to differentiate the behavior of a sub-class from its super-class. First, we could simply add a new message and a corresponding method to the definition of the sub-class. For example, we might add a message to our definition of the "square" sub-class that automatically draws the largest circle that just fits inside the square and touches each side exactly once. The super-class rectangle could not have had this message because it's geometrically impossible for a circle within a rectangle to touch each side exactly once. But, the square is a special kind of rectangle,³⁹ and this message could be performed on a square.

In addition, object-oriented languages allow a sub-class to *override* methods inherited from the super-class. For example, we would want our square to respond to the message "calculate area" just as we wanted our rectangle to respond to that message. Normally, the sub-class square inherits both the message and the method which contains the actual programming code that tells the object how to respond to that message. In the case of the rectangle, the method for "calculate area" multiplied the rectangle's length by its width. However, we know that the square is a special kind of rectangle whose length is equal to its width. Thus, we

36. Inheritance is also a more powerful tool for achieving meaningful abstractions: Inheritance is the most promising concept we have to help us realize the goal of constructing software systems from reusable parts, rather than hand coding every system from scratch. Procedural abstraction has worked well for some select domains, such as mathematical libraries, but the unit of abstraction is too small, the procedural focus not general enough, and the parameter mechanism too rigid.

Korson, *supra* note 11, at 43.

37. C++ uses the term "derived class." STROUSTRUP, *supra* note 33, at 30.

38. C++ uses the term "base class." *Id.*

39. The "is a special kind of" terminology is a particularly useful way to think of the relationship between sub-classes and super-classes.

might want to override the method for "calculate area" with a new method which calculates the area by simply squaring the value of any side.⁴⁰ When the program sends a "calculate area" message to an object of the "square" class, the object will calculate the area by squaring one side rather than using the method of "length times width" that it initially inherited. The ability to override methods provides an important tool for customizing the behavior of sub-classes and for taking advantage of efficiencies that might be available in sub-classes that cannot be used in the more generic super-classes.

This inheritance concept can be easily applied to the structure of the QuadWorld application. At the highest level, we must define a *root class* which is a class that has no super-classes above it. In this case, we might define a root class, called "quadrilateral," which defines only the most generic properties and behaviors common to all four-sided geometric figures. After that, we could construct the simple inheritance chart that is shown in Figure 1.⁴¹

40. Of course, this example is fairly trivial and would not provide any efficiency improvement, but it nonetheless shows how sub-classes can customize particular inherited behavior. A more useful example would arise when we originally define the rectangle class. While the rectangle class would inherit the methods of the parallelogram class, the formula for calculating the area of a parallelogram is considerably more involved than the simple "length X width" formula that can be used for the rectangle class.

41. This figure is adapted directly from SCHMUCKER, *supra* note 25, at 22.

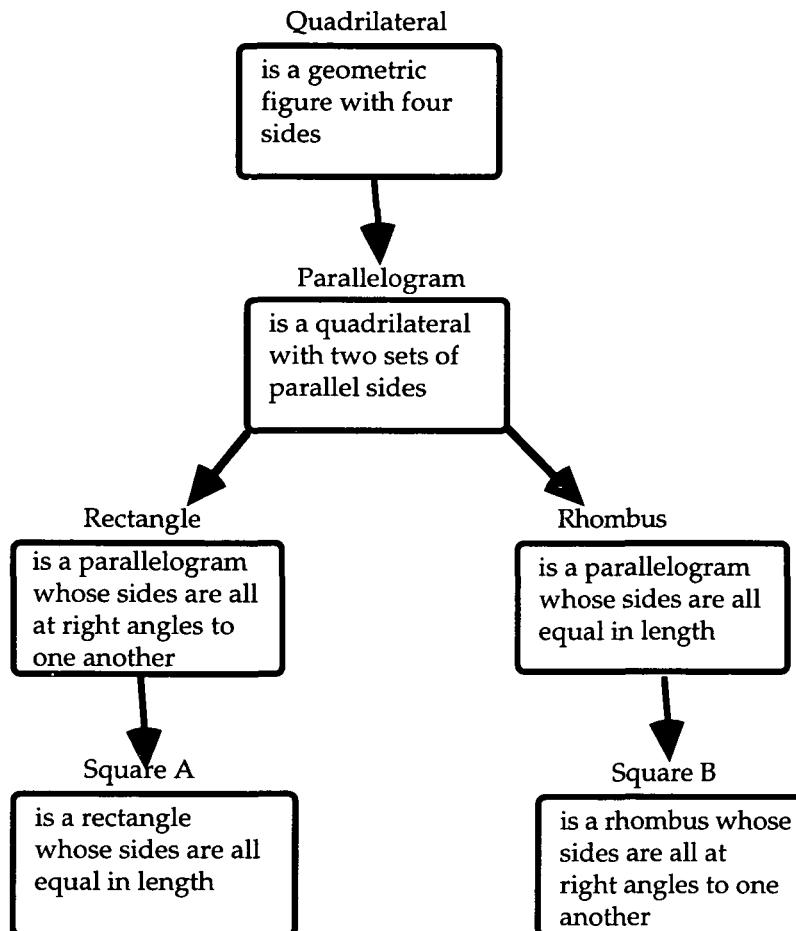


Figure 1: Simple Inheritance Chart for Quad World

A simple inheritance chart thus provides a logical method for organizing the basic relationships among various objects. Moreover, it forces us to recognize common features among different objects and then allows us to take pre-existing objects and modify them to fit our needs by building sub-classes from them. Moreover, "[I]nheritance not only supports reuse across systems, but it directly facilitates extensibility within a given system . . . [I]nheritance minimizes the amount of new code needed when adding additional features . . . and minimizes the amount of new code that must be changed when extending a system."⁴²

42. Korson & McGregor, *supra* note 11, at 43.

Real world behavior does not always fit this simple hierarchical structure, however. The object-oriented model can account for more complex relationships between classes with the concept of *multiple inheritance*.⁴³ Looking at Figure 1, we notice that a square has properties that make it a special kind of rectangle and properties that simultaneously make it a special kind of rhombus. Rather than having two classes of squares as in figure 1, we can improve our model by having a single class, called "square" that inherits from both the rectangle and the rhombus classes. The "square" class will inherit the ability to build a parallelogram with right angles from the rectangle class and will simultaneously inherit the ability to build a parallelogram with equal sides from the rhombus class.⁴⁴ As a result, we can create the "square" class without having to write any new methods, since all of the square's behavior is based on the combined properties of the "rectangle" and "rhombus" classes.⁴⁵ The improved model for QuadWorld is shown in Figure 2.⁴⁶

43. Since multiple inheritance is more difficult for the compiler to handle than simple inheritance, not all object-oriented development packages support multiple inheritance. For example, Symantec's Think C does not support it, while a full implementation of C++ would support it. See THINK C OBJECT-ORIENTED PROGRAMMING MANUAL, *supra* note 33, at 62.

44. SCHMUCKER, *supra* note 25, at 277.

45. If any messages are defined in both the rectangle class and the rhombus, then we must have some way to specify whether we want the square to inherit the rectangle's method for that message or the rhombus' methods. Multiple inheritance allows us to specify either the rectangle or the rhombus as the *primary immediate ancestor class*. In any inheritance conflict, the square will inherit the methods of the primary immediate ancestor class. *Id.* at 278.

46. Adapted from *id.* at 277.

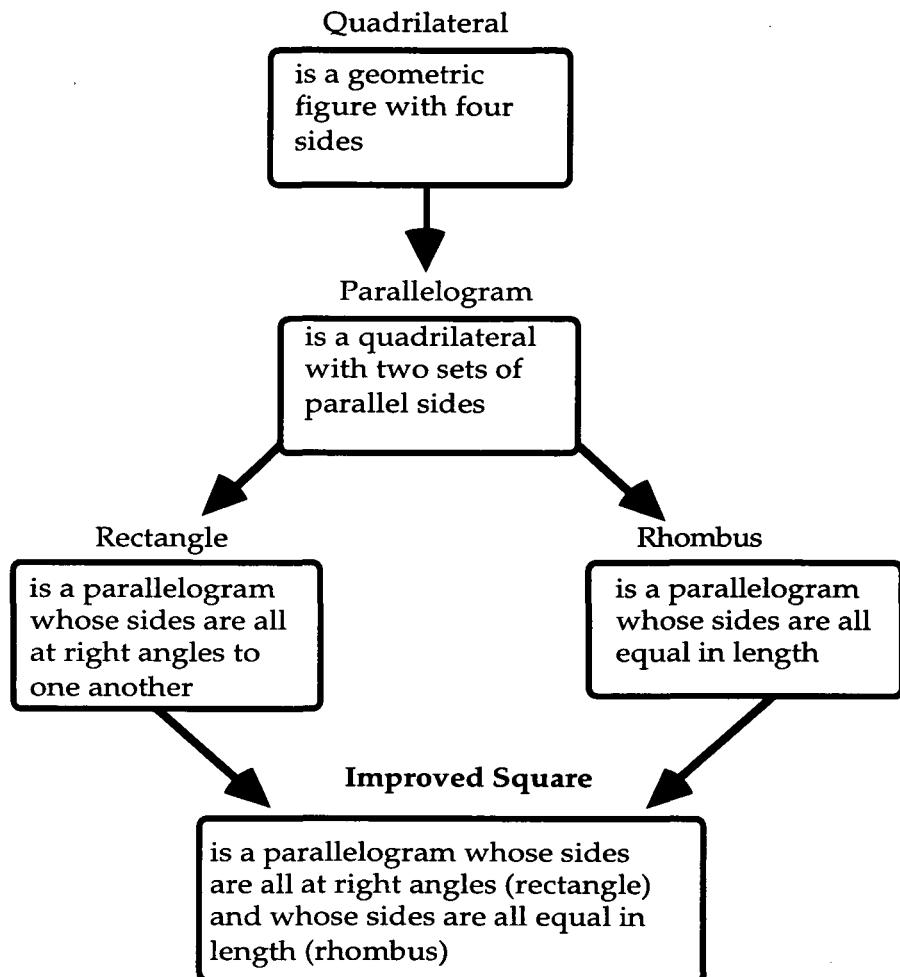


Figure 2: Improving Quad World with Multiple Inheritance

Multiple inheritance is a powerful concept that allows the programmer to model real-world entities which blend the characteristics of two or more super-classes. At the most basic level, multiple inheritance allows us to build classes as simple combinations of existing classes. For example, suppose we were working on an existing system that is designed to track the progress of efforts to save endangered wildlife, and we were instructed to add a class for "leopards."⁴⁷ Suppose further that the system already contained classes for "endangered" and "wild cat." Since leopards are both wild cats and endangered, we would

47. This example is taken directly from Korson & McGregor, *supra* note 11, at 58.

start by defining it as a sub-class of both of these existing classes. In addition, we can use multiple inheritance to model the behavior of entities that *blend* the characteristics of other objects. For example, a "houseboat" is not literally the combination of a house and a boat. Nonetheless, houseboats do possess some of the characteristics of a house and some of the characteristics of a boat. Again, a good starting point for building the "houseboat" class would be to define the class as a sub-class of both a "house" class and a "boat" class.⁴⁸ We could then add the characteristics that make a houseboat something more than the literal combination of a house and a boat by adding new messages and methods and by overriding the behavior of houses and boats that don't really apply to houseboats.⁴⁹

3. POLYMORPHISM AND DYNAMIC BINDING

In conventional programming languages, each variable has a *static type* which is defined when the program is written and remains unchanged while the program is running. In the object-oriented model, each object is defined as belonging to a particular class when the program is written; however, when the program is actually running, objects are not bound to their original class and instead may be treated as if they belong to any sub-class of the original class. Since any object can be treated as belonging to one class at one moment and as belonging to a different class at a later moment, object-oriented programming languages are said to allow for *dynamic typing*.

Polymorphism is simply a more general description of the concept of dynamic typing. The basic idea behind polymorphism is that if "Y inherits from X, [then] Y is an X, and therefore anywhere that an instance of X is expected, an instance of Y is allowed."⁵⁰ For example, consider the QuadWorld program again. Suppose the user had drawn a number of different quadrilaterals on the screen and we wished to display a textual list of all the different types of objects the user had already drawn. We might want to display a message to the user that reads, "you have drawn two squares, three parallelograms and one rectangle." One convenient way for the program to keep track of this information would be to create an object which keeps track of all the quadrilaterals currently displayed on the screen. Here's one way we could define that object:

48. SCHMUCKER, *supra* note 25, at 276.

49. Since houseboats probably have more in common with boats than houses, we would probably use the boat class as the primary immediate ancestor class. *Id.* at 278.

50. Korson & McGregor, *supra* note 11, at 45.

Class Definition of "Current Screen":

Internal Data:

- linked list⁵¹ of square objects currently displayed
- linked list of rectangle objects currently displayed
- linked list of parallelogram objects currently displayed
- linked list of rhombus objects currently displayed
- linked list of quadrilateral objects (not falling into any of the above categories) currently displayed

Messages that the object is able to perform:

- Add a square to the square list, delete a square from the square list
- Add a rectangle to the rectangle list, delete a rectangle
- Add a parallelogram to the parallelogram list, delete a parallelogram
- ... (similar messages for the other shapes)

Polymorphism allows us to find a better way to define this object. First, we note that every shape is a sub-class of the class quadrilateral. That means that wherever we have a data structure (such as a linked list) or a method that expects to use a quadrilateral, it will also accept any sub-class of the quadrilateral class. As a result, we can drastically reduce the complexity of the class "current screen" by redefining it as follows:

Class Definition of "Current Screen" Using Polymorphism:

Internal Data:

- linked list of quadrilateral objects currently displayed

Messages that the object is able to perform:

- Add a quadrilateral to the list, delete a quadrilateral from the list

When the program is actually running, we know that each element will be some type of quadrilateral, but we don't know which elements will contain which type of quadrilateral until the user has drawn some shapes on the screen. At any given moment, each element in the list will have a dynamic type which can be a square, rectangle, parallelogram, rhombus, or quadrilateral depending on which shapes are currently displayed on the screen. For example, if the user has drawn three objects on the screen, a square, a rectangle, and a quadrilateral, the "current screen" object would then contain the linked list shown in Table 1.

51. A linked list is simply a data structure that allows us to store pieces of data in a sequential chain. Each distinct piece of data in the list is called an "element" in the list. In the object-oriented world, each element can be an object.

Element #:	1	2	3
Original Definition:	Quadrilateral	Quadrilateral	Quadrilateral
Dynamic Type:	Square	Rectangle	Quadrilateral

Table 1: Illustration Of Dynamic Typing In A Linked List

Dynamic Binding is closely associated with the idea of polymorphism and dynamic typing. Dynamic binding builds on these concepts by allowing a particular object to respond differently to a particular message depending upon its dynamic type at a given moment while the program is running.⁵² For example, assume that our program accepts a command from the user to display the area of each shape currently on the screen. Dynamic binding makes this feature easy to

Element #:	1	2	3
Original Definition:	Quadrilateral	Quadrilateral	Quadrilateral
Dynamic Type:	Square	Rectangle	Quadrilateral
Class Method Invoked In Response To Message	Square	Rectangle	Quadrilateral
Formula Used In Response To The Message: "Calculate Area"	square of the length	length X width	generic formula applicable to any four-sided shape

Table 2: Illustration Of Dynamic Binding In A Linked List

implement and automatically incorporates the special methods we wrote to take advantage of the fact that there is a simpler formula to calculate the area of a rectangle than the area of a generic quadrilateral.

To implement this feature, the program responds to the user's request to calculate the area of each displayed shape by sending the "calculate area" message to each object in the linked list stored in the "current screen" object. When each object in the list receives the "calculate area" message, it will use the method associated with its dynamic type rather than the method associated with its originally defined class. Thus, the linked list will respond as shown in Table 2.

52. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 63. ("Static binding means that the types of all variables and expressions are fixed at the time of compilation; *dynamic binding* (also called *late binding*) means that the types of all variables and expressions are not known until runtime.").

Dynamic binding and polymorphism provide several immediate advantages. First, they encourage a high degree of generalization by permitting the programmer to write procedures that apply to any quadrilateral, but that respond with the optimal code depending on whether the particular quadrilateral is a square, rectangle, or parallelogram. In the object-oriented world, we simply send the "calculate area" message to some unknown quadrilateral and it doesn't matter that we have no way of knowing what type of quadrilateral will actually receive that message when the program is running. In contrast, traditional programming would require us to add code that essentially said "If the particular quadrilateral is a square then use the procedure for squares, but if the particular quadrilateral is a rectangle then use the rectangle procedure, but if the particular quadrilateral is a parallelogram, then use the parallelogram procedure, . . . otherwise use the generic quadrilateral procedure."⁵³

Moreover, dynamic binding and polymorphism also promote reusability by allowing other programmers to create new sub-classes of the quadrilateral class and know that they will automatically work in any procedure that expects to use the quadrilateral class. The programmers also know that if they have written new methods in the sub-class that override methods in the quadrilateral class the new methods will automatically be used when any procedure sends a message to their objects.

53. Of course traditional programming languages have a shorthand expression for this problem. In Pascal, the programmer would use a "case" statement that lists the names of different procedures for the different quadrilaterals, and in C the programmer would use a "switch" statement that listed the procedures. Nonetheless, these statements are no more than shorthand expressions for the long quotation in the text.

For those readers familiar with Pascal or C, consider a procedure that must re-draw a screen filled with various quadrilaterals. We could implement this procedure in Object Pascal by the following piece of code:

```
for i:= 1 to Number_of_Shapes do  
  current_figure.item(i).draw;
```

{current_figure is an array
of quadrilateral objects, and
draw is a message in each of
the quadrilateral subclasses
that tells the object to draw
itself}

Korson & McGregor, *supra* note 11, at 46. "At each pass through the loop, the code matching the dynamic type of current_figure.item(i) will be called. Note that if additional kinds of shapes are added to the system, this code segment remains unchanged. Contrast the resulting simplicity and extensibility as compared with a traditional case statement design." *Id.*

C. The Process of Designing Software Under the Object-Oriented Model

Given the concepts of objects, classes, inheritance, polymorphism, and dynamic binding, we can formulate an analytical approach for writing software that takes advantage of the object-oriented model. There have been many formal attempts to define an "object-oriented approach" to software design,⁵⁴ this Section outlines the basic features common to most of these models. In reading this Section, compare this model to the traditional model of software design as understood by the *Whelan* court and Nimmer.

1. IDENTIFY THE OBJECTS AND CLASSES THAT COMPRIZE THE "PROBLEM DOMAIN"

The first step in approaching object-oriented design is to learn as much as possible about the problem that the program is supposed to solve ("the problem domain"). As a first approximation, the programmer should approach the problem not as a computer scientist but rather by becoming an expert in a specific domain. For example, a programmer writing a navigational system for an airplane should initially learn from pilots, air controllers, and aeronautical engineers how navigation works:

Essentially, the developer must act as an abstractionist. By studying the problem's requirements and/or by engaging in discussions with domain experts, the developer must learn the vocabulary of the problem domain. The tangible things in the problem domain, the roles they play, and the events that may occur form the candidate classes and objects of our design, at its highest level of abstraction.⁵⁵

Once the developer learns the vocabulary and physical items used by pilots, air controllers, and aeronautical engineers, the programmer can begin to identify specific classes that will be needed to write navigational software. At this stage, the programmer is examining the problem domain from a fairly high level of abstraction.⁵⁶ Various commentators have identified formal categories that may help suggest candidate classes as illustrated in Table 3.

54. See, e.g., BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4; Henderson-Sellers & Edwards, *supra* note 9; Korson & McGregor, *supra* note 11; Ronald J. Norman, *Object-oriented Systems Analysis: A Methodology for the 1990s*, J. Sys. MGMT., July 1991, at 32; Rebecca J. Wirfs-Brock & Ralph E. Johnson, *Surveying Current Research in Object-Oriented Design*, COMM. ACM, Sept. 1990, at 104.

55. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 191.

56. Norman, *supra* note 54, at 33 ("It is not necessary and certainly not required that all possible objects be identified during this step. Only the most intuitive and obvious ones may be identified here, while others or refinements of these may be identified during a later step.").

Physical Items	planes, wings, engines, fuel pump, radio beacon
Roles	pilot, copilot, navigator, passenger
Events	landing, take-off, turning, putting down landing gear
Interactions	clearance from air controller, radio contact, schedules, connections with other planes
Places	airport, destination, origin

Table 3: Types Of Classes That Are Likely To Be Used⁵⁷

This approach has several advantages over traditional software design. First, rather than asking "what tasks must the program perform," object-oriented design asks "how would those who will be relying on this program describe their problem, and what would *they* identify as the major actors (both human and inanimate) in the problem domain." This direct focus on the problem domain forces the programmer to address the specific needs of users in the problem domain before writing any code. In contrast, the traditional programming model promotes an early emphasis on the "tasks" that the software must perform and thereby removes the focus from the problem domain. At an early stage of the design process, the traditional programmer becomes bound to the specific instructions that will be used to write the program, often before potential users have identified all of their requirements. Second, the object-oriented approach helps to reveal commonalities that may exist across similar applications (vertical domain analysis) as well as commonalities that can be reused in different parts of the same application (horizontal domain analysis):

For example, when starting to design a new patient-monitoring system, it is reasonable to survey the architecture of existing systems to understand what key abstractions and mechanisms were previously employed and to evaluate which were useful and which were not. Similarly, an accounting system must provide many different kinds of reports. By treating these reports as a single domain, a domain analysis can lead the developer to an understanding of the key abstractions and mechanisms that serve all the different kinds of reports. The resulting classes and objects reflect a set of key abstractions and mechanisms generalized to the immediate report-generation problem; therefore, the resulting design is likely to be simpler than if each report had been analyzed and designed separately.⁵⁸

57. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 141 (summarizing categories proposed by Shlaer, Mellor, Ross, Coad, and Yourdon); see also Norman, *supra* note 54, at 40, Table 3 (identifying categories as "tangible items," "roles played by people or organizations," "incidents which happen at a specific point in time," "interaction [sic] that have a transaction-like quality," and "specification [sic] that have table-like qualities such as sales offices, state codes, standard industry codes, [and] tax rates").

58. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 142-43.

The object-oriented programmer's first written output is apt to be a rough list of classes and objects whose names imply their basic role in the problem domain and which will be used as the "common vocabulary of discourse among the developers."⁵⁹ Most important, these classes and objects should be subject to continual revision as the programmer follows the other three steps, thus leading to iterative and evolutionary changes in the original model of the problem domain, or what some commentators have called "Round-Trip Gestalt Design."⁶⁰ Steps 2, 3, and 4 in the model are all explicitly designed to foster such reevaluation of the problem domain.

In contrast, the traditional programmer's first written task using top-down design is to produce a rough flowchart of the program, which, by its very nature, is farther removed from the problem domain, closer to the stage of writing actual software code, and more likely to lock the programmer into tight dependencies among different parts of the program. The traditional programmer's tendency is to then parcel out pieces of the project to different programmers based on the original flowchart. While nothing stops the design team from refining the flowchart later, nothing in the traditional top-down model encourages iterative or evolutionary changes in the basic flowchart. In fact, the risk that changes in one part of the program will ripple through all other parts of the program actively discourages such changes.⁶¹

59. *Id.* at 192 (noting also that "[i]n most cases, this step takes a small amount of time relative to the other three steps. Often, a single chief designer will draft a list of candidate classes and objects and then review this list with peers as a kind of sanity check." *Id.* at 191-92).

60. *Id.* at 188 ("This style of design emphasizes the incremental and iterative development of a system through the refinement of different yet consistent logical and physical views of the system as a whole."); see also Henderson-Sellers & Edwards, *supra* note 9, at 148 ("Both top-down analysis and bottom-up class design, seen as the hardest part of the entire object-oriented software life cycle, must therefore be either concurrent or, at least iterative.") (footnote omitted).

61. BOOCHE, OBJECT-ORIENTED DESIGN, *supra* note 4, at 188. The entire "top-down" vs. "bottom-up" approach to design has been the subject of significant debate within the software community. It is important to recognize that object-oriented design is neither "top-down" nor "bottom-up":

Assume that we are faced with the problem of staffing an organization to design and implement a fairly complex piece of computer hardware. We might use horizontal staffing, in which we have a waterfall progression of products, with systems architects feeding logic designers feeding circuit designers. This is an example of top-down design, and requires designers who are "tall skinny men," as Druke calls them, because of the narrow yet deep skills that each must possess. Alternately, we might use vertical staffing, in which we have good all-around designers who take slices of the entire project, from architectural conception through circuit design. The skills that these designers must have leads Druke to call them "short fat men." Unfortunately, given its inherent complexity, software development often demands that we employ "tall fat people."

2. IDENTIFY THE STRUCTURE AND SEMANTICS OF THE OBJECTS AND CLASSES

At this stage, the programmer must specify the behavior and properties that each object will possess. One possible approach is to write "a script for each object, which defines its life cycle from creation to destruction, including its characteristic behaviors."⁶² For example, once we have identified a "radio beacon" object in our navigational software, we might write a script that reads "beacon object is created when a plane is close enough to receive the signal from that beacon, and then the beacon is expected to send a radio signal at a pre-set frequency and at pre-set intervals, and then the beacon is expected to continue this behavior until the plane passes the beacon and leaves the beacon's range." Similarly, we might define a "landing gear" object which is created as soon as the software is running and is expected to be able to keep track of whether the landing gear is up or down, and send an alarm message if the landing gear is in the wrong position. The process of writing these scripts should also force the programmer to reevaluate the original list of objects and classes identified in step 1. For example, when we define the landing gear as being able to send an alarm message, we then realize that we never identified the need for an "alarm bell" object that would receive the alarm message and be used to display an alarm message on the navigator's computer screen. In this way, step 2 is iterative because it forces the programmer to reevaluate the decisions made in step 1.

In addition, identifying the behavior of each object may reveal other sub-classes that could be introduced. For example, an analysis of a bookkeeping program in step 1 might reveal the need for an "invoice" object. However, once we specify the attributes of an invoice in this step, we might realize the need for additional objects, such as "header," "account summary," and "list of transactions," that represent the different sections that make up the invoice.⁶³ Conversely, this analysis

Id.; see also Henderson-Sellers & Edwards, *supra* note 9, at 146 (noting that "object-oriented (OO) design and analysis has many attributes of both top-down and, perhaps predominately, bottom-up design. Since one of the aims of an OO implementation is the development of generic classes for storage in libraries, an approach which considers both top-down analysis and bottom-up design simultaneously is likely to lead to the most robust software systems.").

62. BOOCHE, *OBJECT-ORIENTED DESIGN*, *supra* note 4, at 192 (noting that "[t]his step is much harder than the first and takes much longer. This is the phase in which there may be fierce debates, wailing and gnashing of teeth, and general name-calling during design reviews. Finding classes and objects is the easy part; deciding upon the protocol of each object is hard").

63. Norman, *supra* note 54, at 33 (suggesting that the programmer look for whole-to-part relationships and generalization-to-specialization relationships at this point). While this activity may blur some of the distinctions between step 2 and step 3, such blurring of

could also reveal the need to redefine some of the super-classes. For example:

[A] class of 'bird' with an attribute that "birds can fly" is successful until we consider the Southern Hemisphere and "penguins," "ostriches," "kiwis" etc. In this case, one solution is to introduce an additional level in the inheritance hierarchy by introducing two children classes of class bird as "flying bird" and "non-flying bird" and redefining the parent class to remove the attributes relating to flight. This process tries to develop a logical hierarchy of objects so there are no "missing" objects.⁶⁴

3. IDENTIFY THE RELATIONSHIPS AMONG OBJECTS AND CLASSES

In step 3, the programmer must identify the relationships among the previously identified objects and classes. First, the programmer must develop the inheritance relationships and define the structure of super-classes and sub-classes. In doing so, the programmer is likely to uncover additional "patterns among classes, which cause us to reorganize and simplify the system's class structure, and patterns among cooperative collections of objects, which lead us to generalize the mechanisms already embodied in the design."⁶⁵

One way to formulate a concrete representation of these relationships is by building a *semantic data model*.⁶⁶ For example, if we were designing a program that was intended to control the traffic lights at a busy intersection, we might construct the partial semantic data model shown in Figure 3 in which rectangles denote classes and circles denote the functional relationships between classes connected by arrows.

specific steps is consistent with the iterative and evolutionary style of object-oriented design.

64. Henderson-Sellers, *supra* note 9, at 150 (footnotes omitted).

65. BOOCH, OBJECT-ORIENTED DESIGN, *supra* note 4, at 193.

66. Korson, *supra* note 11, at 47.

Using this data model, we can then determine what messages each object must accept. One useful conceptual device is to consider each message as a "service" that the object is capable of providing to any other object. Then, using the scripts from step 2 to determine what behaviors each object will exhibit, the programmer can determine what kind of services each object will need to fulfill its role.⁶⁸ As would be expected from the theme of iterative development, this process is likely to reveal that certain objects require services that no current object yet provides. The programmer must then either add that service to an existing object or create a new object to handle that service. As an alternative, the relationship between two objects can be considered to be a contract in

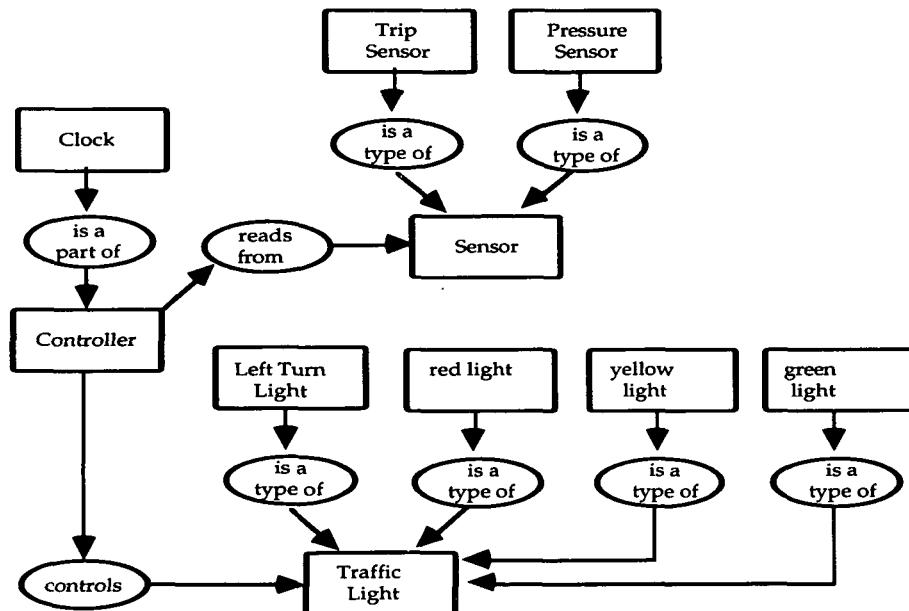


Figure 3: Semantic Data Model for Controlling Traffic Lights⁶⁷

which one object is a "client" that requests certain services from another object which is a "server" and fulfills those requests.⁶⁹ Again, the programmer must make sure that every object that needs a contract for a particular service has a corresponding server object to fulfill that contract.

67. This model is adapted from Korson & McGregor, *supra* note 11, at 48 (fig. 8).

68. Henderson-Sellers & Edwards, *supra* note 9, at 150.

69. Wirfs-Brock & Johnson, *supra* note 54, at 110-11.

4. IDENTIFY THE PUBLIC INTERFACES AND SERVICES PROVIDED BY EACH OBJECT AND CLASS

At this point, we know what role each object plays and what services the object provides to other objects. Using that information, we can define the general type of data structures needed by each object and the methods that the object will need to provide services to other objects. As the internal structure of a particular object is developed, the programmer may discover that this object can be built by using pre-existing libraries of more primitive objects.⁷⁰ At the end of this stage, the programmer can begin writing the actual source code for each method, perhaps treating each method as a miniature program that can be approached using traditional procedural techniques.

IV. COPYRIGHT PROTECTION FOR OBJECT-ORIENTED SOFTWARE

Before examining the scope of copyright protection, it is important to recognize when the fact that the object-oriented model was used to design a piece of software matters and when it does not. If a programmer writes software under the object-oriented model, the programmer will use an object-oriented programming language to write the high level source code for the program. This source code is then translated by a compiler program into object code which is a series of 1's and 0's. These 1's and 0's represent low level commands which the microprocessor can understand and execute. At the same time, the compiler can also produce an assembly language version of the source code. Assembly language is a human-readable listing of object code in which each low-level microprocessor instruction is represented by a single word, such as "jump," "store," or "link." However, once the program has been translated into object code or assembly language, the fact that the original source code was written in an object-oriented programming language is virtually impossible to detect. The microprocessor itself has no concept of object-oriented principles;⁷¹ therefore, the compiler produces a program that in object code form is indistinguishable from a program written according to traditional design methods.

As a result, in a case where the plaintiff alleges that the defendant copied the object code or assembler versions of the program, which can

70. Henderson-Sellers & Edwards, *supra* note 9, at 150.

71. While some computer manufacturers are touting "object-oriented operating systems," this statement does not mean that the microprocessor itself understands object-oriented principles. Instead, this feature means that the operating system is designed so that a program written in high level source code can interact with the operating system by using a pre-defined library of objects that perform the functions of the operating system. For example, the operating system running on a machine with a graphical user interface may supply libraries of objects for windows, icons, and menus.

arise only when verbatim copying of the program has occurred,⁷² the court can proceed without worrying about principles of object-oriented design. But, object-oriented principles are absolutely critical when the plaintiff alleges that the defendant had access to the original source code and copied it. During the trial, both plaintiff and defendant will have to produce the source code for their programs, and the court will have to determine whether the programs are "substantially similar."⁷³ In weighing the evidence of similarity, the court will need to understand how principles of object-oriented design affect that comparison.

The starting point for any discussion on copyright protection for software is 17 U.S.C. § 102(b), which excludes protection for "any idea, procedure, process, system, [or] method of operation."⁷⁴ While the theoretical limits imposed by § 102(b) seem clear, courts and commentators have struggled with the practical application of the section to computer software. This Section will review the two dominant approaches to § 102(b) and determine how they apply to object-oriented software. In addition, this Section analyzes an alternative approach which advocates an ad hoc balancing of the economic effects of protecting the plaintiff's work, but concludes that courts will not adopt this approach because it has no support in copyright doctrine and it is problematic as a matter of innovation policy.⁷⁵ Finally, this Section concludes that since the "behavioral" aspects of software are particularly dominant when the object-oriented model is used to write software, pure copyright doctrine provides almost no protection for the *object-oriented* aspects of software.

72. While it is theoretically possible to work backwards from the assembler version of the source code, this process is virtually impossible for a program of any complexity. Particularly as software projects are increasingly characterized by "programming-in-the-colossal," the assertion that a defendant could have reproduced the plaintiff's source code by disassembly is ludicrous. Thus, a defendant will find copying the object code useful only if the defendant can sell verbatim copies of the plaintiff's program, perhaps in a foreign market where explicit piracy is tolerated. For a discussion of the technical difficulties involved in disassembly, see G. Gervaise Davis III, *Reverse Engineering and the Computer Industry: A Battle Between Legal and Economic Principles* (1991) (unpublished presentation on file with the *High Technology Law Journal*); Ronald S. Laurie, *Protection of Trade Secrets in Object Form Software: The Case for Reverse Engineering*, COMPUTER LAW., July 1984, at 1. But cf. Allen R. Grogan, *Decompilation and Disassembly: Undoing Software Protection*, COMPUTER LAW., Feb. 1984, at 1 (arguing that disassemblers and decompilers may allow object code to be converted so that much of the logic of the program is revealed).

73. See NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.26 ("In many software cases, access is either conceded or easily proved, so that a finding of infringement turns entirely on whether the works are substantially similar.").

74. 17 U.S.C. § 102(b) (1988).

75. The term "innovation policy" is used to denote the best mix of legal incentives that would maximize the total value of new software inventions.

A. *Whelan* and Its Progeny

Whelan Associates v. Jaslow Dental Laboratory represents the broadest approach to protecting computer software. In *Whelan*, the court examined the idea/expression dichotomy⁷⁶ stated in § 102 and concluded that "the purpose or function of a utilitarian work would be the work's idea, and everything that is not necessary to that purpose or function would be part of the expression of the idea."⁷⁷ Since this rule forces the court to focus on *one* idea behind the program, courts applying this test will necessarily define an idea which represents an extremely high level of abstraction. For example, the *Whelan* court characterized the idea behind the plaintiff's program as the efficient operation of a dental laboratory.⁷⁸ At this high level of abstraction, there are of course many ways to write a program which performs that general function and all program elements at lower levels of abstraction would constitute copyrightable expression. As a result, the *Whelan* test protects the "structure, sequence, and organization" of source code as a general rule.

76. The limitations expressed in § 102(b) create what is known as the "idea/expression dichotomy" in copyright law:

The crucial consideration in the analysis that follows is that copyright law protects only an author's original expression, not ideas or elements taken from preexisting works. Infringement is shown by a substantial similarity of *protectable expression*, not just an overall similarity between the works. Thus, before evaluating substantial similarity, it is necessary to eliminate from consideration those elements of a program that are not protected by copyright.

NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.28 to .29.

77. *Whelan Assocs. v. Jaslow Dental Lab.*, 797 F.2d 1222, 1236 (3d Cir. 1986), *cert. denied*, 479 U.S. 1031 (1987).

78. *Id.* at 1238 n.34. Other cases applying the *Whelan* test have provided wide-ranging software protection either by broadly defining the "idea" behind the plaintiff's program or by finding that "other ways" exist to express a particular idea. See, e.g., *Johnson Controls v. Phoenix Control Sys.*, 886 F.2d 1173 (9th Cir. 1989) (general finding that "the structure of the JC-5000S [plaintiff's entire program] is expression, rather than an idea in itself," apparently because "each individual application is customized to the needs of the purchaser. This practice of adaptation is one indication that there may be room for individualized expression in the accomplishment of common functions."); *Lotus Dev. Corp. v. Paperback Software Int'l*, 740 F. Supp. 37, 65-66 (D. Mass. 1990) (repeatedly asking whether there were other ways to express the idea of "an electronic spreadsheet"); *Pearl Sys. v. Competition Elec.*, 8 U.S.P.Q.2d (BNA) 1520, 1524 (S.D. Fla. 1988) (defining the idea behind two subroutines as providing "a method for the user to set a par time" and as allowing "the user to review the shots he or she has fired and to learn of the time that elapsed between each shot"); *Digital Communications v. Softklone Distrib.*, 659 F. Supp. 449, 459 (N.D. Ga. 1987) ("The use of a screen to reflect the status of the program is an 'idea'; the use of a command driven program is an 'idea'; and the typing of two symbols to activate a specific command is an 'idea.' "); *Broderbund Software, Inc. v. Unison World*, 648 F. Supp. 1127, 1133 (N.D. Cal. 1986) (defining the idea of the plaintiff's program as "the creation of greeting cards, banner, posters and signs that contain infinitely variable combinations of text, graphics, and borders").

Courts applying *Whelan* to object-oriented software are likely to protect the basic inheritance relationships among objects. For example, the idea behind the QuadWorld program could be expressed as a program to allow for the efficient drawing of quadrilaterals on a computer screen. Since there are undoubtedly many ways to write such a program, the particular choice of classes, sub-classes, and messages is copyrightable expression. In fact, the court could argue that the idea behind QuadWorld could be achieved using traditional programming techniques, and since this would produce entirely different looking source code from the object-oriented version, that variation alone proves the necessary range of expression to justify copyright protection.⁷⁹

Moreover, the court could use our analysis of the design process to argue that high level inheritance relationships and class structures must be protected by copyright. In *Whelan*, the court justified the protection of structure, sequence, and organization in part on the basis that "among the more significant costs in computer programming are those attributable to developing the structure and logic of the program. The rule proposed here, which allows copyright protection beyond the literal computer code, would provide the proper incentive for programmers by protecting their most valuable efforts."⁸⁰ While the validity of this argument is highly doubtful in light of the Supreme Court's recent decision in *Feist*,⁸¹ lower courts may still be tempted to protect those parts of the plaintiff's software that are the products of significant time and effort. Under our analysis of the object-oriented design process, steps 2 and 3, identifying the structure and semantics of the objects and classes and identifying the relationships among these objects and classes, represent the most difficult parts of the design process.⁸² The decisions made in those steps are critical to the quality of the final program.⁸³ Thus, a court could use this

79. This argument is arguably analogous to one proposed in *Lotus*. The *Lotus* court held that the differences between the user interface for Microsoft Excel on the Macintosh and the user interface for the Lotus program were evidence that there are multiple ways to express the idea of an electronic spreadsheet. *Lotus*, 740 F. Supp. at 65-66. The court was oblivious to the fact that Excel's user interface was entirely attributable to the Macintosh operating system (all programs running on the Macintosh have that same interface) and had nothing to do with how Excel chose to express the idea behind an electronic spreadsheet.

80. *Whelan*, 797 F.2d at 1237.

81. *Feist Publications v. Rural Tel. Serv., Inc.*, 111 S. Ct. 1282, 1290 (1991) (repudiating "sweat of the brow" theories for copyright protection because "the primary objective of copyright is not to reward the labor of authors, but '[t]o promote the Progress of Science and useful Arts'").

82. See *supra* parts III.C.2, III.C.3.

83. While the iterative nature of object-oriented development encourages refinement of the decisions made in steps 2 and 3, the court will see only the final program and thus will not be able to determine which decisions were initially made during the first pass through the design process and which were added later by refining the decisions made in steps 2 and 3. As a result, the references to steps 2 and 3 in this discussion include all decisions

argument to protect the general inheritance relationships between classes, the detailed scripts for each object, and the collections of services each object is expected to provide.

The *Whelan* court's analysis of § 102(b) has been heavily criticized by commentators,⁸⁴ and for good reason. The primary criticism of *Whelan* has focused on *Whelan's* use of a *single* idea existing in each computer program:

The crucial flaw in [*Whelan's*] reasoning is that it assumes only one "idea," in copyright law terms, underlies any computer program, and that once a separable idea can be identified, everything else must be expression. All computer programs are intended to cause the computer to perform some function. The broad purpose that the program serves, be it managing a dental laboratory, automating a factory, or dispensing cash at a bank teller machine, is *an* idea. Other elements of the program's structure and design, however, may also constitute ideas for copyright purposes.⁸⁵

Similarly, in *Computer Associates v. Altai*, a district court adopted this criticism and then used the traditional model of programming to further reveal *Whelan's* flaws:

In the case at bar, Dr. Davis [court-appointed expert] pointed out further technical flaws in the *Whelan* analysis which render its reasoning inadequate. As he so convincingly demonstrated, a computer program is made up of sub-programs and sub-sub-programs, and so on. Each of those programs and sub-programs has at least one idea. Some of them could be separately copyrightable; but many of them are so standard or routine in the computer field as

that fall within the general subject matter of those steps whether or not they were actually made in those steps or at a later time.

84. See, e.g., NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.33 to .34; Richard A. Beutel, *Software Engineering Practices and the Idea/Expression Dichotomy: Can Structured Design Methodologies Define the Scope of Software Copyright*, 32 JURIMETRICS J. 1, 17-20 (1991); Nimmer et al., *supra* note 3, at 629-30, 639; Reback & Hayes, *supra* note 3, at 3-4. Several cases have also rejected *Whelan* or have recognized its existence but then implicitly failed to apply it. See *Computer Assocs. Int'l v. Altai, Inc.*, 982 F.2d 693, 706 (2d Cir. 1992) ("We think that *Whelan's* approach to separating idea from expression in computer programs relies too heavily on metaphysical distinctions and does not place enough emphasis on practical considerations."); *Sega Enter. v. Accolade, Inc.*, 977 F.2d 1510, 1524 (9th Cir. 1992) ("The *Whelan* rule, however, has been widely—and soundly—criticized as simplistic and overbroad."); *Plains Cotton Co-op Ass'n v. Goodpasture Computer Serv., Inc.*, 807 F.2d 1256, 1262 (5th Cir. 1987) (declining "to embrace *Whelan*"), cert. denied, 484 U.S. 821 (1987); *Computer Assocs. Int'l v. Altai, Inc.*, 775 F. Supp. 544, 558-59 (E.D.N.Y. 1991) (describing *Whelan* as setting "forth what now seems to be a simplistic test for similarity between computer programs"), aff'd in relevant part, 982 F.2d 693 (2d Cir. 1992); *Manufacturers Technologies, Inc. v. CAMS, Inc.*, 706 F. Supp. 984, 992 (D. Conn. 1989) (not explicitly rejecting *Whelan* but arguing that the *Broderbund* court's application of *Whelan* to screen displays, "overextended the scope of copyright protection applicable to those screen displays"); *Healthcare Affiliated Servs., Inc. v. Lippman*, 701 F. Supp. 1142 (W.D. Pa. 1988).

85. NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.33 to .34.

to be almost automatic statements or instructions written into a program.⁸⁶

Despite this criticism, *Whelan* has never been overruled and is still the starting point for most discussions of copyright protection for computer software.

B. The Filtering Approach⁸⁷

In contrast to *Whelan's* "one idea" approach, Nimmer starts with the "patterns of abstractions" test⁸⁸ and concludes that the court must apply a series of standard copyright doctrines to filter out unprotectable ideas at each level of abstraction. This test is easy to defend because each filter is closely tied to a specific copyright doctrine and thus forces the court to account for every theory that can limit the number of program elements entitled to protection. Nimmer proposes that the court apply four basic filters: abstract ideas, merger, *scenes a faire*, and public domain. While Nimmer's test was closely tied to the traditional model of software development, it can still be applied to object-oriented software by altering the relative importance of each filter.

1. ABSTRACT IDEAS

Nimmer's first filter revisits the basic problem of separating protectable ideas from non-protectable expression. In the context of traditional software, this filter provides a strong limit on copyright protection because the top-down approach to software development "provides natural divisions, which may correspond to the various levels of abstractions that the court seeks to identify and analyze."⁸⁹ In Nimmer's view, the court can divide the software into programs, sub-

86. *Computer Assocs.*, 775 F. Supp. at 559.

87. This approach was first developed in Nimmer et al., *supra* note 3, at 635-55, and is summarized in NIMMER & NIMMER, *supra* note 3, at § 13.03[F]. The Second Circuit has recently endorsed this approach to substantial similarity. *Computer Assocs.*, 982 F.2d at 706.

88. The test was first developed by Judge Learned Hand:

Upon any work, and especially upon a play, a great number of patterns of increasing generality will fit equally well, as more and more of the incident is left out. The last may be no more than the most general statement of what the play is about, and at times might consist only of its title; but there is a point in this series of abstractions where they are no longer protected, since otherwise the playwright could prevent the use of his "ideas," to which, apart from their expression, his property is never extended.

Nichols v. Universal Pictures Corp., 45 F.2d 119, 121 (2d Cir. 1930), *cert. denied*, 282 U.S. 902 (1931).

89. Nimmer et al., *supra* note 3, at 638 ("[T]he systematic method used to develop computer programs makes the abstractions test facially more applicable to computer software than other types of works. Traditional literary works are not created in such a consistently organized and orderly fashion.").

programs, and sub-sub-programs and then determine at which level the code passes from being an unprotectable idea to being protectable expression.

Two problems bar meaningful application of this filter to object-oriented software. First, Nimmer himself admitted that even in the context of structured top-down programming, the test is not easy to apply.⁹⁰ As one commentator complained, "simply to characterize the filter as eliminating 'abstract ideas' says very little about what is, and is not, an 'idea.' One man's 'abstract idea' may be another's protectable expression."⁹¹ Second, the iterative nature of object-oriented development prevents the court from finding easy lines to draw in determining what is a "level of abstraction." The process of "round-trip gestalt design" will tend to blur meaningful line drawing on the basis of the design process itself.

As an alternative, we could define the levels of abstraction by considering class lists, inheritance relationships, and the semantic data model to each be separate levels of abstraction. However, these lines create extremely broad categories which may encourage the court to find the same single idea behind each level of abstraction. For example, a court examining the traffic light problem might conclude that the idea behind the list of classes is the "efficient management of a traffic intersection." But, if the court then examines the inheritance structure and semantic data model, it seems that the idea at those levels of abstraction is also the efficient management of a traffic intersection. In the context of object-oriented software, this alternative leads courts back to the heavily criticized "one idea" approach of *Whelan*.

2. MERGER

The merger filter operates to exclude elements of the program that can only be expressed in one way.⁹² In the context of computer software, "merger issues may arise in somewhat unusual ways. Although theoretically many ways may exist to implement a particular idea, efficiency concerns can make one or two choices so compelling as to virtually eliminate any form of expression."⁹³ In this category, Nimmer lists such low-level routines as searching and sorting algorithms, which should not be protected, because "the fact that two programs both use the most efficient sorting or searching method available supports an inference of independent creation as readily as it supports one of copying, and thus

90. NIMMER & NIMMER, *supra* note 3, at 13-78.33.

91. Beutel, *supra* note 84, at 23.

92. NIMMER & NIMMER, *supra* note 3, at 13-78.35.

93. *Id.*

is not reliable evidence that copying occurred."⁹⁴ These considerations have finally received explicit judicial recognition by the Ninth Circuit which has expressed the impact of merger even more broadly than Nimmer:

To the extent that there are many possible ways of accomplishing a given task or fulfilling a particular market demand, the programmer's choice of program structure and design may be highly creative and idiosyncratic. However, computer programs are, in essence, utilitarian articles—articles that accomplish tasks. As such, they contain many logical structural, and visual display elements that are dictated by the function to be performed, by considerations of efficiency, or by external factors such as compatibility requirements and industry demands.⁹⁵

For most merger issues, object-oriented software can be analyzed in the same manner as traditional software. Sorting and searching routines would be used primarily by the *internal* implementation of a specific object's methods. Since this internal implementation may itself have been written using traditional structural programming techniques, courts should be able to apply this test without alteration. Similarly, other merger concerns, such as ensuring compatibility with particular hardware and software, should not raise issues unique to object-oriented software.⁹⁶ In general, courts should find the merger filter to be a powerful tool for limiting infringement claims relating to the internal implementations of specific objects.⁹⁷

94. *Id.* at 13-78.36. It must be emphasized that copyright law does not prevent a defendant from producing a substantially similar program, as long as the defendant did not actually copy the plaintiff's work. Copying is an absolute prerequisite for infringement, and the analysis of substantial similarity is used only to raise the inference of copying because direct evidence of copying rarely exists. See, e.g., *Computer Assocs.*, 982 F.2d at 708 ("Since, as we have already noted, there may be only a limited number of efficient implementations for any given task, it is quite possible that multiple programmers, working independently, will design the identical method employed in the allegedly infringing work. Of course, if this is the case, there is no copyright infringement.").

95. *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510, 1524 (9th Cir. 1992); see also *Computer Assocs.*, 982 F.2d at 708 ("[W]hen one considers the fact that programmers generally strive to create programs 'that meet the user's needs in the most efficient manner,' the applicability of the merger doctrine to computer programs becomes compelling. . . [T]he more efficient a set of modules are, the more closely they approximate the idea or process embodied in that particular aspect of the program's structure." (quoting Menell, *supra* note 3, at 1052; citation omitted)).

96. See *Sega*, 977 F.2d at 1526 (allowing intermediate copying in order to ensure compatibility with videogame hardware); *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832 (Fed. Cir. 1992) (same).

97. Merger analysis should not be used when evaluating semantic data models or the general structure of particular object classes and sub-classes. For most sophisticated and complex programs, it is highly unlikely that only one efficient object-oriented structure exists. The analysis of semantic data models is better addressed by the *scènes à faire* doctrine discussed in the next sub-section.

3. SCENES A FAIRE

Scenes a faire represents the most powerful filter for object-oriented software. Nimmer used the term to justify excluding program elements dictated by "external considerations," such as hardware standards, software standards, computer manufacturers' design standards, target industry practices, and computer industry programming practices.⁹⁸ While these considerations can certainly be applied to object-oriented software, traditional case law dealing with *scenes a faire* will actually be more important in eliminating elements of object-oriented software from copyright protection.

Under the *scenes a faire* doctrine, copyright protection is denied for "those elements that follow naturally from the work's theme rather than from the work's creativity."⁹⁹ In the literary context, *scenes a faire* has precluded protection for stock literary devices or stock character types that are inherent in the general theme of the work.¹⁰⁰ For example, in *Shaw v. Lindheim*,¹⁰¹ the court examined the mood, setting, and pace of the plaintiff's and defendant's television scripts and concluded that "[b]oth works are fast-paced, have ominous and cynical moods that are lightened by the [hero's] victory, and are set in large cities. These similarities are common to any action adventure series, however, and do not weigh heavily in our decision."¹⁰²

Particularly in the case of software designed to model real-world behavior, this approach to *scenes a faire* justifies excluding from protection software elements that are dictated by the real-world behavior being modeled. This understanding of the doctrine has already been accepted by several courts evaluating traditional software. For example, in *Data East USA, Inc. v. Epyx, Inc.*,¹⁰³ the Ninth Circuit analyzed two computer karate games and concluded that infringement could not be based on program elements that "encompass the idea of karate."¹⁰⁴ In doing so, the court approved of the district court's finding that:

[T]he visual depiction of karate matches is subject to the constraints inherent in the sport of karate itself. The number of combatants, the stance employed by the combatants, established and recognized moves and motions regularly employed in the sport of karate, the regulation of the match by at least one referee or judge, and the manner of scoring by points and half points are among the constraints inherent in the sport of karate. Because of these

98. NIMMER & NIMMER, *supra* note 3, § 13.03[F] at 13-78.36 to .43.

99. Nimmer et al., *supra* note 3, at 642.

100. See *Hoehling v. Universal City Studios, Inc.*, 618 F.2d 972 (2d Cir. 1980), *cert. denied*, 449 U.S. 841 (1980).

101. 919 F.2d 1353 (9th Cir. 1990).

102. *Id.* at 1363.

103. 862 F.2d 204 (9th Cir. 1988).

104. *Id.* at 209.

constraints, karate is not susceptible of a wholly fanciful presentation.¹⁰⁵

Similarly, in *Plains Cotton Co-op Ass'n v. Goodpasture Computer Serv., Inc.*,¹⁰⁶ the Fifth Circuit refused to find infringement because the "appellees presented evidence that many of the similarities between the GEMS and Telcot programs are dictated by the externalities of the cotton market."¹⁰⁷ As a result, the plaintiff could not claim protection for program elements that were designed to imitate a "cotton recap sheet," because that was a stock element in the real-world cotton market and necessary to any program trying to model that market.¹⁰⁸ Finally, in *Q-Co Industries, Inc. v. Hoffman*,¹⁰⁹ the court examined two tele-prompting programs and found no protectable expression because "the same modules would be an inherent part of any prompting program. Their order and organization can be more closely analogized to the concept of wheels for the car rather than the intricacies of a particular suspension system."¹¹⁰

These cases provide strong authority for excluding many of the object-oriented elements in a program that models real-world behavior. For example, in QuadWorld, the entire class and inheritance structure flows directly from the natural relationships between squares, rectangles, parallelograms, and quadrilaterals which, in turn, are dictated by formal mathematical definitions in the real world. Similarly, in the traffic light control program, nothing in the semantic data model would be protectable because these relationships are dictated by the functional behavior of trip sensors, controllers, and traffic lights. Finally, the relevance of *scenes à faire* to object-oriented software is further underscored by our approach to design in step 2, in which we wrote "scripts" for each object, making it fairly easy for a court to compare each object to a "stock character" in the real-world system being modeled.

In most situations, the list of services that each object must provide will be largely dictated by these scripts and hence will be unprotectable. In certain cases, it might be possible to identify certain low-level objects¹¹¹

105. *Id.*

106. 807 F.2d 1256 (5th Cir. 1987), *cert. denied*, 484 U.S. 821 (1987).

107. *Id.* at 1262.

108. *Id.* at 1262 n.4.

109. 625 F. Supp. 608 (S.D.N.Y. 1985)

110. *Id.* at 616 (citation omitted).

111. By low-level objects, I mean certain objects which are simply building blocks in constructing more complex objects which model real-world properties. At this low level, the building block may be sufficiently removed from real-world behavior to render *scenes à faire* inapplicable. However, even these objects may often be taken from libraries of reusable objects and should be excluded from protection because they do not satisfy copyright's originality requirement. See *infra* Section IV.B.4 discussing the public domain filter.

that do not directly model real-world behavior and could therefore escape the *scènes à faire* filter. In general, however, the only elements that will survive this filter are low-level implementations of specific methods; at that level, those portions of code resemble traditional programs and embody few object-oriented principles.

4. PUBLIC DOMAIN

The "public domain" filter will also be extremely important in analyzing object-oriented software. Since object-oriented design focuses on reusable software components, many complex object-oriented programs will take advantage of existing objects that have been written for other programs. In some cases, these objects may be taken from public domain libraries, such as those provided on electronic bulletin boards. As Nimmer notes, "It is axiomatic that material in the public domain is not protected by copyright even when incorporated into a copyrighted work."¹¹² As a result, the court must eliminate any objects taken from public domain when determining which elements of the program are protectable.

However, the bulk of reusable objects may not come from entirely "public" sources. These reusable objects may come from vendors selling libraries of pre-defined objects on a license basis, particularly in the case of graphical user interfaces and database systems. These vendors clearly intend that their libraries will be incorporated into commercial products.¹¹³ Nonetheless, these objects should not be included in the scope of the copyright protection for the final commercial product, as they would not be original to the programmer claiming authorship of the final product, and hence could not pass copyright's threshold test for originality.¹¹⁴ As a result, the court must treat the use of licensed objects

112. NIMMER & NIMMER, *supra* note 3, at 13-78.43 (citing *Sheldon v. Metro-Goldwyn Pictures Corp.*, 81 F.2d 49, 54 (2d Cir. 1936), *aff'd*, 309 U.S. 390 (1940)); *see also Computer Assocs. Int'l v. Altai, Inc.*, 982 F.2d 693, 710 (2d Cir. 1992) (public domain "material is free for the taking and cannot be appropriated by a single author even though it is included in a copyrighted work").

113. In addition to the previously discussed object libraries for implementing the Macintosh user interface, see *supra* note 6, vendors are hawking a wide variety of object libraries for use in commercial applications. A quick perusal of advertisements and articles in any programming trade magazine will confirm the growth of this industry. *See, e.g., COMM. ACM*, Oct. 1991.

114. Copyright protection is allowed only for *original* works of authorship. 17 U.S.C. § 102(a) (1988). At a minimum, original authorship means that the programmer did not directly take the expression from any other source, whether public or not. *Feist Publications v. Rural Tel. Serv., Inc.*, 111 S. Ct. 1282, 1287 (1991) ("The sine qua non of copyright is originality. To qualify for copyright protection, a work must be original to the author. Original, as the term is used in copyright, means only that the work was independently created by the author (as opposed to copied from other works), and that it possesses at least some minimal degree of creativity." (citation omitted)).

on the same basis as truly "public domain" objects. In both cases, copyright protection would not be available for any object which the programmer seeking protection did not write.

C. Economic Balancing Approach

While no court has yet adopted the economic balancing approach, several recent commentators on software protection have suggested answering the idea/expression problem by balancing the copyright plaintiff's creative contribution against the loss to society from granting the plaintiff a monopoly over particular software code. In one version of this approach, the court would determine the existence of protectable expression by following a two-step test:

The first step is for the court to define as specifically as possible the thing that the defendant has taken from the plaintiff . . . the second step is to decide whether that thing is original to the plaintiff . . . That is, to get that thing the defendant took, did the plaintiff invest costly creative effort that presumptively relied on the promise of copyright? If so, judgment properly goes to the plaintiff, because, in conclusory terms, the defendant has taken the plaintiff's expression. Or did the plaintiff get that thing by copying it effortlessly from existing and available sources, or by otherwise responding entirely to incentives other than copyright? If so, judgment properly goes to the defendant because, again stating it in conclusory terms, the defendant took only the plaintiff's idea.¹¹⁵

In another version of the economic balancing approach, the court would divide the plaintiff's program at different levels of abstraction and then determine the dividing line between idea and expression by "balancing the need to provide an incentive to authors against the cost to society of losing the free use of the author's work at that level of expression."¹¹⁶

While the economic balancing approach seems intriguing as a matter of innovation policy, courts are not likely to endorse it primarily because it is not supported by copyright doctrine. Both versions require the court to parse the plaintiff's work in a manner similar to the "patterns of abstractions" test first articulated in *Nichols v. Universal Pictures Corp.*¹¹⁷ However, both versions ultimately depart from copyright doctrine by requiring the court to balance the economic return necessary to induce the author to produce a particular type of work against the cost to society of granting that author a monopoly over particular expression at a particular level of abstraction. This equation confuses the distinctions between copyright and patent law. In patent law, the author's creative

115. Wiley, *Copyright at the School of Patent*, 58 U. CHI. L. REV. 119, 158-59 (1991).

116. Reback & Hayes, *supra* note 3, at 5.

117. 45 F.2d 119, 121 (2d Cir. 1930), cert. denied, 282 U.S. 902 (1931).

contribution is assessed by the requirements of utility, novelty, and non-obviousness.¹¹⁸ The cost to society is controlled by requiring the inventor to define the invention with specific claim language sufficiently narrow to avoid the prior art and by requiring that those claims be supported by the specification, thus ensuring that most patents will have a fairly narrow scope. Moreover, the costs of protection are offset by the societal benefits resulting from full disclosure of the underlying technology in the patent specification.

In contrast, copyright asks little of the author except that the work not be copied from any other source and that the work reflect at least minimal creativity.¹¹⁹ While copyright law limits the monopoly costs to society by allowing independent creation to be an absolute defense to infringement, it provides no doctrinal tools for defining the scope of the monopoly against potential defendants who have had access to the work. As a result, the economic approach is difficult to support with copyright doctrine. In fact, the author of the first version acknowledged this dilemma and explicitly developed his test by applying the "good sense" of patent doctrine in order to "rationalize" copyright doctrine.¹²⁰

Even if the economic approach could somehow be justified under traditional copyright doctrine, it is not clear that the economic approach would be particularly desirable as a matter of policy. Under the first version, the court would face the elusive task of *objectively* determining whether the plaintiff would have authored the code appropriated by the defendant in the absence of copyright incentives. Beyond the obvious evidentiary problems in this analysis, the process of analyzing incentives *ex post* leads unavoidably to circular reasoning. Whether the plaintiff was motivated by the promise of copyright depends to a large degree on the generally perceived rule of law regarding the scope of software copyrights. But, at the same time, the purpose of the two-part test itself is to determine the proper scope of the idea/expression dichotomy and hence announce a new rule of law. This dilemma is further exacerbated by the small number of software copyright cases that result in published decisions.

The second version of the economic approach faces similar problems. First, the *ad hoc* nature of the inquiry makes it difficult for

118. See 35 U.S.C. §§ 101-103 (1988).

119. While the *Feist* case may have raised the standard of originality required by copyright, it did not raise that standard anywhere close to requiring an analysis of creative contributions.

120. Wiley, *supra* note 115, at 120 ("Using an economic perspective on innovation policy, this Part defends the notion that we should regard core portions of patent doctrine as intellectual successes worthy of imitation. Most fundamentally, patent law establishes a set of sensible and efficient *incentives* to creation. Copyright should learn this basic lesson, for a focus on sound incentives would give copyright doctrine the coherence it now lacks.").

software companies to make rational business decisions based on which aspects of their own software and their competitor's software are protectable in copyright.¹²¹ Moreover, the formulation of the test suggests a false empiricism. Even for software products aimed at mature business markets, it will be extremely difficult to determine the "cost" to society of granting the plaintiff a monopoly. More fundamentally, even if this cost could be accurately calculated, it must be balanced against the purely speculative "creative contribution" of the author, which inevitably invites judgments which are nothing more than a determination that the plaintiff's work is novel, and non-obvious, and therefore worthy of protection. This type of analysis is better left to patent law, where more precise standards exist for determining non-obviousness and where the court has the benefit of an initial expert analysis performed by the patent examiner.¹²²

D. Copyright Doctrine Properly Applied Provides Little Protection for Object-Oriented Software

While Nimmer's filtering test is closely linked to traditional copyright doctrine, it may present an unnecessary exercise in the case of object-oriented software. As a practical matter, Nimmer's filters will exclude from protection nearly every element that makes a particular program object-oriented in design. More important, these elements are precisely the elements which reveal software's "behavioral" rather than "textual" nature and which render object-oriented programs generally unsuitable for copyright protection.

While most courts recognize that computer programs are utilitarian articles, most infringement cases require the court to analyze only the textual representation of the program's structure, sequence, and organization as embodied in source code. As a result, most courts focus on the textual embodiment of software and quickly lose sight of the behavioral nature of software. This distinction was first recognized in *Computer Associates International v. Altai, Inc.*,¹²³ in which the district court adopted the findings of a court-appointed expert who explained that:

a computer program must be viewed both as text and as behavior.

The text perspective focuses upon the object code and source code A computer program, however, is more than a collection of zeros and ones. When properly loaded into a computer and

121. Beutel, *supra* note 84, at 27 (noting that "[j]ust as the application of the antitrust 'rule of reason' has taken years of dissection and analysis to take form, so too would the eventual parameters of software copyright under the policy-balancing approach set forth in the Reback/Hayes Abstractions Test").

122. Critics of software patents often question the competence of software examiners in analyzing software issues. This problem is discussed further *infra* Part V.

123. 775 F. Supp. 544 (E.D.N.Y. 1991), *aff'd in relevant part*, 782 F.2d 693 (2d Cir. 1992).

provided with appropriate input from, for example, the keyboard, the program behaves. In a word processing program, for example, text can be deleted, blocks of text can be moved, formatting of documents can be changed; all sorts of operations can be instituted; and these can only be described as behavior.¹²⁴

While the court used this analysis primarily to criticize *Whelan* for failing to distinguish between the static, textual view of the program and the dynamic, behavioral view,¹²⁵ the court also recognized that the behavioral aspect of software creates a much more fundamental problem when viewed against the statutory limits imposed by § 102(b):¹²⁶

Going beyond Dr. Davis' analysis, the court notes a possible statutory difficulty that arises when we recognize, as we must, that a computer program "behaves." . . . Since the behavior aspect of a computer program falls within the statutory terms "process", "system", and "method of operation", it may be excluded by statute from copyright protection. . . . Fortunately, this court need not wrestle with that possible development in the law of intellectual property, because CA's rights in this case are fully protected by viewing the ADAPTER program as text.¹²⁷

Although the *Computer Associates* court did not have to resolve this question, courts dealing with object-oriented software must address the behavioral nature of software. Courts will be confronted with this problem from three different angles. First, if courts approach object-oriented programs as they would approach programs written under the traditional model, they will find that the structure, sequence, and organization of the source code tell us little about the inheritance relationships and class structures in the program, which may be the part of the program that the plaintiff most wants to protect. Moreover, the closer the plaintiff adhered to the object-oriented model in the program's creation, the more pronounced this phenomenon will be. In fact, since programs that make good use of polymorphism and dynamic binding must include source code that is highly generalized, the source code behind the best written programs will tell us the least about the objects in the program.

Faced with this problem, the court will then have to focus on the specific class definitions used to specify inheritance relationships, messages accepted, and method implementations. However, while these definitions are expressed in the English words used by a particular programming language, they are simply a shorthand description of a

124. *Id.* at 559.

125. *Id.* at 560.

126. 17 U.S.C. § 102(b) provides: "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work."

127. *Computer Assocs.*, 775 F. Supp. at 560.

highly specific system. When we define the class, "rectangle" as a sub-class of the class "parallelogram," there is nothing expressive, in the copyright sense, about that definition. The term "sub-class" is a shorthand instruction that tells the compiler "whenever you see a rectangle, have it behave just like a parallelogram, except when you receive a message which has been overridden in the definition of the rectangle class, then use a different behavior."

Third, the court will have to approach the program by examining the high-level relationships among different classes and objects because the textual descriptions of a particular class cannot be protected,. In fact, the court may well be tempted to examine substantial similarity by asking the parties to create a semantic data model of each program, on the theory that if the semantic data models are substantially similar, the programs must be substantially similar. However, this approach effectively creates copyright protection for semantic data models themselves, a result which cannot be justified under fundamental copyright principles. True, a programmer who draws a semantic data model can claim a copyright in the pictorial representation that the programmer used to express the model; that programmer can prevent others from copying the *picture*. However, the copyright in the picture cannot be used to indirectly grant protection over the model itself, a result which follows directly from *Baker v. Selden*:¹²⁸

The copyright of a work on mathematical science cannot give to the author an exclusive right to the methods of operation which he propounds, or to the diagrams which he employs to explain them, so as to prevent an engineer from using them whenever occasion requires.¹²⁹

At a more fundamental level, the semantic data model cannot be used to determine infringement because it is simply a list of the constituent elements of a particular system. Copyright protection cannot be used to provide a monopoly over these elements, a point which was

128. 101 U.S. 99 (1879) (holding that copyright can reside in a particular explanation of a system, but not in the system itself). *Baker* is generally regarded as the inspiration for § 102(b). See Amicus Curiae Brief of Copyright Law Professors at 5, *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 799 F. Supp. 203 (D. Mass. 1992) (No. 90-1162-K) [hereinafter Copyright Professors' Amicus Brief] ("It is to cases such as *Baker v. Selden* and its progeny that courts should look in interpreting section 102(b) and its exclusion of systems and methods from the scope of copyright protection available to works of authorship." (citation omitted)).

129. *Baker*, 101 U.S. at 103; see also Copyright Professors' Amicus Brief, *supra* note 128, at 6 n.3 ("[T]he [Baker] Court pointed out that in most instances, useful arts were embodied in wood, metal, or stone, and what had given plausibility to Selden's claim was that his useful art was embodied in a writing. Nevertheless, the Court stated 'the principle is the same in all. The description of the art in a book, though entitled to the benefit of copyright, lays no foundation for an exclusive claim to the art itself.' " (quoting *Baker*, 101 U.S. at 105)).

recently restated in an amicus curiae brief submitted by eleven well-respected copyright professors in *Lotus v. Borland*:¹³⁰

It is in the nature of a method or system to have constituent elements, some of which may be quite detailed in character. In the "Shorthand cases," courts will decline to extend copyright protection not only to the set of abstract rules that a shorthand system developer might have devised for condensing words or phrases, but also to the vocabulary resulting from the implementation of these rules. Both are constituent elements of the system which copyright law will not protect.¹³¹

This analysis can be directly applied to a semantic data model. For example, in the traffic light program, the semantic data model tells us "a clock is a part of a controller, and a controller reads from a sensor which can be either a pressure sensor or a trip sensor." This semantic data model equally describes the real-world physical system and the system for modeling that behavior on a computer. Just as the traffic light and controller are constituent elements in the real-world traffic intersection, the representations of those entities as objects and classes are constituent elements of a *system* for modeling the behavior of a traffic intersection on a computer.

Finally, the semantic data model is exactly what its title implies, an attempt to explain a detailed system in words and pictures. The plaintiff presenting a semantic data model as the basis for proving substantial similarity is not arguing that the defendant used the same words and pictures to depict the system, but rather that the defendant used the same *system* of classes and inheritance relationships in writing the allegedly infringing program. As soon as the plaintiff presents a semantic data model as the basis for infringement, the court must recognize that the plaintiff is seeking protection for the constituent elements of a particular object-oriented system, a right which has no basis in copyright law.

The preceding analysis shows that copyright law does not protect the high-level relationships among objects. The fact that these relationships may represent the bulk of the programmer's effort and innovation during design is irrelevant in determining the scope of protection under copyright doctrine.¹³² If protection for such behavioral elements in object-oriented software is available, it can only be achieved through the patent system.¹³³

130. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 799 F. Supp. 203 (D. Mass. 1992) (granting partial summary judgment).

131. Copyright Professors' Amicus Brief, *supra* note 128, at 7.

132. See *supra* note 81.

133. See *Baker*, 101 U.S. at 105 ("The description of the art in a book, though entitled to the benefit of copyright, lays no foundation for an exclusive claim to the art itself. The object of the one is explanation; the object of the other use. The former may be secured by copyright. The latter can only be secured, if it can be secured at all, by letters-patent.");

V. PATENT PROTECTION FOR OBJECT-ORIENTED SOFTWARE

In some ways, the patentability of object-oriented software is easier to analyze than the patentability of traditional software. In cases involving traditional software, the primary question has been whether the software recites "a mathematical algorithm."¹³⁴ If it does, then the software is not patentable; otherwise the software is patentable subject matter.¹³⁵ Indeed, this analysis would still apply to a patent claim that was drawn to the low-level internal implementation of a specific method in an object-oriented program, since that portion of the program operates on the same principles as traditional software. In that case, the court would also have the benefit of examining the claim in light of a substantial body of critical commentary that has been written on the patentability of traditional software.¹³⁶ The more interesting question is what higher-level elements of the object-oriented model could qualify as patentable subject matter.

A. Patentable Subject Matter

The most promising candidate for protection is a patent claim drawn to a semantic data model. In fact, a purely textual description of a semantic data model would read very much like a standard apparatus claim. In the case of the traffic light example, we could construct a patent

Computer Assocs. Int'l v. Altai, Inc., 775 F. Supp. 544, 560 (E.D.N.Y. 1991) (noting in the context of the problems raised by the behavioral aspects of software, "indeed, it has been suggested that computer software is better protected by patent law than by copyright law"), *aff'd in relevant part*, 982 F.2d 693 (2d Cir. 1992).

134. *Gottschalk v. Benson*, 409 U.S. 63 (1972).

135. The Federal Circuit and its predecessor courts have devised a two-part test, the "Freeman-Walter" test, to determine whether a particular software claim is drawn to patentable subject matter. In the first step, the court must determine whether the claim directly or indirectly recites a mathematical algorithm. If it does not, then the claim is drawn to patentable subject matter. However, even if the claim does recite a mathematical algorithm, it may still be patentable if the claim "implement[s] the algorithm in a specific manner to define structural relationships between the elements of the claim in the case of apparatus claims, or limit or refine physical process steps in the case of process or method claims." *In re Walter*, 618 F.2d 758, 767 (C.C.P.A. 1980). See generally *In re Iwahashi*, 888 F.2d 1370 (Fed. Cir. 1989); *In re Pardo*, 684 F.2d 912 (C.C.P.A. 1982); *In re Abele*, 684 F.2d 902 (C.C.P.A. 1982); *In re Freeman*, 573 F.2d 1237 (C.C.P.A. 1978); *PTO Report On Patentable Subject Matter: Mathematical Algorithms and Computer Programs*, 38 Pat. Trademark & Copyright J. (BNA) 563 (1989) [hereinafter PTO Report].

136. See, e.g., Donald S. Chisum, *The Patentability of Algorithms*, 47 U. PITTS. L. REV. 959 (1986); Pamela Samuelson, *Benson Revisited: The Case Against Patent Protection for Algorithms and Other Computer Program-Related Inventions*, 39 EMORY L.J. 1025 (1990); Randall M. Whitmeyer, Comment, *A Plea for Due Processes: Defining the Proper Scope of Patent Protection for Computer Software*, 85 NW. U. L. REV. 1103 (1991) [hereinafter Comment, *A Plea for Due Processes*].

claim for a real-life intersection control system that read something like this:

A traffic control apparatus consisting of:

a trip sensing means and a pressure sensing means, and a controller device which is operably connected to receive signals from said sensing means, and operably connected to send signals to a sequential display of different colored lights.

In the case of a real-world traffic control system, this claim would certainly recite patentable subject matter. In the case of object-oriented software, this claim is a close description of our semantic data model. Of course, in the case of software, the "trip sensing" means refers not to a physical object but to a location in the computer's memory that is designed to model the behavior of the real-world "trip sensing" means. While the case law on this issue is somewhat confused, a strong case can be made for holding that the above claim should be patentable subject matter whether it refers to the computer model of a traffic intersection or the physical apparatus used in the real world.

1. EXISTENCE OF MATHEMATICAL ALGORITHM

The initial inquiry for computer program related inventions focuses on the existence or absence of a mathematical algorithm.¹³⁷ If the claimed invention is drawn at the level of the semantic data model, no mathematical formulas will appear in the claim. Because the object-oriented model emphasizes encapsulation of data and procedures, the implementation of simple mathematical formulas should be hidden in the internal methods of each class and is generally invisible in the semantic data model.¹³⁸ At this point, the fact that the semantic data model may still embody a "non-mathematical" algorithm, in the broad sense, does not disqualify it from patent protection.¹³⁹

137. *Iwahashi*, 888 F.2d at 1374 ("[T]he proscription against patenting has been limited to *mathematical* algorithms and abstract *mathematical* formulae which, like the laws of nature, are not patentable subject matter.") (emphasis in original); *PTO Report, supra* note 130, at 570 ("The major (and perhaps only) exception in the area of computer processes is the mathematical algorithm . . . If a computer process claim does not contain a mathematical algorithm in the Benson sense, the second step of the Freeman-Walter-Abele test is not reached, and the claimed subject matter will usually be statutory.").

138. One might question whether the specification describing a semantic data model would be sufficiently enabling under 35 U.S.C. § 112. In most cases, the mathematical formulas necessary to construct a working program will be obvious to those skilled in the art. In those cases where the implementation is not obvious, the PTO could require the applicant to disclose those formulas in the specification, perhaps as part of the best mode requirement.

139. All apparatus claims could be considered to follow an algorithm in the broad sense of the term. See *Iwahashi*, 888 F.2d at 1375 ("[T]he fact that the apparatus operates according to an algorithm does not make it nonstatutory.").

The Federal Circuit has been quick to grant claims in which a single system of physical elements and a computer program are drafted as a single claim.¹⁴⁰ In such cases, “[t]he claim as a whole certainly defines [an] apparatus in the form of a combination of interrelated means and we cannot discern any logical reason why it should not be deemed statutory subject matter . . .”¹⁴¹ For many claims drawn to object-oriented programs, a strong argument can be made that the relationships between objects act much like the interaction between physical elements of a real-world apparatus in which different operational “means” send signals to one another and respond accordingly. In fact, as illustrated above by the sample claim for a traffic light control system, a single claim could equally describe the semantic data model or the real-world system itself.¹⁴² The close identity between the description of a real-world system and the object-oriented program which models that system reinforces the argument that a claim based on the object-oriented program presents the same statutory subject matter as a conventional claim for the physical system.

Finally, while software that models or controls real world objects presents the best candidate for patentable subject matter, protection may also be available for object libraries which have no real-world counterparts, such as object-oriented graphical user interfaces and database systems. In these cases, the software still represents the interactions of various “means” designed to control the internal workings of a general purpose computer. The Federal Circuit has already

140. See *id.* (auto-correlation unit for use in pattern recognition); *In re Abele*, 684 F.2d 902 (C.C.P.A. 1982) (software program for improved CAT-scan process); *In re Taner*, 681 F.2d 787 (C.C.P.A. 1982) (software which improved seismic exploration by translating spherical seismic waves into plane or cylindrical waves); *In re Freeman*, 573 F.2d 1237 (C.C.P.A. 1978) (software for controlling conventional phototypesetter).

141. *Iwahashi*, 888 F.2d at 1375.

142. The fact that such a claim could be drafted also implies that a single claim would grant the inventor a monopoly over both the real-world physical system and the object-oriented model of that system, a result which undoubtedly raises alarms in certain circles. However, three considerations mitigate the danger of this result. First, the single claim will have to withstand prior art from both the computer science field and field relating to the real-world physical system. See *infra* Section V.B. Few claims will be non-obvious when tested against such a wide range of prior art. Second, the claim grants a monopoly only to the extent that the invention is enabled by the specification under the standards in 35 U.S.C. § 112. In many cases, the inventor may be able to describe how to write the object-oriented program but will be unable to explain how to actually build some of the elements in the physical system. For example, while we can easily model a “trip sensor” in a program, it may be much more difficult to build one that works consistently when embedded in a roadway. Particularly since the vast majority of claims will have to be written in “means-plus-function” form, the specification will sharply limit the actual scope of the monopoly granted by the single claim. Finally, if the inventor has presented a single claim which is truly non-obvious and which enabled both the computer and physical versions of the systems, then the inventor has really *invented* both systems and should be entitled to protection over both.

recognized the patentability of pure software claims which direct the way the computer manages data internally.¹⁴³

2. MENTAL STEPS

The mental steps doctrine was historically used to deny patent protection for process claims involving simple measurements, calculations, and interpretations of data that could just as easily be performed by a human using paper and pencil.¹⁴⁴ However, the C.C.P.A. may have broadened the doctrine in 1982 when it denied patent protection to an expert system for neurological diagnosis on the basis that "their invention is concerned with replacing, in part, the thinking processes of a neurologist with a computer."¹⁴⁵ Moreover, the Federal Circuit has implicitly used the doctrine to invalidate a claim for an invention designed to determine whether any complex system is in a normal or abnormal state.¹⁴⁶

In its broadest form, the mental steps doctrine would deny patent protection to any expert system. Since object-oriented development emphasizes approaching the project from the perspective of an expert in the problem domain, semantic data models may mirror the mental process that an expert in that field would use to solve problems. However, many object-oriented programs will be able to survive the mental steps doctrine for two reasons. First, the software claims recently invalidated under the mental steps doctrine involved the calculation of a discrete result and in general seemed close to a simple process of mental calculations.¹⁴⁷ In contrast, many object-oriented programs will model the operation of systems with continuous behavior that produce no discrete "answer" to a problem. For example, the "QuadWorld" program solves no specific problem, but rather provides a system for drawing and manipulating a variety of shapes. It is difficult to conceive of QuadWorld

143. See *In re Pardo*, 684 F.2d 912, 913 (C.C.P.A. 1982) (invention which "converts a computer from a sequential processor . . . to a processor which is not dependent on the order in which it receives program steps"); *In re Bradley*, 600 F.2d 807 (C.C.P.A. 1979) ("firmware" designed to improve performance of multi-tasking), *aff'd sub nom.* *Diamond v. Bradley*, 450 U.S. 381 (1981).

144. Samuelson, *supra* note 136, at 1034-38.

145. *In re Meyer*, 688 F.2d 789, 795 (C.C.P.A. 1982).

146. *In re Grams*, 888 F.2d 835, 840 (Fed. Cir. 1989) (analogizing to *Meyer* and finding the existence of an algorithm in part because "the objective [in *Meyer*] of identifying malfunction is similar to the objective here of identifying abnormality").

147. In fact, the Federal Circuit never explicitly mentioned the mental steps doctrine but rather denied the claims because "[f]rom the specification and the claim, it is clear to us that applicants are, in essence, claiming the mathematical algorithm, which they cannot do . . ." *Grams*, 888 F.2d at 840; see also Comment, *A Plea for Due Processes*, *supra* note 136, at 1122.

as a series of discrete mental steps that could be performed to achieve the same result.

Second, the C.C.P.A. did not extend the doctrine to cases where the system could theoretically be performed as a series of mental steps but, as a practical matter, would be too complex to implement with pen and paper.¹⁴⁸ Similarly, many object-oriented programs will reflect complex relationships between elements of an extremely large system. At the level of "programming in the colossal", few object-oriented systems can be reduced to a series of human mental process steps. As a result, the mental steps doctrine should not provide an independent bar to patentability.

3. METHOD OF DOING BUSINESS

If the method of doing business limitation were applied seriously, it would exclude from protection any computer program that implemented familiar business systems, a category that could ensnare many object-oriented database systems. However, the doctrine is of questionable validity¹⁴⁹ and has only been weakly applied in computer cases. For example, in *Paine, Webber v. Merrill Lynch*,¹⁵⁰ a district court noted the existence of the doctrine, but dismissed it because

[t]he product of the claims of the '442 patent effectuates a highly useful business method and would be unpatentable if done by hand. The C.C.P.A., however, has made clear that if no *Benson* algorithm exists, the product of a computer program is irrelevant, and the focus of analysis should be on the operation of the program on the computer.¹⁵¹

On this basis, the court upheld a claim for a computer program that implemented Merrill Lynch's "Cash Management Account System" which allowed customers to combine brokerage, money market,

148. See *In re Toma*, 575 F.2d 872 (C.C.P.A. 1978) (allowing claims for a computer process of translating from any source language to any target language by examining a language dictionary, examining the syntax of the source and then producing a complete sentence in the target language).

149. Arthur J. Hansmann, *Method of Doing Business*, 50 J. PAT. OFF. SOC'Y 503, 504 (1968) ("Except for dicta, one can conclude that there is no basis in existing law for the rejection of claims as being directed to a 'method of doing business.'"); David J. Meyer, Note, *Paine, Webber, Jackson and Curtis, Inc. v. Merrill Lynch, Pierce, Fenner & Smith: Methods of Doing Business Held Patentable Because Implemented on a Computer*, 5 COMPUTER L.J. 101, 103-04 n.13 (1984) (reviewing the cases cited in *Merrill Lynch* and concluding that "examination of these cases reveals that the issue of patentable subject matter was never actually decided. Rather, the patent claims were held invalid for 'lack of invention.' . . . The issue of the patentability of a method of doing business was discussed only in dictum . . ."); Comment, *A Plea for Due Processes*, *supra* note 136, at 1119 ("[I]t is unclear whether this doctrine ever really existed . . .").

150. *Paine, Webber, Jackson & Curtis, Inc. v. Merrill, Lynch, Pierce, Fenner & Smith, Inc.*, 564 F. Supp. 1358 (D. Del. 1983).

151. *Id.* at 1369.

checking, and credit cards into one integrated account. Similarly, the C.C.P.A. has allowed claims for a program to control the optimal operation of plants, such as oil refineries, at multiple locations,¹⁵² and for a program that produced architectural specifications and project control instructions.¹⁵³ Finally, in the only case to invalidate a business methods program, the C.C.P.A. did not even mention the doctrine but declared the claim invalid on the basis that the claim recited and preempted a specific algorithm.¹⁵⁴

These cases indicate that the doctrine may have little effect on inventions related to computer programs in general. Moreover, object-oriented programs will be affected even less by the doctrine than traditional software. As discussed in the next Section, if the program merely implements a familiar business method, then prior art relevant to general business methods and practices will invalidate the claim on § 103 grounds.¹⁵⁵ Thus, courts evaluating such claims will never have to reach the business methods question, since it will be easier to resolve the issue on § 103 grounds and confine the analysis of patentable subject matter to determining whether the claim recites a mathematical algorithm.

B. Non-Obviousness and the Relevant Prior Art

Even though semantic data models qualify as patentable subject matter, few patents will actually be issued because few will pass the non-obviousness requirements of 35 U.S.C. § 103. While critics of software patents have claimed that the Patent Office lacks the expertise or the database files to accurately evaluate prior art for software patents,¹⁵⁶ this problem is considerably less severe in the case of object-oriented software. Since a standard for determining the relevant fields of prior art is "whether it deals with a problem similar to that being addressed by the

152. *In re Deutsch*, 553 F.2d 689 (C.C.P.A. 1977).

153. *In re Phillips*, 608 F.2d 879 (C.C.P.A. 1979).

154. *In re Maucorps*, 609 F.2d 481, 486 (C.C.P.A. 1979) (claim for a program that determined the optimal organization of a sales force).

155. See 35 U.S.C. § 103 (1988).

156. See, e.g., Brian Kahin, *The Software Patent Crisis*, TECH. REV., April 1990, at 53, 55.

The search [for software prior art] is extraordinarily difficult because the field's printed literature is thin and unorganized. Software documents its own design, in contrast to physical processes, which require written documentation. Also, software is usually distributed without source code under licenses that forbid reverse engineering. This may amount to suppressing or concealing the invention and therefore prevent the program from qualifying as prior art. . . . Many programmers suspect that patent examiners lack knowledge of the field, especially since the Patent Office does not accept computer science as a qualifying degree for patent practice

....

inventor,"¹⁵⁷ the examiner will have to search for references not only in the computer science area, but also in the literature relating to the real-world problem being addressed by the software. Because the programmer first approached the project in step 1 of our object-oriented design model by learning as much as possible from experts in the problem domain itself,¹⁵⁸ the examiner will also have to use the full range of literature in the problem domain as prior art. This search may prove fairly easy since the applicant's duty of candor will require the programmer to unilaterally disclose to the patent office all the sources used in developing the project.¹⁵⁹

The prior art problem will also be less significant because the examiner will be better equipped to determine non-obviousness.¹⁶⁰ For example, in the traffic light problem, the examiner will compare the semantic data model to literature that describes the operation of the physical entities that operate traffic lights in the real world. In general, this literature will reflect principles of electrical and mechanical engineering that are more familiar to most examiners than principles of computer science. The examiner can quickly determine whether the software is merely a straightforward model of the physical system and therefore obvious to the hypothetical person of ordinary skill in the art of object-oriented design. In many cases, the claim for the semantic data model will read almost exactly like a claim for a well-documented physical system and thus will quickly appear obvious to the examiner.

Under this analysis, the elements of object-oriented software that most embody object-oriented design are patentable subject matter. As a practical matter, however, only the small percentage of semantic data models that are truly non-obvious, in the face of an extremely broad range of relevant prior art, will be granted patents.

157. Union Carbide Corp. v. American Can Co., 724 F.2d 1567, 1572 (Fed. Cir. 1984) ("The determination that a reference is from a nonanalogous art is therefore two-fold. First, we decide if the reference is within the field of the inventor's endeavor. If it is not, we proceed to determine whether the reference is reasonably pertinent to the particular problem with which the inventor was involved." (quoting *In re Wood*, 599 F.2d 1032, 1036 (C.C.P.A. 1979)); see also *Bott v. Four Star Corp.*, 218 U.S.P.Q. (BNA) 358, 368 (E.D. Mich. 1983) ("The test for relevant or analogous prior art is 'similarity of element, problems, and purposes.' 'Analogous art is that field of art which a person of ordinary skill in the art would have been apt to refer in attempting to solve the problem solved by a proposed invention.'" (citations omitted)), *aff'd*, 732 F.2d 168 (Fed. Cir. 1984).

158. See *supra* Section III.C.1.

159. 37 C.F.R. § 1.56 (1992).

160. An often-heard complaint against traditional software patents is that examiners untrained in computer science "naturally have a lower standard in determining the hypothetical 'person having ordinary skill in the art,' and are thus more apt to grant patents for obvious processes." Kahin, *supra* note 156, at 55.

VI. IMPLICATIONS FOR INNOVATION POLICY

This article has presented the most important concepts of object-oriented programming and discussed one model of the object-oriented design process. A thorough understanding of object-oriented design shows that copyright protection cannot be justified for the elements of the software that make it object-oriented. In addition, one product of the object-oriented design process, the semantic data model, may be used to draft a patentable claim, if it is sufficiently innovative relative to an extremely broad range of prior art.

These conclusions are based on the application of existing legal doctrines, rather than on an analysis of which legal regime provides better protection as a matter of innovation policy. While the industry is still too new to perform any worthwhile empirical analysis, several intuitive observations about the advantages of each legal regime suggest that limited patent protection provides reasonable incentives for innovation.

Legal protection for software designed according to object-oriented principles is important only at the margins of innovation. Assuming that verbatim copying and distribution can be prevented,¹⁶¹ most companies that write commercial software are not motivated by the promise of broad legal protection for their products but by economic returns that result from being first in the market or being the first to introduce programs with new features into an existing market. While there is no method to test this hypothesis, the nature of object-oriented development itself suggests that companies will be motivated to innovate without extensive legal protection. Object-oriented programming will be adopted because it allows complex programs to be created with fewer errors, encourages the use of existing software components, leads to easier software maintenance, and eases the process of improving the software in subsequent versions. As a result, companies utilizing object-oriented techniques will produce better products and face lower development costs than companies using traditional techniques. This improvement in software "manufacturing" will provide most of the incentive necessary to stimulate innovations in object-oriented software. Indeed, these incentives are evident in the growth of the Object Management Group

161. Of course, outright piracy destroys all incentives for innovation. However, verbatim copying and distribution is easy to prohibit, at least within the United States, through traditional copyright protection for literal source and object code. As in the prior discussion of legal doctrines, this discussion is concerned with the more interesting problems presented by copyright and patent protection for the object-oriented elements of software.

(OMG), a "technology endorsement group" whose membership includes almost 200 leading software companies.¹⁶²

Against this background of strong innovation, legal protection can provide some additional incentives. Analyzing legal protection from the standpoint of innovation policy suggests that copyright protection would be inappropriate for object-oriented software¹⁶³ while patent protection may be justified for highly innovative programs.

Copyright protection has been popular for traditional software primarily because it is easy to obtain, it provides strong protection against literal and non-literal copying, and it still allows for a defense of independent creation. However, copyright protection presents serious problems when applied to the aspects of a particular program that make it object-oriented. When applying for a copyright, the author need not make any attempt to define the scope of the copyright being claimed. The author simply submits a copy of the source code or object code of the program with the registration form, and copyright protection instantly attaches. As the Second Circuit recently noted, "we think that copyright registration—with its indiscriminating availability—is not ideally suited to deal with the highly dynamic technology of computer science."¹⁶⁴

The net result of this process is that the scope of any particular copyright is not defined until litigation occurs and, even then, is only defined relative to the particular program accused of infringement. As a result, business competitors cannot legitimately plan future products since they cannot be sure of how to "design around" an existing copyright. This problem is most acute when protection is claimed for the non-literal elements of a program, such as the semantic data model. While the independent creation defense provides some protection for competing software developers, the high mobility of software engineers¹⁶⁵ combined with the questionable legal status of reverse engineering techniques¹⁶⁶ may render any protection highly illusory.

162. Object Management Group Purpose and Definition Statement (1992) (on file with author). OMG was originally formed in April 1989 by Data General, Hewlett-Packard, Sun, Canon, American Airlines, Unisys, Philips, Prime, Gold Hill, Soft-Switch, and 3-Com. Major software players such as AT&T, Digital, NCR, Borland, Microsoft, and IBM have subsequently joined.

163. Remember that copyright protection would still be available to prevent verbatim copying of source and object code, thus supplying the necessary prerequisite to innovation discussed earlier.

164. Computer Assocs. Int'l v. Altai, Inc., 982 F.2d 693, 712 (2d Cir. 1992).

165. Absent costly clean room development procedures, the accused infringer may find it extremely difficult to prove that every engineer involved in the project was completely ignorant of the plaintiff's copyrighted program.

166. Once reverse engineering has been used, the defendant can no longer claim independent creation. Even if the final program is non-infringing, the process of reverse engineering itself could constitute copyright infringement. At the present time, reverse engineering is a risky business strategy. However, the judicial attitude toward reverse

Moreover, the basic fit between copyright protection and continuing innovation must be questioned. Even though the Supreme Court's decision in *Feist* breathed new life into the originality requirement, copyright makes little distinction between the protection afforded to trivial innovations and the protection given to major innovations. If courts adopt broad *Whelan*-style protection for semantic data models, then fairly trivial applications of object-oriented principles are likely to be granted strong protection.¹⁶⁷ This protection will not be offset by the benefits of disclosure of the innovation since commercial programs are distributed in object code form only and copyright registration can be obtained without disclosing the source code to the public. Since any object-oriented innovations occur at the source code level, the public gains no new knowledge from the grant of copyright. At most, the public gains access to a commercial product that might not have been created without the promise of copyright. On balance, copyright protection seems likely to stifle competition and discourage continuing innovation.

In contrast, while the patent examination process makes patent protection more difficult and costly to obtain, this process also addresses the primary deficiencies in copyright protection. First, the inventor must specifically define the scope of the software invention through technical claim language. This language is likely to be narrowed during the examination process in order to overcome prior art rejections. As a result, only highly innovative programs will be granted protection, and the scope of that protection will be sharply limited by prior art. Business competitors can then rationally plan competing products by performing patent searches and then determining how to design around existing patents.¹⁶⁸ If designing around an existing patent is not feasible, the precise definition provided by the claim language will make it easier for the parties to estimate the value of the patent and negotiate licenses. Finally, since patent protection is given only as the *quid pro quo* for full disclosure of the innovation, the costs of protection are offset by dissemination of the new technological knowledge behind the invention as well as by dissemination of the commercial product itself.

engineering may be changing,. Two appellate courts have recently applied the "fair use" doctrine to allow reverse engineering in certain contexts. See *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510, 1520 (9th Cir. 1992), amended, 1993 U.S. App. Lexis 78 (9th Cir. 1993); *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832 (Fed. Cir. 1992).

167. See, e.g., *Computer Assocs.*, 982 F.2d at 712 (noting that "serious students of the industry have been highly critical of the sweeping scope of copyright protection engendered by the *Whelan* rule, 'in that it enables first comers to "lock up" basic programming techniques as implemented in programs to perform specific tasks'" (quoting Menell, *supra* note 3, at 1087; citations omitted)).

168. Admittedly, business competitors still face some risk since patent applications are kept secret during the examination process.

Patent protection does present several drawbacks. First, the costs of obtaining protection or defending a potential infringement suit may deter some smaller companies from innovation. Second, the lack of an independent creation defense sharply increases the societal costs of protection by stifling innovation that would have occurred in the absence of patent protection. Finally, the seventeen-year term for the patent monopoly is excessively long given the short product cycles for most software projects. Nonetheless, if patent examiners utilize a broad range of relevant prior art both to reject applications that are not highly innovative and to limit the scope of patents actually granted, then patent protection is a more effective incentive for innovation than copyright protection.

COMMENT

THE EXPERIMENTAL USE EXCEPTION TO INFRINGEMENT APPLIED TO FEDERALLY FUNDED INVENTIONS

SUZANNE T. MICHEL [†]

Table of Contents

I.	INTRODUCTION	369
II.	BACKGROUND OF EXPERIMENTAL USE EXCEPTION	371
	A. Creation and Early Development.....	371
	B. The Federal Circuit	374
III.	PROBLEMS WITH COMMON LAW DOCTRINE	376
	A. Uncertainty for Universities and Federal Laboratories.....	376
	B. Foreclosure of New Inventions When A Basic Technique is Patented	386
	C. Polymerase Chain Reaction Example	387
IV.	PAST PROPOSALS AND CRITIQUE.....	388
	A. Proposals For a Broad Exception.....	388
	B. Incentives of the Patent System.....	391
	C. A Critique of the Broad Exception	394
V.	NEW PROPOSALS.....	397
	A. Non-profit Researchers Allowed Broad Exception	397
	B. Government-Funded Inventions Subject to a Broad Exception.....	400
VI.	CONCLUSION	409

I. INTRODUCTION

With one minor exception¹ the patent statutes do not suggest any instance in which use of a patented invention is not infringement. According to 35 U.S.C. § 154, "[e]very patent shall contain . . . a grant to the patentee . . . of the right to exclude others from making, using or

© 1993 Suzanne T. Michel.

[†] J.D. candidate 1993, School of Law (Boalt Hall), University of California at Berkeley; Ph.D. 1989, Yale University; B.S. 1984, Northwestern University.

1. 35 U.S.C. § 271(e) provides that it is not an act of infringement to make, use, or sell a patented invention for purposes reasonably related to obtaining FDA approval of drugs.

selling the invention throughout the United States." Section 271(a) provides that "whoever without authority makes, uses or sells any patented invention . . . infringes the patent."

In spite of the seemingly unyielding dictate of the statutes, courts have recognized experimental use as an exception to infringement. Use of a patented invention "for the mere purpose of philosophical experimentation, or to ascertain the verity and exactness of the specification" is exempt from infringement.² While it is well settled that a patented invention may be made and used to test the verity and exactness of the specification, the scope of the "philosophical experimentation" prong of the exception is much less clear. The Federal Circuit has called this prong "truly narrow."³ To be deemed philosophical experimentation, the experiment must be "for amusement, to satisfy idle curiosity, or for strictly philosophical inquiry."⁴ The exception does not "allow a violation of the patent laws in the guise of 'scientific inquiry,' when that inquiry has definite, cognizable and not insubstantial commercial purposes."⁵ Part II of this Comment describes the history and current scope of the experimental use exception.

In view of this narrow interpretation of the "philosophical experiment" prong of the experimental use exception, several commentators have called for a legislative broadening of the exception to encompass all activity short of commercialization.⁶ A House bill, the Research, Experimentation and Competitiveness Act of 1990, also proposed broadening the exception.⁷

Those proposing the broad exception point to two key problems which they contend the broad exception would either clarify or solve. First, it is unclear whether university and other non-profit research done under contract with industry or with a purpose to patent the results is "strictly for philosophical inquiry." The uncertain limits of the doctrine might chill research or lead to litigation. Second, when a patent owner controls important information, that control might prevent a subsequent researcher from building on the information in a way that benefits society. The broad exception would allow subsequent research on patented inventions and would clarify the position of non-profit

2. Sawin v. Guild, 21 F. Cas. 554, 555 (C.C.D. Mass. 1813) (No. 12,391).

3. Roche Prods., Inc. v. Bolar Pharmaceutical Co., 733 F.2d 858, 863 (Fed. Cir.), cert. denied, 469 U.S. 856 (1984).

4. *Id.*

5. *Id.*

6. Rebecca S. Eisenberg, *Patents and the Progress of Science: Exclusive Rights and Experimental Use*, 56 U. CHI. L. REV. 1017 (1989); Ned A. Israelsen, *Making, Using, Selling Without Infringing: An examination of 35 U.S.C. Section 271(e) and the Experimental Use Exception to Patent Infringement*, 16 AM. INTELL. PROP. L. ASS'N Q.J. 457 (1989).

7. H.R. 5598, 101st Cong., 1st Sess. §§ 401-403 (1990).

researchers. Part III describes the conditions which caused these two problems.

The task at hand is to find the wisest limits for the exception while providing a workable solution to the problems of foreclosed research and the uncertain position of non-profit researchers. Any proposal must take into account the economics and incentives of the patent system. Part IV critiques the wisdom of the proposals for a generally applicable broad exception. Part IV also argues that a generally applicable broad experimental use exception weakens the incentives to invent, to develop and to disclose provided by the patent system to too great an extent when applied to patents resulting from private research efforts.

Instead, this Comment proposes in Part V that the experimental use exception (extending up to commercialization) be made applicable only in the special circumstances in which its harm to patent incentives is minimal compared to the resulting benefits. First, university and other non-profit researchers should be allowed the advantage of the broad exception. This first proposal clarifies the position of non-profit researchers with minimal harm to the patent holder. Second, any party should be allowed to use a patented, federally funded invention in research and development. This second proposal provides a number of benefits without the disincentives which result when a broad experimental use exception is applied to privately funded patents. For instance, federally funded inventions will not foreclose subsequent research, but federal grantees will not lose their incentive to invent and disclose because those incentives come from outside the patent system.

II. BACKGROUND OF EXPERIMENTAL USE EXCEPTION

Understanding how the critique and proposals presented by this Comment fit into the framework of the patent laws first requires understanding the judicially created experimental use exception.

A. Creation and Early Development

The experimental use doctrine as a defense to patent infringement originated in 1813 in *Whittemore v. Cutter*, an opinion written by Justice Story while sitting on the Massachusetts Circuit Court.⁸ The defendant in that case challenged a jury instruction that "the making of a machine fit for use, and with a design to use it for profit, was an infringement of the patent right."⁹ Justice Story approved the instruction on the grounds that "it could never have been the intention of the legislature to punish a man, who constructed such a machine merely for philosophical experiment, or

8. *Whittemore v. Cutter*, 29 F. Cas. 1120 (C.C.D. Mass. 1813) (No. 17,600).

9. *Id.* at 1121.

for the purpose of ascertaining the sufficiency of the machine to produce its described effects."¹⁰

Justice Story referred to this exception again in *Sawin v. Guild*.¹¹ In holding that the defendant's use of patented machines constituted patent infringement, he noted that the machines had been used for profit rather than "for the mere purpose of philosophical experimentation, or to ascertain the verity and exactness of the specification. . . . In other words, that the making must be with an intent to infringe the patent-right, and deprive the owner of the lawful rewards of his discovery."¹² Even though experimental use was not an issue in either case, meaning that the exception originated in dicta, by 1861 the law on this subject was deemed "well-settled."¹³

Very few early cases applied the experimental use doctrine created by Justice Story to excuse use of a patented invention that would otherwise constitute infringement.¹⁴ Even so, the second prong of Justice Story's test which allows activity for "ascertaining the verity and exactness of the specification" does appear to be "well settled." A party may wish to challenge a patent as invalid for not being enabling or useful and therefore must use the invention without a license to assemble proof of this invalidity. A party may also wish to test a patent before taking a license. Although there is little case law on the point, most commentators agree that this sort of activity is and should be protected by the exception.¹⁵

The scope of the "philosophical experiment" prong is much less clear. The cases that applied this prong simply concluded that the use in question was "experimental" without offering an elaboration of that term.¹⁶ The commercial character of a use or the commercial intent of a user usually forfeited the protection of the doctrine in other early cases.¹⁷ Overall, these early cases provide little guidance in setting the contours of the exception today.

Two more recent cases developed the "philosophical experiment" prong more fully, but neither found the doctrine to be applicable. In *Pitcairn v. United States*, the Court of Claims considered whether

10. *Id.*

11. 21 F. Cas. 554 (C.C.D. Mass. 1813) (No. 12,391).

12. *Id.* at 555 (citation omitted) (citing *Whittemore*).

13. *Poppenhusen v. Falke*, 19 F. Cas. 1048, 1049 (C.C.S.D.N.Y. 1861) (No. 11,279).

14. The history of the experimental use exception from its creation to its application by the Federal Circuit is described elsewhere. See Ronald D. Hantman, *Experimental Use as an Exception to Patent Infringement*, 67 J. PAT. & TRADEMARK OFF. SOC'Y 617 (1985). Accordingly, this Comment presents only a summary.

15. Eisenberg, *supra* note 6, at 1074.

16. See Israelsen, *supra* note 6, at 460 n.11.

17. See *id.* at 460 n.14.

helicopters produced under contract for the United States infringed patents that had been previously declared valid by that court.¹⁸ The court rejected the government's argument that the helicopters were purchased for testing and experimental purposes and therefore did not infringe.¹⁹ The court held that "[t]ests, demonstrations and experiments of such nature are intended uses of the infringing aircraft manufactured for the defendant and are in keeping with the legitimate business of the using agency."²⁰ The helicopters were not built solely for experimental purposes and thus were excluded from the exception.

In *Pfizer, Inc. v. International Rectifier Corp.*, a federal district court held International Rectifier (IR) in contempt of court for violating an injunction which ordered IR to cease manufacture, use and sales of doxycycline, a pharmaceutical compound patented by Pfizer.²¹ In spite of the injunction, IR had continued to manufacture doxycycline in order to conduct various tests such as bioequivalency and serum level tests.²² IR also shipped doxycycline to laboratories in and out of the United States accompanied by a notice that the compound constituted laboratory samples for experimental purposes only.²³

IR defended its activities on the grounds that they were solely experimental, and that the compound was never sold in the United States after the injunction. The court held these arguments to be "utterly without merit."²⁴ The court interpreted the history of the experimental use doctrine to suggest that "the underlying rule of permissible experimental use demands there must be no intended commercial use of the patented article, none whatsoever, if the exception is to be recognized at all."²⁵ Because IR's activities were for the purpose of competing with Pfizer after its patent expired, the court held IR in contempt. In addition, the court ordered IR to destroy all the doxycycline it possessed as well as all data it illicitly acquired regarding doxycycline.²⁶

Both *Pitcairn* and *Pfizer* make clear that when a use is consistent with the "legitimate business" of the infringer or has an ultimate commercial purpose, the use is not "philosophical experimentation" and falls outside of the exception.

18. *Pitcairn v. United States*, 547 F.2d 1106 (Ct. Cl. 1976), *cert. denied*, 434 U.S. 1051 (1978).

19. *Id.* at 1124-25.

20. *Id.* at 1125-26.

21. *Pfizer, Inc. v. International Rectifier Corp.*, 217 U.S.P.Q. (BNA) 157 (C.D. Cal. 1982).

22. These tests are required for FDA approval of generic drugs.

23. *Id.* at 158-59.

24. *Id.* at 160.

25. *Id.* at 161.

26. *Id.* at 163.

B. The Federal Circuit

In *Roche Products, Inc. v. Bolar Pharmaceutical Co.*, the only Federal Circuit²⁷ case discussing at length the scope of the experimental use doctrine, the court interpreted the doctrine narrowly.²⁸ Bolar had imported five kilograms of Roche's patented compound flurazepam hydrochloride which Roche sold as a sleeping pill, Dalmane. Bolar used the compound to conduct the bioequivalency studies required for FDA approval with an eye toward marketing a generic version of the drug when Roche's patent expired a year later. Roche argued that this use constituted infringement, but the district court held that the use of a patented drug for testing related to FDA drug approval during the last six months of the patent term was *de minimis*, experimental and noninfringing.²⁹

The Federal Circuit reversed, calling the experimental use exception "truly narrow."³⁰ The court's analysis first addressed the statute, noting that "[s]ection 271(a) prohibits, on its face, any and all uses of a patented invention," but admitted that the definition of "use" is a matter of judicial interpretation. The court cited *Pitcairn* for both the proposition that experimental use may be a defense to infringement and as setting forth the controlling law.³¹ The court quoted *Pitcairn*'s statement that "[t]ests, demonstrations, and experiments . . . [which] are in keeping with the legitimate business of the . . . [alleged infringer]" are infringements for which '[e]xperimental use is not a defense.'³²

Bolar did not come within the exception because its use was "not for amusement, to satisfy idle curiosity, or for strictly philosophical inquiry."³³ The court explained:

[U]nlicensed experiments conducted with a view to adaption of the patented invention to the experimenter's business is a violation of the rights of the patentee to exclude others from using his patented invention. . . . We cannot construe the experimental use rule so broadly as to allow a violation of the patent laws in the guise of "scientific inquiry," when that inquiry has definite, cognizable and not insubstantial commercial purposes.³⁴

27. The Federal Circuit, established in 1982, has jurisdiction over all appeals in cases "arising under" the federal patent laws. 28 U.S.C. § 1295 (1988).

28. *Roche Prods., Inc. v. Bolar Pharmaceutical Co.*, 733 F.2d 858 (Fed. Cir.), *cert. denied*, 469 U.S. 856 (1984).

29. *Roche Prods., Inc. v. Bolar Pharmaceutical Co.*, 572 F. Supp. 255 (E.D.N.Y. 1983), *rev'd*, 733 F.2d 858 (Fed. Cir.), *cert. denied*, 469 U.S. 856 (1984).

30. *Roche*, 733 F.2d at 863.

31. *Id.* at 861, 863.

32. *Id.* at 863 (quoting *Pitcairn v. United States*, 547 F.2d 1106, 1125-26 (Ct. Cl. 1976), *cert. denied*, 434 U.S. 1051 (1978)).

33. *Id.*

34. *Id.*

Nor did the court consider the use *de minimis* even though the quantity used was small, because the testing could have had a significant economic impact on Roche if Bolar released the generic drug on the market earlier than it would have absent the infringement.³⁵

1. THE OVERRULING OF ROCHE V. BOLAR

Shortly after the *Roche v. Bolar* decision, Congress passed the Drug Price Competition and Patent Term Restoration Act of 1984³⁶ which legislatively overruled that decision. That law exempts from infringement activity which is "reasonably related" to seeking FDA approval for a generic drug. The portion of the bill codified as 35 U.S.C. § 271(e)(1) states that "[i]t shall not be an act of infringement to make, use, or sell a patented invention . . . solely for purposes reasonably related to the development and submission of information under a federal law which regulates the manufacture, use or sale of drugs."

The scope of the exemption is fairly narrow. The legislative history indicates that only a limited amount of testing to establish the bioequivalency of a generic drug substitute is permitted.³⁷ Whether an activity is "reasonably related" to seeking FDA approval has been narrowly interpreted in the case law.³⁸

The legislation is interesting because it demonstrates a Congressional attitude which is willing to allow exceptions to infringement under some circumstances. The committee report states that the exemption did not substantially interfere with the rights of the patent holder because "[t]he patent holder retains the right to exclude others from commercial markets during the life of the patent."³⁹ In spite of this statement, Congress concurrently enacted a law which extended the patent grant for human drugs and other products which must undergo federal approval before marketing to compensate patentees for the time lost in which they can monopolize the market.⁴⁰ Patent owners essentially receive an extension of the patent term in exchange for their toleration of infringing use which enables a competitor to market a product as soon as the pertinent patent expires. This trade-off implies

35. *Id.* at 866.

36. 35 U.S.C. § 271(e) (1988).

37. H.R. REP. NO. 857 pt. 2, 98th Cong., 2d Sess. 8 (1984), reprinted in 1984 U.S.C.C.A.N. 2686, 2692.

38. *Scripps Clinic v. Genentech, Inc.*, 666 F. Supp. 1379, 1396 (N.D. Cal. 1987) (a multiple purpose use of a patented invention is not exempted where only one purpose is reasonably related to FDA testing). However, the Supreme Court's decision in *Eli Lilly v. Medtronics* affirms a Federal Circuit decision to extend the scope of 271(e) to include the testing of medical devices. 872 F.2d 402 (Fed. Cir. 1989), *aff'd*, 496 U.S. 661 (1990).

39. H.R. REP. NO. 857 pt. 2, *supra* note 37, at 8, reprinted in 1984 U.S.C.C.A.N. at 2692.

40. *Id.* at 14, reprinted in 1984 U.S.C.C.A.N. at 2691; 35 U.S.C. § 156 (1988).

that Congress may have conflicting views as to whether the harm to the patentee caused by the exempted experimental testing is as insubstantial as the legislative history suggests. The nature of an experimental use exception's interference with the patent right is discussed below in Section IV.C.

III. PROBLEMS WITH COMMON LAW DOCTRINE

Although the holding of *Roche* was overruled through legislation, that case is illustrative of the Federal Circuit's attitude toward the experimental use exception as a defense to infringement. The rationale of *Roche* remains the common law of experimental use in contexts other than the limited conditions of section 271(e). Given the narrow limits which that case places on the doctrine, any activity with a long-range profit motive or with any profit potential is unlikely to fall within the exception. Corporate research will nearly always be "in keeping with the legitimate business of the alleged infringer."

This narrow interpretation of the "philosophical experiment" prong of the experimental use exception engenders two related problems. First, it is unclear whether university research done under contract with industry or with a purpose to patent the results is "strictly for philosophical inquiry." Second, when a patent owner controls important information, that control might prevent a subsequent researcher from building on the information in a way that benefits society. The uncertain limits of the doctrine might chill research or lead to unnecessary litigation.

A. Uncertainty for Universities and Federal Laboratories

The extent to which use of a patented invention is permissible noninfringing experimentation when conducted by nonprofit researchers such as universities and federal labs remains unclear. Only one 1935 case, *Ruth v. Stearns-Roger Manufacturing Co.*, has addressed the issue of whether university use can be infringement. The district court in that case held that use of an infringing machine by the Colorado School of Mines was experimental and exempt from infringement because the machines were used in a laboratory and were cut up and changed from day to day. The school used the machines in furtherance of its educational purpose.⁴¹

Whether all research conducted in universities and federal laboratories today can be categorized as "philosophical experiments" is extremely problematic given the Federal Circuit's narrow interpretation

41. *Ruth v. Stearns-Roger Mfg. Co.*, 13 F. Supp. 697, 703 (D. Colo. 1935), *rev'd on other grounds*, 87 F.2d 35 (10th Cir. 1936).

of that term. To understand how university research, which would appear to epitomize "philosophical experimentation," could fall outside the exception, we must examine the trend toward patenting and licensing university research and the relationships universities have forged with industry. It is through the universities' own attempts to monopolize research results and collaborate with the commercial sector that they have potentially lost claim to the experimental use exception. Part I through Part III below describe the current landscape of industry-university, industry-federal laboratory relationships. Part IV explains why these relationships make application of the experimental use exception uncertain.

1. PATENTING AND LICENSING BY UNIVERSITIES

Prior to the 1980 and 1984 amendments to the patent laws, patents resulting from federally funded research belonged to the government, who often licensed them on a royalty-free, non-exclusive basis, although policies varied depending on the granting agency.⁴² The government had a poor record for advancing the development of its patents. For instance, in 1976, less than four percent of the twenty-eight thousand patents held by the federal government were commercially developed.⁴³

The perceived need for the 1980 and 1984 amendments was prompted in part by the concern that federally funded research was not being efficiently commercialized because a company wishing to use that research confronted "a bewildering array of 26 different sets of agency regulations governing their rights to use such research."⁴⁴ In response, the amendments created a single, uniform national policy. Non-profit research institutions and small businesses now retain the rights to patents resulting from federally funded research which they perform. The amendments also give universities the right to own inventions made in federally owned research facilities run by the university under contract with the government.

The amendments encourage government-funded researchers to patent resulting inventions by simplifying the bureaucratic obstacles to licensing and by allowing the patent holder to keep the royalties.⁴⁵ Private industrial firms can exclusively license these patents from the

42. James A. Dobkin, *Patent Policy in Government Research and Development Contracts*, 53 VA. L. REV. 564 (1967) (describing policies of the AEC, NASA, the FAA, the Department of Defense, and the Department of Health, Education and Welfare).

43. S. REP. NO. 480, 96th Cong., 1st Sess. 2 (1979).

44. H.R. REP. NO. 1307, 96th Cong., 2d Sess. 2 (1980), reprinted in 1980 U.S.C.C.A.N. 6460, 6461.

45. See *id.* at 5, reprinted in 1980 U.S.C.C.A.N. at 6464.

university or another government contractor for specific uses they intend to commercialize.⁴⁶

Congress designed the amendments to encourage private industry to commit the capital necessary to develop government-funded inventions to the point of commercial application. Supporters of the amendments argue that without the profit incentive provided by exclusive rights, commercial development lags and research results do not become socially useful. The Secretary of Commerce stated, "Direct access to the university and the university's right to transfer the results of its research on an exclusive basis is an important incentive for business to invest in the further development and commercialization of new technologies."⁴⁷

Thus, the patent system accomplishes the policy goal of transferring the products of university research to the public by allowing a university to license its inventions.⁴⁸ However, the license must be exclusive before companies will invest in development. Inventions arising from university research are often at an early stage of development and the licensee may need to do further development simply to identify a commercial product. Because biotechnology products in particular require expensive regulatory approval, it is difficult to find a licensee who is willing to make the required investment without receiving an exclusive license.⁴⁹

In the past the university scientific community viewed private ownership of discoveries as contrary to the university's mission and the public interest.⁵⁰ Especially in the biomedical fields, some researchers held a belief that new knowledge should be made as widely available as possible to serve humanity.⁵¹ This attitude has changed for several reasons, making universities increasingly likely to patent publicly and privately funded research.

First, the view that basic research should be freely available to everyone was predicated on the assumption that the work being done

46. 35 U.S.C. § 202(c)(7) (1988). The government retains a royalty-free worldwide license to practice the invention or have it practiced for the government. *Id.* § 202(c)(4). In addition, the government has march-in rights that terminate the rights of the contractor if the contractor does not effectively attempt to apply the invention. *Id.* 35 U.S.C. §§ 202(c)(8), 203.

47. S. REP. NO. 662, 98th Cong., 2d Sess. 4 (1984), reprinted in 1984 U.S.C.C.A.N. 5799, 5803.

48. Phyllis S. Lachs, *University Patent Policy*, 10 J.C. & U.L. 263, 276-77 (1983). Of course, this argument assumes that the private sector would not commercially develop the university invention absent an exclusive license.

49. See DAVID DICKSON, THE NEW POLITICS OF SCIENCE 91 (1984); Joyce Brinton, *Biotechnology Licensing: Issues from the University Perspective*, 16 AM. INTELL. PROP. L. ASS'N Q.J. 479, 484 (1988).

50. DICKSON, *supra* note 49, at 89-90; BERNARD BARBER, SCIENCE AND THE SOCIAL ORDER 130 (1952).

51. MARTIN KENNEY, BIOTECHNOLOGY: THE UNIVERSITY-INDUSTRIAL COMPLEX 32 (1986).

had no immediate commercial value. When this premise broke down in fields like molecular genetics, and laboratories produced results with commercial value, various entrepreneurial interests insisted that results be privatized.⁵² Consequently, patent protection for basic research discoveries with potential commercial value has become more commonplace. This is especially true in biotechnology-related fields where the dividing line between basic and applied research is not clear.⁵³ Academic and industrial scientists often work on the same or closely related problems.⁵⁴

Second, universities had little incentive to pursue patent rights before the 1980 amendments because the common practice of government agencies supporting the research was to require that the patent be assigned to the government and then freely licensed.⁵⁵

Because the amendments allow the universities to keep royalties, they are looking to licenses as a way to supplement government money for research. As government support of university research has decreased in terms of constant dollars, the cost of scientific research has rapidly escalated.⁵⁶ Erich Bloch, then director of NSF, testified before a Senate Committee that the federal government is unable to meet all research needs of the universities and, therefore, the universities have a continuing need for additional funding.⁵⁷

Allowing universities to patent and license faculty inventions has produced a number of success stories for different universities. The Cohen-Boyer gene-splicing patent which forms the basis of the biotechnology industry is expected to bring more than \$100 million in royalties to the University of California and Stanford.⁵⁸ The Massachusetts Institute of Technology registers more patents than any other university, over one hundred per year, and licenses up to 53% of

52. *Id.* at 107.

53. DICKSON, *supra* note 49, at 75-76.

54. *Id.* at 74-75; see David Blumenthal et al., *Industrial Support of University Research in Biotechnology*, 231 SCIENCE 242 (1986).

55. Dobkin, *supra* note 42, at 568-84, 591-607.

56. Lachs, *supra* note 48, at 268.

57. *National Science and Technology Issues: Hearing Before the Senate Committee on Commerce, Science and Transportation*, 101st Cong., 2nd Sess. 22 (1990) (statement of Erich Bloch, Director, National Science Foundation) [hereinafter *Technology Issues Hearing*]. For instance, the NIH budget has been rising rapidly, from \$3.2 billion in 1980 to \$7.5 billion in 1990. However, the soaring cost of doing research, the fact that more money is tied up in long-term grants, and the increasing number of scientists applying for grants have created a money drought, especially for younger scientists. NIH research grants account for about 75% of all biomedical research funds provided by the federal government and private nonprofit sources. Gina Kolata, *Beginning Scientists Face a Research Fund Drought*, N.Y. TIMES, June 5, 1990, at C1.

58. Marjorie Shaffer, *All About University Patents: When Research Labs Go After Business*, N.Y. TIMES, Feb. 23, 1992, § 3, at 10; see also DICKSON, *supra* note 49, at 90.

those. In 1991, M.I.T. grossed \$5.5 million from its licensing activities.⁵⁹ Some forty companies employing more than one thousand people have been started based on M.I.T.-licensed technology.⁶⁰

Third, the requirement of the 1980 amendments that universities share royalties with inventors gives researchers an incentive to be alert to patent rights.⁶¹ Universities generally include a patents rights clause in employment contracts with faculty so that the patent must be assigned to the university. Often the university awards between one third and one half of any resulting royalty to the inventor, with the remainder going to the university.⁶² Consequently, the inventor profits from any licensing.

2. UNIVERSITY-INDUSTRY RELATIONSHIPS

For universities, patents provide more than just royalty income. Patents are also a means of strengthening ties with industry and gaining private support for academic research.

Universities are contracting with industry to conduct specific research with the understanding that the industrial firm receives the right to license and commercially develop the results.

In the past, university-industry agreements were generally of a small scale and seldom controversial.⁶³ The situation began to change in the mid-1970s at a time when universities experienced economic pressures from rising operating costs coupled with federal funding that failed to keep pace with the expanding number of scientists. In this atmosphere, university faculty and administrators welcomed increased collaboration with and funding from industry.⁶⁴ Industrial support of academic research made up only 3.8% of the total university research budget in 1980 but has been generally increasing since then.⁶⁵ A 1984 study reveals that industry may be funding as much as one fourth of all biotechnology research in universities.⁶⁶ Fueling industry's increased

59. M.I.T. netted only \$500,000 from the \$5.5 million it grossed from royalty licenses in 1991 due to the costs associated with filing and licensing patents and the \$1 million it distributed to hundreds of individual scientists. Shaffer, *supra* note 58, at 10.

60. *Id.*

61. 35 U.S.C. § 202(c)(7)(C) (1988).

62. Lachs, *supra* note 48, at 281, 285-86 (recommending that universities include a patent rights clause in their employment contracts).

63. DOROTHY NELKIN, SCIENCE AS INTELLECTUAL PROPERTY 18 (1984).

64. *Id.* Universities are partially motivated to accept the corporate sponsorship in order to keep their best scientists, who may move to another university or to industry if denied the corporate funding. KENNEY, *supra* note 51, at 62.

65. KENNEY, *supra* note 51, at 35; NELKIN, *supra* note 63, at 18, 23. Industry contributed \$667 million to university research in 1986. Gretchen Morgenson, *In Pecunia Veritas?*, FORBES, Nov. 1988, at 204, 208.

66. Blumenthal et al., *supra* note 54, at 244.

funding is the fact that the gap between basic research and commercial interest has "dramatically narrowed."⁶⁷

University-industry relationships take a variety of forms ranging from contributions⁶⁸ to contract research to long-term agreements. Contract research requires that a university laboratory conduct specific experiments, such as testing the safety of new chemical, for a fee. In another type of contract arrangement, the university provides continuing education for the industrial personnel.⁶⁹

In some long term university-industry agreements the corporation directs a grant to a specific investigator. In others, a review board consisting of university and corporate members selects the projects to be funded. Alternatively, a research institute may be organized to be distinct from the university.⁷⁰

Businesses see the long-term university-corporate contract as a means to purchase access to university scientists. In exchange for funding, the corporation receives the research skill of the principal investigator and her entire laboratory staff, including graduate students and postdoctoral researchers. The company often expects to be intimately familiar with ongoing research, giving the company tremendous opportunities for access to the researcher's ideas.⁷¹

Companies fund university research to enhance their competitive position and this requires that research results be patented.⁷² The 1980 and 1984 patent amendments are a cornerstone of university-industry cooperation.⁷³ (Since projects may be both publicly and privately funded, the amendments will apply.) Cooperative arrangements between an industry sponsor and a university usually provide that the sponsor acquires either patent rights, patent ownership or, more commonly, an

67. DICKSON, *supra* note 49, at 74.

68. A corporation may simply make a contribution to the university, either an undirected contribution which the university may use as it sees fit or a directed contribution which targets a substantive research area or supports a specific professor. These relatively small grants do not provide the corporation with any unique claim on research results though they may provide a personal relationship with the professor. KENNEY, *supra* note 51, at 37-38.

69. *Id.* at 39-40.

70. Lachs, *supra* note 48, at 279.

71. KENNEY, *supra* note 51, at 68, 240. The long-term one university-one corporation agreement has been more common in biotechnology than in other fields. The early development of biotechnology took place in universities rather than commercial laboratories and a significant portion of cutting-edge research remains in universities. *Id.* at 55. For example, in 1982 Yale University announced a \$1,100,000 research contract with Celanese Corporation which mandated that Yale conduct specific research for the company on the composition and synthesis of enzymes. Lachs, *supra* note 48, at 280. Other specific agreements are described elsewhere. See, e.g., Kenney, *supra* note 51, at 58-72.

72. *Id.* at 61.

73. DICKSON, *supra* note 49, at 89.

exclusive license, to all inventions made in connection with the project agreement.⁷⁴

Universities are not always on the receiving end in a university-industry relationship. Universities like Harvard, Johns Hopkins and the University of Texas have formed or invested in for-profit venture capital companies to develop their researchers' work. Universities may also "accept stock in startup companies instead of a licensing fee," hoping to "benefit when a company becomes successful."⁷⁵

University-industry relationships may also involve only individual faculty members, rather than the university as an entity. Faculty consulting⁷⁶ for corporations can grow into a more intimate relationship, such as a position on a scientific advisory board for which the advisor receives stock. An important factor in convincing professors to affiliate with a company has been the provision of equity interest, giving the professor a stake in the success of the company.

Professors may also start firms in partnership with venture capital. While retaining their faculty status, university faculty have acted as founders and consultants for new ventures hoping to commercialize the practical applications of their research.⁷⁷ In some cases professors who are heavily involved in company management leave the university completely, especially in the electrical engineering and computer science fields.⁷⁸ Many biology professors, however, retain their university connection. All of the earliest genetic engineering companies were founded by professors who completed the initial research in university laboratories. For instance Genentech, co-founded by Herbert Boyer to exploit the Cohen-Boyer gene splicing patent, did not have a laboratory in its early stages, so Boyer's campus laboratories at UCSF were used. The company granted Boyer \$200,000 to perform specific research.⁷⁹

3. FEDERAL LABORATORY-INDUSTRY RELATIONSHIPS

The 1984 amendment to the patent laws works with the Federal Technology Transfer Act (FTTA) to promote a greater flow of technology from federal laboratories to the private sector by giving the laboratories the authority and incentive to work with the private sector. The 1984

74. KENNEY, *supra* note 51, at 58, 63, 64, 65.

75. Shaffer, *supra* note 58, at 10.

76. KENNEY, *supra* note 51, at 101-03. Faculty consulting for government and corporations has long been a part of the academic world. Within the traditional limits which dictate that consulting should not interfere with academic responsibilities, consulting alone has had little effect upon the university. *Id.* at 91.

77. Brinton, *supra* note 49, at 480; see also Lachs, *supra* note 48, at 289-90.

78. KENNEY, *supra* note 51, at 100.

79. *Id.* at 94-95. For a list of professors having both university and lucrative industry connections, see Morgenson, *supra* note 65, at 208.

amendments require federal laboratories to identify and seek patents for inventions with commercial potential and authorize laboratories to negotiate and issue patent licenses on those inventions.⁸⁰ The FTTA establishes a mechanism and offers incentives for government laboratories to enter into cooperative research and development agreements (CRADAs) with industry. The Act allows the over 700 federal laboratories to accept funds, services, and property from a private firm in exchange for an exclusive license to any patent rights resulting from the cooperation.⁸¹ In addition, federal employees whose inventions are commercially developed receive a percentage of the royalties.⁸²

Federal laboratories had previously been generally unsuccessful in transferring research results to the private sector. Although an agency supporting research in federal laboratories could allow a private company to develop a patented invention, agencies were often so slow in doing so that companies lost interest. For example, between 1977 and 1985, the Department of Energy received 135 requests to use patent rights to inventions made at contractor-operated facilities, but by December 1985 had responded to only fifty-five of them.⁸³

The importance of federal laboratories having the authority to grant exclusive licenses was emphasized by Ronald Hart, director of the National Center for Toxicological Research. He stated that the inability to grant exclusive licenses in the past has meant the research "was everybody's property and so nobody's product." Hart contends that "many inventions that could have improved public health simply languished."⁸⁴

4. UNCERTAINTY UNDER THE EXPERIMENTAL USE EXCEPTION

The activities and relationships described above raise the issue of what circumstances constitute infringement when a university researcher uses a patented invention without a license. When will that use be exempted as "philosophical inquiry," and when will that use be deemed commercial so that the use falls outside the narrow experimental use

80. 35 U.S.C. § 207 (1988); S. REP. NO. 662, *supra* note 47, at 4, 6, reprinted in 1984 U.S.C.C.A.N. at 5802, 5804.

81. Federal Technology Transfer Act of 1986, Pub. L. No. 99-502, 100 Stat. 1785 (1986); 15 U.S.C. § 3710(a) (1988); 35 U.S.C. § 207 (1988); S. REP. NO. 662, *supra* note 47, at 6-10, reprinted in 1984 U.S.C.C.A.N. at 5804-08; *Technology for Sale*, SCI. AM., May 1987, at 62.

82. 15 U.S.C. § 3710(b) (1988); *Technology for Sale*, *supra* note 81. When Oak Ridge National Laboratory was designated as a guinea-pig laboratory in 1984 to test the incentives, laboratory employees sought 30% more patent applications in two years. *Id.*

83. S. REP. NO. 662, *supra* note 47, at 5, reprinted in 1984 U.S.C.C.A.N. at 5803; *Technology for Sale*, *supra* note 81.

84. *Technology for Sale*, *supra* note 81.

exception? (Research in federal laboratories raises the same concerns and difficulties as research in universities and is not referred to separately.)

In some cases, non-licensed research use of a patented invention will lead to another invention which the university may patent and license. By generating funds from licensing, the university imparts a profit motive to the research work. Of course, the university can argue that because any license fees it receives are not distributed as profits, but instead are used to support its educational and research missions, the use of the original patented invention was an exempted experimental use.

But when industry licenses a university owned invention with a view towards commercializing the invention, the university research takes on a "definite, cognizable and not insubstantial commercial purposes."⁸⁵ Although the university does not itself participate in the commercialization, the subsequent industry use taints the university research with a commercial purpose. University research most likely loses its status as an experimental use once the results of that research are licensed to a commercial organization. This is especially true when an industrial firm contracts with a university with an understanding that the industrial sponsor will receive an exclusive license on any patented results. Any different conclusion would allow industry to easily circumvent the patent laws when it wanted to conduct research on a patented invention by contracting with a university lab to do the work.

If universities insist that industrial firms license their patented inventions, even for experimental purposes, it will not be surprising if industrial firms eventually demand equal, reciprocal treatment for their own inventions. Thus, industrial firms may begin raising the issue of whether university use is truly experimental.

Universities apparently do expect industrial firms to license their technologies for experimental use. To generate royalties, Joyce Brinton, Director of Harvard University's Office for Patents, Copyrights and Licensing, suggests a licensing strategy which clearly contemplates that industry must license technology it wishes to research, even when no specific commercial application has yet been identified. Brinton suggests that the university set financial terms that encourage licensees to experiment with the invention. For example, Brinton suggests that the university could charge a small initial fee for time-limited research followed by a larger fee to continue R&D. The hope is that a licensee will find the invention useful and sufficiently inexpensive so that the company will not search for an alternative.⁸⁶

85. Roche Prods., Inc. v. Bolar Pharmaceutical Co., 733 F.2d 858, 863 (Fed. Cir.), *cert. denied*, 469 U.S. 856 (1984).

86. Brinton, *supra* note 49, at 487.

In another example, Stanford announced in 1985 that seventy-three companies had taken licenses under the Cohen-Boyer patents⁸⁷ for recombinant DNA technology. As of that time however, only about ten companies had reportable sales using the technology. It logically follows that a substantial number of licensees were using the technology on an experimental basis.⁸⁸

Scientists working in a university or federal laboratory setting are likely to be concerned with both the traditional rewards of the scientific community and the incentives provided by the patent system.⁸⁹ The patent incentive may cause individual scientists to choose experiments with a profit motive rather than strictly for philosophical inquiry. This individual profit motive can also cause the research in question to lose its status as an experimental use.

By profiting directly from any licensing of their inventions,⁹⁰ individual university scientists may lose their status as pure experimental researchers entitled to the exception. But the profit motive may have an even deeper effect on research choices. It is now accepted practice for university researchers to profit directly from the results of academic research through various types of commercial ventures. University scientists may have equity interests in companies sponsoring or developing their work.⁹¹ In this case, the scientist's and the company's objectives will be closely aligned, tempting the scientist to ensure the company's success by tilting her academic research agenda. This can lead to a conflict of interest for the researcher. For instance, the scientist may use students and university equipment for private gain, divide work time in a way that slights the university, shift research to commercial goals, transfer patent rights from the university to the company and suppress research results.⁹² Certainly, this type of activity has a commercial purpose and cannot be characterized as an experimental use.

As research use of patented inventions by non-profit institutions becomes more of a threat to the interests of patent holders, the patentees may either demand licensing fees or attempt to enjoin the research. For instance, Johnson & Johnson sent letters to researchers at universities and government laboratories warning them that research use of the company's patented cells may infringe the company's patent rights. Johnson & Johnson's position was that using the patented invention to

87. Hantman, *supra* note 14, at 643.

88. *Id.*

89. Rebecca S. Eisenberg, *Proprietary Rights and the Norms of Science in Biotechnology Research*, 97 YALE L.J. 177, 195 (1987).

90. The amendments to the patent laws require that the inventor receive a share of the royalties when any university invention is licensed. 35 U.S.C. § 202(c)(7)(C) (1988).

91. NELKIN, *supra* note 53, at 20; Blumenthal et al., *supra* note 54.

92. KENNEY, *supra* note 51, at 112-13.

make a commercial improvement would be an infringement, even if made in a university.⁹³ The licensing and contracting activities described above increase the uncertainty of the outcome of this type of threat to the universities and the federal labs.

B. Foreclosure of New Inventions When A Basic Technique is Patented

Occasionally the subject of a patent is a basic technique or piece of information which lays the foundation for further discovery and developments. The consequences of patenting discoveries that are significant to both basic and applied research are not well understood.⁹⁴ Patents present a difficult dilemma for the progress of science. Patentees can retard further scientific progress by prohibiting the use of their patented invention in subsequent research.⁹⁵ For instance, patent holders can prohibit the research use of their inventions in order to prevent subsequent researchers from developing non-infringing substitutes which would of course undermine the value of the patent.

Subsequent researchers might obtain licenses, but as explained below, this is problematic in a non-profit research institution conducting basic research.⁹⁶ Of course, holders of patents on basic techniques might not enforce their rights against subsequent researchers. The patent holder may not be aware of the research or may not object.⁹⁷ Furthermore, the amount of damages available from research-oriented infringement may make a suit financially unsound, especially if the research does not threaten the commercial interests of the patent holder.⁹⁸

This problem is related to the uncertain situation in which non-profit researchers find themselves. A basic technique is precisely the type of invention that a university researcher would wish to investigate further. Again, the patent holder may not object as suing a university would undoubtedly create adverse publicity. However, the university is in an uncertain situation due to the confused scope of the experimental use exception, especially if it wishes to patent and license the results of the subsequent research.

93. Jeffrey L. Fox, *Patents Encroaching on Research Freedom*, 224 SCIENCE 1080 (1984).

94. Rebecca S. Eisenberg, *Patenting the Human Genome*, 39 EMORY L.J. 721, 740 (1990).

95. *Id.* at 742; Eisenberg, *supra* note 89, at 177.

96. See *infra* notes 147-50 and accompanying text.

97. Eisenberg, *supra* note 94, at 742-43.

98. *Id.* at 743. In *Roche*, however, the Federal Circuit noted that the financial harm to the patentee from even the minimal use at issue there would not be inconsequential. *Roche Prods., Inc. v. Bolar Pharmaceutical Co.*, 733 F.2d 858, 866 (Fed. Cir.), *cert. denied*, 469 U.S. 856 (1984).

C. Polymerase Chain Reaction Example

The actions of Hoffmann-LaRoche controlling the use of the polymerase chain reaction (PCR) demonstrate how a patented invention can cause uncertainty for non-profit researchers and foreclose scientific progress if other researchers are not allowed to use the invention.

The polymerase chain reaction amplifies minute amounts of genetic material into measurable quantities, enabling scientists to identify and study extremely small traces of genetic material. For instance, if the procedure is repeated twenty times, one million copies of DNA are produced from a single original piece of DNA.⁹⁹ The technique has been analogized to taking a needle in a haystack and using it to generate a whole stack of needles.

PCR is having a revolutionary impact in the biomedical sciences. It is crucial to the human genome project and has great potential for the examination of genetic defects and the identification of infections, viruses and cancers. For instance, PCR plays an increasingly major role in the study and detection of AIDS. PCR can also be used in forensic criminal investigations to obtain DNA fingerprints from blood, semen or hair samples.¹⁰⁰

Dr. Kary Mullis, then at Cetus Corporation, invented PCR in 1983.¹⁰¹ In 1991, Hoffmann-LaRoche paid \$300 million to acquire the patented technology from Cetus.¹⁰² The invention is unusual in that it was made by a biotechnology company but is now in wide use for both academic research and commercial purposes. Roche has not attempted to block any basic research, but the company only recently agreed to relax its hold on PCR rights for certain applications after lobbying by genetic-testing laboratories and researchers, including James Watson. The company dropped the idea of demanding an up-front licensing fee and an annual minimum payment by academic and non-profit institutions.¹⁰³ However, the very fact that Roche considered the plan demonstrates the company's attitude that academic research does constitute infringement. Although

99. The first step in using PCR is to heat the DNA so that the two strands of the double helix come apart. Primers, short pieces of DNA, are attached to each strand to identify the portion of DNA to be copied and to provide chemical instructions for the next step. In the next step, the enzyme DNA polymerase assembles a matching strand of DNA along each original strand, thereby producing two new double helixes of the target DNA. Kary B. Mullis, *The Unusual Origin of the Polymerase Chain Reaction*, SCI. AM., Apr. 1990, at 56; Harold M. Schmeck Jr., *New Test That Finds Hidden AIDS Virus Is a Sleuth With Value in Many Fields*, N.Y. TIMES, June 21, 1988, at C1.

100. Schmeck, *supra* note 99.

101. Mullis, *supra* note 99.

102. Lawrence M. Fisher, *Making a Difference: Chance of a Lifetime*, N.Y. TIMES, Feb. 16, 1992, § 3, at 12.

103. Jeff Johnston & Deborah M. Barnes, *Roche Loosens its Grip on PCR Licensing*, J. NIH RES., Mar. 1992, at 46.

adverse publicity may often force a company to relent in similar circumstances, the uncertain position of academic and non-profit research institutions indicates a potential for future problems.

Roche also considered but abandoned requiring its licensees to report back any new technology based on PCR. Biotechnology companies warned that the requirement would impede improvements by others in PCR techniques, discouraging commercial laboratories from developing new genetic tests since they would be competing directly with Roche's own PCR based products.¹⁰⁴ The proposal shows how a patentee can influence subsequent research.

IV. PAST PROPOSALS AND CRITIQUE

A serious problem results when a patent prevents subsequent research by others. This is especially problematic when basic techniques and information are patented. Although the common law experimental use exception answers this concern in some instances, its scope is limited. Moreover, the full reach of the exception is unclear, especially as it applies to research at universities and federal labs when the ultimate goal is to patent and license research results.

As a solution, commentators and legislators have suggested broadening the experimental use exception to allow any research use short of commercialization. The broad exception would allow subsequent research on patented inventions and would clarify the position of non-profit researchers.

A. Proposals For a Broad Exception

The House Committee on the Judiciary reported favorably on the Patent Competitiveness and Technological Innovation Act of 1990 (H.R. 5598) and recommended that the bill pass in September 1990.¹⁰⁵ Title IV of the bill entitled "The Research, Experimentation and Competitiveness Act of 1990" proposed a broad experimental use exception. The proposed legislation would add section 271(j) to the patent code, which would state:

It shall not be an act of infringement to make or use a patented invention solely for research or experimentation purposes unless the patented invention has a primary purpose of research or experimentation. If the patented invention has a primary purpose of research or experimentation, it shall not be an act of infringement to manufacture or use such invention to study, evaluate or characterize

104. *Id.* at 46-47.

105. H.R. REP. NO. 960, 101st Cong., 2d Sess. 1 (1990).

such invention or to create a product outside the scope of the patented invention to which subsection (e)(1) applies.¹⁰⁶

The Committee Report describes the amendment as a clarification of case law, although this description is inaccurate as explained below. The Report states, "It is a central tenet of American patent law that there is a right to use scientific information to create new and better inventions in competition with the patented invention."¹⁰⁷ The Report delineates several acts which do not constitute infringement under the proposed legislation:

- (1) testing an invention to determine its sufficiency or to compare it to prior art; (2) tests to determine how the patented invention works;
- (3) experimentation on a patented invention for the purpose of improving on it or developing a further patentable invention; (4) experimentation for the purpose of "designing around" a patented invention;
- (5) testing to determine whether the invention meets the tester's purposes in anticipation of requesting a license; and
- (6) academic instructional experimentation with the invention.¹⁰⁸

Business testing is clearly not an experimental use.¹⁰⁹ Examples 1, 2 and 5 fall under Justice Story's second prong of experimental use, that of testing the verity of the specification. Example 6 is arguably a "philosophical experiment." However, examples 3 and 4, when performed with an ultimate commercial motive, are excluded from the experimental use exception as interpreted in *Roche v. Bolar*. Using a patented invention to improve or design around the invention is infringement when the effort is in keeping with the legitimate business of the user.¹¹⁰ Thus, the legislative proposal presents a broadening, rather than a codification of the current experimental use doctrine.

The report justifies the broad exception partly on the grounds that it clarifies that university research use of a patented invention is an experimental use and not infringement.¹¹¹ Government and university scientists should not be confused about the permissible boundaries of their research because the confusion may chill research and

106. *Id.* at 55-56. The research exemption does not apply to research tools (like microscopes and mice) except to allow study of an invention to create a second invention that falls outside the scope of the original patent. In other words, if an invention's primary use is as an aid or tool in research, so that the research use is not directed towards improving the tool, the use should not fall under the exemption. *Id.* at 56. The legislative overruling of *Roche Products, Inc. v. Bolar Pharmaceutical Co.*, 35 U.S.C. § 271(e), is not altered by the proposed amendment. *Id.*

107. *Id.* at 41.

108. *Id.* at 44-45.

109. *Id.* at 45-46.

110. *Roche Prods., Inc. v. Bolar Pharmaceutical Co.*, 733 F.2d 858, 863 (Fed. Cir.), *cert. denied*, 469 U.S. 856 (1984).

111. In addition, supporters of a broad exception worry that legitimate scientific activities are driven outside the U.S. where the work is exempt from U.S. patent laws. H.R. REP. NO. 960, *supra* note 105, at 44.

experimentation.¹¹² If university research were clarified as experimental, then researchers could work with patented materials without paying a licensing fee.¹¹³ Almost all university research could be described under examples 3 and 4 as improving or designing around the invention if it could not be described as academic instructional experimentation according to example 6. As long as the university refrained from "business testing" it would not infringe.

A broad experimental use exception would also lessen the foreclosure of subsequent research when a basic technique or information is patented. A subsequent researcher could improve on or design around the base patent without risking infringement.

Commentators have made recommendations which generally correspond to the proposed legislation broadening the exception by allowing experimentation on a patented invention for the purpose of improving or "designing around" it, even with the long term goal of an eventual business use.¹¹⁴ Given the support for a broad experimental use exception voiced by practitioners,¹¹⁵ scholars¹¹⁶ and Congress,¹¹⁷ it is likely that Congress will eventually enact a bill similar to the one quoted above. The task at hand is to find the wisest limits for the exception while providing a workable solution to the problems of foreclosed research and the uncertain position of non-profit researchers. Any proposal must take into account the economics and incentives of the patent system. Accordingly, the following Part describes those incentives before attempting a critique of the proposals for a broad experimental use exception.

112. *Id.* at 43-44.

113. *Id.* at 43.

114. One commentator argues that the case law supports a position that the experimental use exception applies to testing a patented invention for adaptation to the experimentor's business provided that the experimental use does not result in a "use for profit." He contends that, based on case law and the policies behind the patent system, the exception ought to apply to infringement while developing new uses and improvements for the patented technology if the infringer does not make or attempt to make a monetary profit during the infringement. Hantman, *supra* note 14, at 644. Given the language in *Roche v. Bolar*, however, it is doubtful that the Federal Circuit will follow this interpretation. See *supra* notes 27-35 and accompanying text; see also Irving N. Feit, *Biotechnology Research and the Experimental Use Exception to Patent Infringement*, 71 J. PAT. & TRADEMARK OFF. SOC'Y 819, 835-37 (1989). Accordingly, any proposals on the scope of the experimental use exception are best viewed as suggestions for changes to the patent laws.

115. See generally Feit, *supra* note 114, at 819; Israelsen, *supra* note 6.

116. Eisenberg, *supra* note 6.

117. H.R. REP. NO. 960, *supra* note 105.

B. Incentives of the Patent System

Invention and innovation¹¹⁸ contribute to the public good and should be encouraged. However, a free enterprise economy underinvests in research because it is risky and sometimes unrewarded.¹¹⁹ Moreover, when something of value is discovered in a free enterprise economy, competitors may quickly appropriate the innovation, thus decreasing the inventor's profits by driving down prices. Thus, underinvestment in research and development is expected because the social returns on these activities are greater than the private returns. Therefore, society must promote invention and innovation because profit-maximizing firms would often choose to be imitators rather than innovators.¹²⁰ The goal of the patent system is to make the level of investment in R&D closer to its social value.

The patent system advances this goal in three ways; by promoting invention, by encouraging the development and commercialization of the invention and by encouraging public disclosure of the invention.¹²¹ The patent system promotes invention that might not occur otherwise by granting the inventor an exclusive right in her invention and presumably the ability to exploit monopoly profits.¹²² The lure of monopoly profits stimulates the invention of new products and processes. A patent enables an inventor to capture returns from her investment by preventing others from appropriating the invention and driving down the price of the final product. In this way the patent system curtails free-riding that would otherwise discourage research and development activities.¹²³

The incentive provided by the patent system to develop an invention to the point of commercialization is as important as the

118. As a matter of definition, "invention" will be used to refer to the first implementation of the inventor's idea, but not necessarily a commercial product. An "innovation" is the commercially practical form of the invention. An innovation may significantly differ from an invention due to the changes necessary to turn the invention into a commercial product. Robert P. Merges, *Commercial Success and Patent Standards: Economic Perspectives on Innovation*, 76 CAL. L. REV. 803, 807 (1988).

119. Kenneth J. Arrow, *Economic Welfare and the Allocation of Resources for Invention*, in THE RATE AND DIRECTION OF INVENTIVE ACTIVITY: ECONOMIC AND SOCIAL FACTORS 609, 619-24 (1962).

120. See MORTON I. KAMIEN & NANCY L. SCHWARTZ, MARKET STRUCTURE AND INNOVATION 112-13 (1982).

121. F.M. SCHERER, INDUSTRIAL MARKET STRUCTURE AND ECONOMIC PERFORMANCE 440 (2d ed. 1980).

122. A patentee's actual power over price varies widely from case to case depending on the availability of substitutes and elasticity of demand for the patented product. *Id.* at 449.

123. Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265, 266 (1977).

incentive to invent.¹²⁴ The first patentable invention is frequently discovered years before the first significant commercial product is marketed.¹²⁵ In many cases, an invention requires extensive development before any commercial application is possible. For example, this was the case with the laser, the transistor, nylon and xerography.¹²⁶ The patent system furnishes an incentive to develop early inventions by allowing an innovator to recoup development costs through monopoly profits and by protecting an innovator's central ideas during the development process. In the modern industrial world, development accounts for more than three fourths of all industrial R&D expenditures.¹²⁷ Developing an invention to the point of commercial applicability is both costly and risky. To recoup their development costs plus a premium for risk, developers must be able to sell the product at a price which exceeds production costs. The patent system prohibits imitation of the new product which would otherwise drive down the price in a competitive market, thereby allowing the innovator to recoup development costs.¹²⁸

Because most inventors apply for patents at the beginning of a long process leading from invention to innovation, the incentive to develop can be characterized as an incentive to complete the project. The incentive stems from the knowledge that the exclusive rights granted by the patent protect the inventor from appropriation of the invention's central ideas while the inventor converts the invention to an innovation. Through the patent the inventor carves out an area which she exclusively is allowed to develop.¹²⁹

124. By enacting the 1980 Amendments to the Patent Act, which allow small businesses and universities to own and license patents resulting from federally funded research, Congress recognized the importance of patent protection as an incentive to invest in development. *See supra* notes 45-49 and accompanying text.

125. Inventors file patent applications before an invention is marketable because the laws and rules governing the patent system encourage early filing in three ways. First, the inventor can obtain a patent before the invention is commercially practical because she need not actually construct the invention to file the application, she need only supply instructions that enable one skilled in the art to make it. 35 U.S.C. § 112 (1988). Second, although a patent is awarded to the first to invent rather than the first to file, late filers carry the burden of proving they were the first to invent in an interference proceeding to determine the first to invent between parties having filed applications on the same invention. 35 U.S.C. § 102 (1988). Third, publication or other public disclosure of the invention bars a patent after one year. 35 U.S.C. § 102(b) (1988). If an inventor wishes to make an invention public she has only one year to file the application.

126. JOHN JEWKES ET AL., THE SOURCES OF INVENTION 263-410 (1959); Kitch, *supra* note 123, at 271-72, 276.

127. SCHERER, *supra* note 121, at 440.

128. *Id.* at 444. Of course, the patent owner has an incentive to make the required investment in development and maximize the value of the patent only if the patent will cover the commercial embodiment. Otherwise, the development investment may produce information as to product manufacture and design that would be appropriable by competitors. Kitch, *supra* note 123, at 276-77.

129. Merges, *supra* note 118, at 840.

In exchange for the exclusive grant, society gains the resources that cost-saving innovations release for alternative uses and the introduction of superior products and more efficient processes which would not otherwise have been made or made much later. In addition, society may freely use the invention once the patent expires.¹³⁰

The patent system also offers the important incentive to disclose the invention. By offering protection from appropriation by others, even after the invention becomes public, patent protection leads inventors to make public what they might otherwise keep secret.¹³¹ The inventors' other major method of protecting inventions from appropriation, trade secrets, are risky because once the information is disclosed, even unintentionally or without consent, proprietary protection is no longer possible. Moreover, the inventor has no protection if another party independently creates the invention.¹³²

An issued patent is public so other firms can direct their work so as not to duplicate work already done by the patent holder. When trade secrets protect the research, however, a competing firm striving for the same result will likely duplicate efforts made by others.¹³³ Thus, it is reasonable to assume that research which was kept secret due to a lack of patent protection would also be duplicated by competitors hoping to achieve the same product. Accordingly, trade secret protection arguably leads to more duplication than patent protection.

Another benefit of the patent system is that it generates R&D activity by creating a desire to design around another firm's patent to avoid infringement and royalty payments. Sometimes design around the patent yields a superior product or process.¹³⁴ In addition, companies may conduct R&D to guard against being foreclosed from an area of technology by the patent rights of another. A company's own patents give it a bargaining chip in negotiating with firms holding complementary patents.¹³⁵

130. SCHERER, *supra* note 121, at 442.

131. *Id.* at 441. A successful patent application must disclose a novel and nonobvious invention in a way that enables one skilled in the art to which the invention pertains to practice the invention. 35 U.S.C. §§ 102, 103, 112 (1988).

132. Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 476 (1974).

133. Kitch, *supra* note 123, at 278.

134. SCHERER, *supra* note 121, at 446-47. However, some commentators have criticized the competitive and duplicative research devoted to design around activity as being, on the whole, wasteful. See William S. Comanor, *Research and Competitive Product Differentiation in the Pharmaceutical Industry in the United States*, 31 ECONOMICA 372, 381-84 (1964). Scherer concludes that it is impossible to determine whether the benefits of design around research exceed the cost of the duplication. SCHERER, *supra* note 121, at 446. This criticism overlooks the importance of the incentive to disclose.

135. SCHERER, *supra* note 121, at 446. Cross-licensing is a common way to bargain with patents.

C. A Critique of the Broad Exception

1. THE HARM TO THE PATENT INCENTIVE

The broad experimental use exception, which would allow any person to use a patented invention in research and development up to the point of commercialization, does clarify the position of non-profit researchers and partially alleviate the foreclosure of subsequent inventions. Unfortunately the broad exception will also discourage inventors from using patent protection in the future. If a broad exception discourages use of the patent system, it decreases public disclosure of new inventions and reduces the incentive to invent and develop in industries where patent protection is especially important.¹³⁶ For these reasons the Patent and Trademark Office opposes legislation creating a research exemption.¹³⁷

The broad exception harms the incentive to invent. Allowing an experimental use exception in a commercial setting allows subsequent inventors to free ride on the original inventor's work if the subsequent inventor can use the original invention to improve on and design around the original inventor's crude models. This free riding can make it impossible for the original inventor to appropriate returns on early R&D investment.¹³⁸ A broad experimental use exception could lead to less R&D rather than more if inventors refuse to invest in inventive activity due to free riding.

The broad exception also harms the incentive to develop already patented inventions to the point of commercialization. As discussed above, a patentable invention often requires extensive development before any commercial application is possible. In this situation the patent serves as an incentive to complete the project because it guards against appropriation of the inventor's central idea while she converts the invention to a marketable innovation. A commercial experimental use exception destroys this important control over the patented idea. If others can use the patented invention to design around, improve and ultimately supersede the invention with a superior alternative, the use contributes another degree of uncertainty to the profitability of the final product. Therefore, the patentee has less reason to develop and commercialize the invention. But without commercialization the patentee

136. For an analysis of this question, see Jordan P. Karp, *Experimental Use as Patent Infringement: The Impropriety of a Broad Exception*, 100 YALE L.J. 2169 (1991).

137. H.R. REP. NO. 960, *supra* note 105, at 8 n.25 ("It could diminish the strong incentive provided by the patent system.").

138. See Ben T. Yu, *Potential Competition and Contracting in Innovation*, 24 J.L. & ECON. 215, 237 (1981).

cannot recoup her investment in researching and developing the invention, discouraging both invention and innovation.

The proposal for a broad experimental use exception amounts to an unlicensed appropriation of the patent during the research and development stages. At a minimum the appropriation deprives the patentee of royalties for the research use even though the patentee is clearly contributing something of value to the infringer's work. Even if an attempt to design around or improve a patent is unsuccessful, the infringer would have used the patented invention with an eye toward future profit without compensating the patent holder. The patentee loses some of the reward of the patent monopoly that encourages invention and development. Allowing a competitor to benefit from a patentee's inventive activity without any payment lessens the patentee's economic return on the inventive investment. The result may be to reduce inventive activity in industries that rely heavily on patent protection.¹³⁹

Because a broad experimental use exception encourages design around and improvement activity it also harms the incentive to disclose and encourages secrecy. Inventors may choose secrecy as an alternative to patent protection if their patent disclosure can make the invention and eventual innovation less valuable. The extent to which the patent disclosure facilitates designing around the patent influences the choice between obtaining a patent and maintaining the invention as a trade secret.¹⁴⁰

Disclosure through patenting is preferable to secrecy for at least two reasons, even if the patented invention cannot freely be used in research. First, the patent avoids duplication of the same work by competitors by informing them of the results. Second, the patent informs competitors of what technology is available for licensing and cross-licensing.

Firms have significant incentives to enter licensing and cross-licensing agreements.¹⁴¹ A patentee may license an invention that it

139. The differences among innovation processes in different industries and the differences in the importance of patents to these industries makes the incentives provided by the patent system more relevant in some cases than in others. Merges, *supra* note 118, at 846 n.181 and references therein. For instance, patents constitute a critical dimension of product differentiation in the pharmaceutical industry, but are less important in the mechanical and electronic industries. See Edwin Mansfield, *Patents and Innovation: An Empirical Study*, 32 MGMT. SCI. 173 (1986).

140. Richard C. Levin et al., *Appropriating the Returns from Industrial Research and Development*, 1987 BROOKINGS PAPERS ON ECON. ACTIVITY (Special Issue on Microeconomics) 783, 805.

141. See JULIAN LOWE & NICK CRAWFORD, INNOVATION AND TECHNOLOGY TRANSFER FOR THE GROWING FIRM 1 (1984) (licensing adds another resource available to a growing company); Note, *An Economic Analysis of Royalty Terms in Patent Licenses*, 67 MINN. L. REV. 1198, 1226 n.166 (1983) (small firms need licensing due to their lack of marketing organization necessary for commercialization of an invention); Michael L. Katz & Carl Shapiro, *On Licensing of Innovation*, 16 RAND J. ECON. 504 (1985); Eric von Hippel,

cannot develop and commercialize itself or that the licensee can develop more efficiently. Royalty-free cross-licensing schemes can increase efficiency of industry-wide R&D.¹⁴² Cross-licensing agreements also provide a constructive way to avoid stalemates between complementary patent portfolios.¹⁴³ Many firms may cross-license or license because they are risk-averse and prefer to preserve a delicate balance of relations within an industry or to assure future licensing by rivals.¹⁴⁴ Some firms would rather risk losing monopoly profits from their own invention by cross-licensing than be excluded from the market by a competitor's important innovation.¹⁴⁵

The prevalence of licensing and cross-licensing agreements also indicates that in many cases a patent holder will not in fact foreclose further development and improvements on the patent. If the patent holder licenses the invention, the licensee can proceed. Thus, disclosure through the patent system promotes more innovation than does secrecy.

2. THE HARM TO SMALL FIRMS

In general, the patent system appears to be of more value in stimulating invention and innovation by small rather than large firms. Because the market position of a small firm is more vulnerable to imitation by large firms, patents do more to protect their market position. In addition, small firms will likely be slower at penetrating new markets through innovation, given their lack of distribution channels and market acceptance as compared to large firms.¹⁴⁶ For these reasons, anyone proposing changes to the patent laws should be especially cognizant of their effect on small firms.

The broad exception would be especially harmful to small firms, research centers and universities that invent important advances but do not have the resources to develop the advance into a commercial product.¹⁴⁷ In this case the patent holder may license the invention to an entity that can develop a commercial product.¹⁴⁸ Licensing can be a

Cooperation Between Rivals: Informal Know-How Trading, 16 RES. POL'Y 291 (1987) (licensing allows research and development costs to be split between firms); Merges, *supra* note 118, at 868-69; Yu, *supra* note 138, at 234-36 (licensing can be used to reduce litigation and preserve industry relations).

142. Michael L. Katz, *An Analysis of Cooperative Research and Development*, 17 RAND J. ECON. 527 (1986).

143. SCHERER, *supra* note 121, at 452.

144. Merges, *supra* note 118, at 868.

145. Karp, *supra* note 136, at 2185.

146. SCHERER, *supra* note 121, at 449.

147. For a description of a university's expectations that industrial firms will license a university's patents in order to experiment with and develop those inventions, see *supra* notes 85-86 and accompanying text.

148. Note, *supra* note 141, at 1200, 1226 n.166.

central part of a firm's innovation policy, crucial to the firm's survival and growth.¹⁴⁹ If the developer can claim it is experimenting by improving on the original patent, it will not have to take a license unless the final product infringes the original patent.

The patent portfolio of a start-up company is often its major asset, enabling it to raise venture capital for highly speculative R&D before the company has a product ready for market. If larger, better funded companies were able to use the patented inventions of start-ups to design their own commercial alternatives, the smaller companies' patent portfolio would be less attractive and less able to attract funding. Overall, this scenario would have an important dampening effect on innovation because small firms contribute significantly toward the creation of new products and processes.¹⁵⁰

V. NEW PROPOSALS

A generally applicable broad experimental use exception weakens the incentives to invent, to develop, and to disclose provided by the patent system to too great an extent when applied to patents resulting from private research and development efforts. If the disclosure of a patented invention can be used to make the patent less valuable, inventors will have less incentive to make risky investments in ground breaking research. Inventors may also choose secrecy over patents. The broad exception does, however, provide a solution for the uncertain position of non-profit researchers and the foreclosure of subsequent research based on a patented inventions.

In response to these conflicting values this Comment proposes that the broad experimental use exception (extending up to commercialization) be made applicable in the special circumstances in which the harm to the patent incentive is minimal compared to the resulting benefits. First, university and other non-profit researchers should be allowed the advantage of the broad exception. Second, the broad exception should be applied to all patented, federally funded inventions so that any party can use these inventions in research and development.

A. Non-profit Researchers Allowed Broad Exception

Because one of the problems with the common law experimental use exception is that the status of non-profit researchers such as universities is uncertain due to their patenting and licensing activities, a simple remedy is to grant these researchers the benefit of a clarified

149. LOWE & CRAWFORD, *supra* note 141, at 45.

150. SCHERER, *supra* note 121, at 416-17; Karp, *supra* note 136, at 2183.

experimental use exception, which would exempt them from infringement when studying and improving a patented invention.¹⁵¹

The exemption should extend only to research use and not to commercialization of a product. However, the exemption will provide the certainty that non-profit researchers seek because they do not engage in commercialization anyway. Most importantly, the exempted work cannot harm the patentee because the research institution cannot commercialize a non-infringing design-around.

Although researchers could theoretically license the patents used in their research, in reality licensing patented inventions used in basic research poses special problems. The need to obtain licenses would add significant administrative and financial burdens to researchers in fields where patent protection is widespread.¹⁵² Most research builds on many prior discoveries. If a significant number of these are patented, obtaining licenses on each would generate mounting royalty and transaction costs. Society should question whether paying royalties to further basic research is the best use of scarce resource funds. Licensing would add another expense to conducting research at a time when funding is not keeping pace with needs.¹⁵³

The unpredictable nature of basic research may make it impossible to determine significantly in advance what patents will be needed, leading to delays if the researcher must negotiate a license in the midst of a project. In addition, the contribution of that patent to the research (and, therefore, a fair royalty rate) may also be impossible to determine in advance.¹⁵⁴ More likely, the scientist without astute legal advice will not realize that she is using a patented invention. She risks a law suit when she makes public her methods and results.¹⁵⁵

1. A LIMITATION ON THE EXEMPTION

The close relationships between industry and universities require that the proposed exemption be subject to a key limitation. A university's exempted research does not directly harm the patentee because the university cannot commercialize any results. However, those results could be patented and exclusively licensed to the original patentee's competitor. An industrial firm could simply move any work that it

151. The exemption would not cover routine laboratory use of patented research tools. For instance, patented microscopes are routinely used in certain types of research and their main purpose is to serve as a research tool. To exempt these tools from infringement would make those patents worthless without furthering technical progress.

152. Eisenberg, *supra* note 94, at 743.

153. See *supra* notes 56-57 and accompanying text.

154. Eisenberg, *supra* note 89, at 217.

155. These issues will be of less concern for the industrial scientist with legal advice and working on more predictable development projects.

wished to do on a patented invention to a university setting. The firm would support the university research and if a patentable design-around resulted, the firm would take a license from the university. The end result would be that university experimental use would be generating the same disincentive that corporate experimental use generates. Industries with university ties could easily circumvent the spirit of the special university exemption.

A reasonable limitation can prevent this easy circumvention. When university (or other non-profit) research takes advantage of the exemption by using a patented invention, the results should be used by industry only under specific circumstances. The industry must operate as though it wished to conduct itself the research done by the university. If work done by the university would have infringed if conducted by a private firm, then the firm must negotiate a license with the patent holder before the firm can use the university results. This license should be required whether or not the research results infringe. Of course if the results do infringe, the firm must also request a license for use of the results. Requiring a license even for use of non-infringing results insures that industry will not tempt university laboratories into creating a design-around which the industrial firm then commercializes.

It is possible that the patent holder will refuse to grant a license in this situation. That does not mean, however, that the university research can never be put to practical use. The holder of the patent on which the research was based may use the results. If the university patents its results, the original patent holder may take a license from the university in order to commercialize those results. This proposal is easily implemented given the university-industry relationships described above.¹⁵⁶

The openness required in a university means that any infringement can be detected and monitored by the patentee. The detection prevents the scenario described above in which infringing research is set in a university and later moved to industry for commercialization. Universities encourage the publication and free sharing of research results.¹⁵⁷ Traditionally, the scientific community has operated on a principle of communal ownership of research results rather than private ownership. According to the norms that guide the scientific community, all discoveries build on previous work and contribute to further discoveries. To extend knowledge, scientists must dedicate their scientific findings to the scientific community.¹⁵⁸

156. See *supra* notes 63-78 and accompanying text.

157. DICKSON, *supra* note 49, at 89.

158. Robert K. Merton, *The Normative Structure of Science*, in THE SOCIOLOGY OF SCIENCE 267, 273-75 (Norman W. Storer ed., 1973).

In exchange for dedicating discoveries to the public, the scientific community offers professional recognition and esteem to those who make original contributions to the store of knowledge.¹⁵⁹ Scientists will publish results as quickly as possible to avoid being scooped by researchers doing competing work, making sure the first disclosure will be original.¹⁶⁰ Thus, the traditional reward structure of science encourages original research and prompt disclosure.¹⁶¹ Although university researchers are influenced by the rewards available from privatization of results, the forces supporting openness are also very strong.

To summarize this first proposal, university researchers will be allowed to conduct research using patented inventions without being subject to infringement. However, any company that wishes to use the results of this exempted research must obtain a license on the underlying patent, even if the university results do not infringe.

B. Government-Funded Inventions Subject to a Broad Exception

The first proposal clarifies the position of university and other non-profit researchers by exempting them from the restrictions of the patent system. However, it is inequitable to allow an institution to gain from the existence of the patent system without submitting to the restrictions that are necessary for the system to exist in the first place. This is especially true given that individual university members can personally profit from their university work, either through royalty payment or through equity interest in a company benefiting from university research.

For this reason, the first proposal should be implemented only in tandem with the second. The second proposal applies the broad experimental use exception to patents resulting from federally funded research so that the patent can be used without liability for infringement up to the point of commercialization. The proposal exempts any researcher, whether for profit or not, from infringement when using a federally funded invention.

159. Robert K. Merton, *Priorities in Scientific Discovery*, in THE SOCIOLOGY OF SCIENCE 286 (Norman W. Storer ed., 1973).

160. Robert K. Merton, *Behavior Patterns of Scientists*, in THE SOCIOLOGY OF SCIENCE 327 (Norman W. Storer ed., 1973).

161. Competition among scientists for recognition is intense and can lead to temporary secrecy, but the norms of science place limits on that secrecy. KENNEY, *supra* note 51, at 108-10. For instance, in 1971 James Watson, then director of Cold Spring Harbor Laboratory, pressured an NIH researcher into delivering a viral strain to Cold Spring Harbor researchers which the NIH team was withholding for competitive reasons. Watson threatened to report the withholding to the director of NIH, the journal *Science* and Congress (on the grounds that results from publicly funded research was not being made available). This threat alone was enough. *Id.* at 109.

This proposal provides a number of benefits without the disincentives which result when a broad experimental use exception is applied to every patent. For instance, federally funded inventions will not foreclose subsequent research, but federal grantees will not lose their incentive to invent and disclose.¹⁶²

1. ISSUES IN FEDERALLY FUNDED RESEARCH

In spite of the advantages which supported passage of the 1980 patent law amendments, private ownership of federally funded work is not without critics. Critics point out that the government does not own any proprietary rights resulting from the research it funds, despite the fact that federal funds account for about half of the over \$100 billion dollars spent each year on research and development in the United States.¹⁶³ The federal government provides two-thirds of the funds for basic research in most years¹⁶⁴ and spends between \$12 and \$14 billion a year on research at universities alone.¹⁶⁵

In dissenting from the House report recommending passage of the 1980 amendments, Rep. Jack Brooks articulated the major flaws in the amendments. He criticized the government patent policy as giving away rights that properly belonged to the taxpayers who funded the research. He argued that there was no reason to assume the new policy would spur productivity since the federal government was already funding half the research and development in the United States. Companies and institutions were accepting the money and producing results without the lure of patents and exclusive licensing rights. Making technological advances available to all offers a greater potential for increased productivity than does offering exclusive rights. Granting exclusive rights to the developer of federally funded research only restricts the number of potential producers.¹⁶⁶

Rep. Brooks acknowledged that when a private company takes risks in developing new products it deserves the exclusive rights and profits that may result. However, when the government risks the taxpayers' money, the rewards should go to all people. Granting a monopoly privilege is justified when a private entity risks considerable sums on research and development because it allows the developer to recapture the investment through commercialization and monopoly profits. But

162. See *supra* notes 157-61 and accompanying text.

163. H.R. REP. NO. 960, *supra* note 105, at 42; *Technology for Sale*, *supra* note 81, at 62.

164. Claude E. Barfield, *Forum: The Truth About Research*, N.Y. TIMES, Mar. 4, 1990, § 3, at 25.

165. Shaffer, *supra* note 58.

166. H.R. REP. NO. 1307, *supra* note 44, at 29-30, reprinted in 1980 U.S.C.C.A.N. at 6487-88.

when the taxpayer bears the financial risk, the justification for granting a monopoly does not exist. Moreover, the new policy creates a disincentive for private investment whenever federal money is available, since exclusive rights would still be available, but without the financial risk.¹⁶⁷

Critics outside Congress made similar claims. Consumer activist Ralph Nader wrote that the new patent laws represented a "massive giveaway" of property which was contrary to the public interest.¹⁶⁸ By allowing federally funded research to be patented by the grantee and exclusively licensed to private firms for commercial development, the government forces the public to pay twice for that research, once through the federal funds and once when purchasing the product. However, without an exclusive license, a company will not invest the capital necessary to commercially develop a product and research results are never made useful.¹⁶⁹

While the private sector should certainly be encouraged to develop government-funded basic inventions, it would be incongruous with the purpose behind public funding of research to allow that development to foreclose further basic research. Much of the research funded by the federal government is basic research which lays the foundation for further discovery and developments. Federal funding by the National Institutes of Health (NIH) and the National Science Foundation (NSF) built the basic scientific knowledge from which commercial biotechnology developed. Funding of basic research was meant to create the technical base necessary to understand and cure diseases.¹⁷⁰ For instance, information concerning human DNA sequences, the discovery of which is funded as part of the human genome project, is vital to the future course of basic research in the biomedical sciences.¹⁷¹ Allowing this information to be patented and exclusively licensed can retard further scientific progress by prohibiting its use in subsequent research. The problem is less acute with regard to privately funded patents since these tend to cover applied inventions rather than basic information.¹⁷²

In the future the government may fund more industry research and allow the firms doing the work to keep any resulting patent rights. This allocation of rights raises the same concerns discussed above when universities keep patent rights in work sponsored by the government.

167. *Id.* at 30-31, reprinted in U.S.C.C.A.N. at 6489.

168. DICKSON, *supra* note 49, at 92.

169. See *supra* notes 47-49 and accompanying text.

170. KENNEY, *supra* note 51, at 241.

171. Eisenberg, *supra* note 94, at 780. Most biotechnology research makes use of patented inventions. See Fox, *supra* note 93.

172. Although a few industries devote 10% of their research and development budgets to basic research, most corporations routinely spend about 3.8%. Barfield, *supra* note 164.

Concern over foreign competition¹⁷³ and the United States' leadership role in technology¹⁷⁴ have prompted calls for the federal government to increase participation with the private sector in research and development activities.¹⁷⁵ For instance, Erich Bloch, former director of the National Science Foundation, encouraged Congress to provide partial support for research on generic technologies at the pre-competitive stage. Generic technologies are those that promise to benefit a wide range of industries. Precompetitive R&D lies between laboratory discoveries, on the one hand, and proprietary product development, on the other. Technical work at this stage focuses on overcoming basic engineering obstacles and barriers which threaten to slow the commercialization and production of new technologies.¹⁷⁶ In a hearing before the Senate Committee on Commerce, Science and Transportation, Mr. Bloch stated:

Investments [in generic technologies] are often too high and too risky for private companies because the technologies are evolving too quickly, or as in some important industries, U.S. companies have lost the necessary technology base or cannot afford to compete with foreign competitors that are funded by their Governments in these endeavors. The federal government must provide the kind of support often available to our foreign competitors in these critical technologies essential to entire industries and the industrial sectors.¹⁷⁷

In response to this type of sentiment, Congress passed the Technology Competitiveness Act of 1988¹⁷⁸ which created the Advanced

173. In a report entitled "Gaining New Ground: Technology Priorities for America's Future," the private Council on Competitiveness partly blamed U.S. industry's loss of market share in technology-intensive products on programs sponsored by the governments of other major industrialized countries. Those programs have used R&D funding, public-private technology consortia, infrastructure programs and tax policy to improve the technological competitiveness of their industries. S. REP. NO. 157, 102d Cong., 1st Sess. 4 (1991).

174. Critics of U.S. policy point to a 1990 Department of Commerce study identifying 12 key emerging technologies: advanced materials, biotechnology, digital imaging technology, superconductors, advanced semiconductor devices, high-density data storage, high-performance computing, medical devices and diagnostics, optoelectronics, sensor technology, artificial intelligence and flexible computer-integrated manufacturing. The study concluded that the United States is losing relative to Europe in 3 of the 12 technologies, remaining steady in 6 and gaining in 3. The United States is also losing relative to Japan in all but two technologies—flexible computer-integrated manufacturing and artificial intelligence. *Id.* at 4.

175. The federal government spends relatively little to support industry-led technology projects. In 1988 only 0.2% of the federal R&D budget promoted industrial development. *Id.* at 5.

176. *Id.* at 3.

177. *Technology Issues Hearing*, *supra* note 57, at 6 (statement of Erich Bloch, Director, National Science Foundation).

178. Pub. L. No. 100-418, §§ 5101-5164, 102 Stat. 1107, 1426-51 (codified at scattered sections of 15 U.S.C.).

Technology Program (ATP) to assist industry-led precompetitive R&D projects in developing new generic technologies.¹⁷⁹ Proposed amendments to the Technology Competitiveness Act, the American Technology Preeminence Act, seek to strengthen the ATP. The amendments state that any intellectual property rights arising from work done under an ATP grant will belong to the company doing the work to avoid deterring participation by private companies.¹⁸⁰

It is possible that in the future, the federal government will be funding more industrial research and allowing the firms to own patent rights in the results of that research. However, exclusive ownership defeats the purpose of government funding research for generic technologies which are meant to benefit a range of industries. Others must be allowed to build on that work.

2. BENEFITS OF THE PROPOSAL

The exception proposed here softens the blow of having the public pay twice for research by ensuring that the research will stimulate rather than foreclose further inventions. To the extent that prior researchers might otherwise charge royalties, free access provides a subsidy for subsequent research.¹⁸¹ It is fitting that publicly funded research should continue providing a subsidy for subsequent research since the point of government-funded research is to stimulate further research.

Society gains no benefit in exchange for granting a patent when an invention would be conceived and developed without patent protection.¹⁸² As discussed below, federal grantees have incentives to invent and disclose outside the patent system.¹⁸³ If an inventor secures patent protection on the invention anyway, society pays the monopoly costs without the corresponding trade-off of an invention that would not have been available otherwise.¹⁸⁴ The suggested changes to the patent laws decrease the social costs of federally funded inventions that would have been created without the patent incentive.

Since federally funded inventions would be created without patent protection, one could argue that they should not be patented. However, patents are also important for providing an incentive to develop an invention, and the ability to grant exclusive licenses on government-funded, patented inventions is necessary to encourage licensees to further

179. In March 1991, the Department of Commerce announced 11 ATP awards. S. REP. NO. 157, *supra* note 173, at 3.

180. *Id.* at 17.

181. Eisenberg, *supra* note 6, at 1057.

182. SCHERER, *supra* note 121, at 443-47.

183. See *supra* notes 157-61 and accompanying text.

184. SCHERER, *supra* note 121, at 443.

develop and commercialized those inventions.¹⁸⁵ The patent laws need only protect the incentive to develop, and not the incentive to invent federally funded inventions. Thus, the weaker patent incentive suggested here should suffice.

The proposed exception also helps resolve a conflict in the scientific community on how best to promote scientific progress. The university scientist's interest in both traditional scientific rewards and patent rewards presents a conflict between the free contribution of knowledge to the community and the private ownership of that knowledge. The fundamental conflict occurs because disclosure through traditional publication marks the end of exclusivity whereas disclosure through patents marks the beginning of exclusivity. Through publication, a researcher contributes the knowledge to the whole community whereas through a patent a researcher stakes a claim in the right to exclude others from using the invention.¹⁸⁶ The conflict arises from the divergent views of how to best promote scientific progress.¹⁸⁷ The patent laws are based on the theory that exclusive rights will provide an incentive to create. The traditions of the scientific community derive from the assumption that free access to discoveries best promotes further progress.¹⁸⁸ The proposal allows free access for federally funded inventions for research purposes while still allowing exclusive ownership for commercial development purposes. Since most academic, basic research is federally funded, the proposal here solves the most perplexing branch of the conflict.

Finally, allowing private industry to freely research with results that will mostly come from universities and other non-profit researchers equitably balances the proposal above which exempts non-profit researchers from infringement in the usual course of their work.

3. EFFECT ON FEDERAL GRANTEES

The broad exception proposed here would not affect the incentive to invent and to disclose provided by the patent system when federal funds are involved. Because university and other non-profit researchers have significant incentive to invent and publish outside the patent system, weakening their patent rights will not effect their productivity.

185. See Kitch, *supra* note 123, at 287.

186. Eisenberg, *supra* note 89, at 217.

187. Scientists seeking patent protection may delay publication of research discoveries that are ripe for reporting to the scientific community but are not ripe for patent protection. *Id.* at 216.

188. Merton, *supra* note 158, at 273; Merton, *supra* note 160, at 346-52; Barber, *supra* note 50, at 197, 198-206; Eisenberg, *supra* note 6, at 1048-49.

The proposed exception will probably not affect federal grantees' incentive to invent. University researchers' major incentive to invent comes from their professional reward structure. Tenure, awards and other forms of professional recognition are based on the researcher's inventiveness. For example, Dr. Jonathan King, a biology professor at the Massachusetts Institute of Technology, has stated, "The extraordinary development of genetic engineering was the fruit of 40 years of public investment in university research by the federal government. The whole notion that you need the profit motive for scientific innovation is spurious. Biotechnology grew up without patenting or proprietary knowledge."¹⁸⁹

Small businesses which receive federal funds but keep any resulting patent rights incur less risk than firms which provide sole support for risky research. It is unlikely that a company would turn down federal grants due to this exception. Thus, weakening the patent rights of these small businesses should not affect their incentive to invent.

Weakening the patent will not discourage university scientists from disclosing their research results because universities can use secrecy to only a limited extent. As discussed above, the reward structure of science places a high value on openness and the disclosure of research results.¹⁹⁰ Moreover, universities will not be forced to choose between secrecy and inadequate patent protection because universities worry much less than private industry about whether a subsequent researcher designs around or improves on its patent. This is true because it cannot commercialize the patent in any event.

This proposal to apply the broad exception to federally funded patents is consistent with the policies of the major scientific funding organizations. NSF guidelines state that the policy allowing awardees to retain intellectual property rights does not reduce the responsibility of researchers and institutions to make results, data, and collections available to the research community.¹⁹¹ The NSF expects significant findings from research it supports to be promptly submitted for publication. It expects investigators to share with other researchers within a reasonable time, the data, samples, physical collections, and other supporting materials created or gathered in the course of the work. It also encourages awardees to share software and inventions, once appropriate protection for them has been secured, or otherwise act to make the innovations they embody widely useful and usable.¹⁹²

189. Shaffer, *supra* note 58.

190. See *supra* notes 157-61 and accompanying text.

191. NATIONAL SCIENCE FOUND., NO. NSF 90-77, GRANTS FOR RESEARCH AND EDUCATION IN SCIENCE AND ENGINEERING: AN APPLICATION GUIDE 15 (1990).

192. *Id.* at 14.

Usually, the financial impact of this proposal on the non-profit research institution will be insignificant. Some universities may be currently licensing their patented technologies to industries conducting research and development with the invention. This proposal will cause the universities to lose that source of funding. However, the loss of funds will rarely be significant. Licensing fees for this type of use are generally lower than fees for a commercial use.¹⁹³

Moreover, most patent programs are actually not financially successful. John Preston, head of M.I.T.'s technology licensing office, warns that universities should not count on a large amount of funds from patent licenses. M.I.T. netted only \$500,000 from the \$5.5 million it grossed from royalty licenses in 1991 due to the costs associated with filing and licensing patents and the \$1 million it distributed to hundreds of individual scientists. Most universities and government laboratories license as few as one percent of their issued patents.¹⁹⁴ In fact, few universities have made or expect to make any significant money from their patents, and some are losing money from their programs.¹⁹⁵

4. EFFECT ON EXCLUSIVE LICENSEES

Companies which exclusively license patents resulting from federally funded research present a less clear picture because the broad exception may affect the incentive to develop. The licensee hopes to have an exclusive market position after developing the patented invention to the point of commercialization. The broad exception supports this goal by not allowing any other firm to commercialize the patent. However, the exception may allow other researchers or firms to design around the patent and create a noninfringing product which replaces the licensee's product in the marketplace.

If the experimental work by a non-licensee results in a product which designs around the basic research and does not infringe the original patent, then federal funds spurred more development at the private level, which is the point of government supported research. A licensee will be forced to accept the possibility of increased design around activity.

In response, the licensee and licensor can account for the added risk through decreased royalty rates. By licensing technology for which it did not have to risk capital to invent, at a rate which accounts for market risks, the licensee is in a better position than if it had been forced to make the original invention itself. Presumably, once the invention is made, development, though expensive, will be less risky than invention. The

193. Brinton, *supra* note 49, at 487.

194. Shaffer, *supra* note 58.

195. DICKSON, *supra* note 49, at 91.

incentive to develop should still be strong enough to motivate the technology transfer from the government contract researcher to the marketplace.

Alternatively, the experimental research by a non-licensee may lead to an improvement that still infringes the underlying patent. Because research in universities is heavily supported by public funds, a university must always consider the public interest in its licensing activities.¹⁹⁶ Thus, the university should certainly license the patent to the developer of the improvement so that the improvement can be marketed. If the university refuses the government can exercise its march-in rights.¹⁹⁷ A problem arises when the base patent has already been exclusively licensed to another firm. The first licensee can block the improvement, but the problem can be resolved through purchase of the improvement patent or through a cross-licensing agreement under which the exclusive licensee and the holder of the improvement patent are both allowed to use the improved technology.¹⁹⁸

Alternatively, the original patent holder, the university, might plan ahead for such an occurrence and incorporate a provision into the exclusive licensing agreement mandating cross-licensing in the event of a significant improvement. A university may view this as part of its responsibility for acting in the public interest.

5. PRACTICAL POINTS

A research project may be funded by both federal and private sources. Therefore it is necessary to set a minimum amount of government funding before the exception becomes applicable, perhaps 50%. Implementing this proposal will not be difficult. Government grant applications require that the applicant report all other sources of funding for a project, and thus the total funding for any project is a matter of public record. For instance, whenever Department of Energy grantees issue any document describing projects funded in whole or in part with DOE money, the document must state the percentage of the total cost financed with DOE money.¹⁹⁹ Thus, mechanisms for keeping track of percent funding have been designed and must be in place for some non-profit research.

A grant application must include a detailed description of the proposed project, including the objectives of the project and the applicant's plan for carrying it out. Applications must also include a

196. Brinton, *supra* note 49, at 483.

197. 35 U.S.C. § 203 (1988).

198. SCHERER, *supra* note 121, at 444.

199. OFFICE OF ENERGY RESEARCH, DEPARTMENT OF ENERGY, NO. DOE/ER-0249, APPLICATION AND GUIDE FOR THE SPECIAL RESEARCH GRANT PROGRAM 32.

detailed budget, with supporting written justification sufficient to evaluate the costs of the proposed project.²⁰⁰ Awardees must report on their progress to the funding agency²⁰¹ and clear any major changes in the research plan with the government funding agency.²⁰² Consequently, even if the research does not go as planned it is possible to track which results were funded by which grant.

To facilitate full use of the exception, the pertinent regulations should require that patent applicants disclose the percent of government support for the invention when the application is filed. The disclosure of government funding should be subject to rules of candor before the Patent Office to insure compliance by patentees. If the application issues, a statement on the patent cover will notify interested parties that this patent is available for experimental use.

VI. CONCLUSION

The proposals laid out here attempt to strike a balance between maintaining a strong patent incentive, answering the concerns of non-profit researchers, and ensuring that the patent system forecloses a minimum of new inventions.

Arguably, society might gain new products through a generally applicable broad experimental use exception by allowing infringing design around activity that would not occur otherwise. But society would also sacrifice some level of inventive activity by allowing the broad exception. No patent system can exist without foreclosing some subsequent work; short-term inefficiencies must be traded for long-term progress. Economist Joan Robinson described the paradox inherent in the patent system:

A patent is a device to prevent the diffusion of new methods before the original investor has recovered profit adequate to induce the requisite investment. The justification of the patent system is that by slowing down the diffusion of technical progress it ensures that there will be more progress to diffuse. . . . Since it is rooted in contradiction, there can be no such thing as an ideally beneficial patent system, and it is bound to produce negative results in

200. *Id.* at 35; NATIONAL SCIENCE FOUND., *supra* note 190, at 3.

201. NSF grantees must submit an annual technical progress report, which summarizes activity during the past year, identifies any scientific developments and describes any problems encountered. After the expiration of the grant, the investigator submits a Final Project Report which contains the results of the supported activity. NATIONAL SCIENCE FOUND., *supra* note 190, at 14. The Department of Energy has identical requirements. OFFICE OF ENERGY RESEARCH, *supra* note 198, at 31.

202. OFFICE OF ENERGY RESEARCH, *supra* note 198, at 35; NATIONAL SCIENCE FOUND., *supra* note 190, at 13.

particular instances, impeding progress unnecessarily even if its general effect is favorable on balance.²⁰³

Proponents of the broad experimental use exception can point to instances in which negative results produced by the patent system would have been alleviated by a broad exception. It is important to keep the "contradiction" of the patent system in mind, however, when judging any proposed weakening of the patent system. For instance, Scherer concludes that "society gains unambiguously from inventions and innovations induced or hastened by the grant of patent rights."²⁰⁴ We must look beyond the particular negative instances to the generally favorable effect of the entire system.

203. JOAN ROBINSON, THE ACCUMULATION OF CAPITAL 87 (1956).

204. SCHERER, *supra* note 121, at 443.