Berkeley

[technology law Journal]

VOLUME 26 NUMBER 2 2011

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF LAW
BOALT HALL

Production: Produced by members of the *Berkeley Technology Law Journal*. All editing and layout done using Microsoft Word.

Printer: Joe Christensen, Inc., Lincoln, Nebraska.
Printed in the U.S.A.
The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Library Materials, ANSI Z39.48—1984.

Berkeley Technology Law Journal
U.C. Berkeley School of Law
Student Center, Ste. 3
Berkeley, California 94720-7200
btlj@law.berkeley.edu
www.btlj.org

BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 26 NUMBER 2 SPRING 2011

TABLE OF CONTENTS

ARTICLES

SUBSCRIBER INFORMATION

The *Berkeley Technology Law Journal* (ISSN1086-3818), a continuation of the *High Technology Law Journal* effective Volume 11, is edited by the students of the University of California, Berkeley School of Law (Boalt Hall) and is published four times each year (May, August, November, February) by the Regents of the University of California, Berkeley. Periodicals Postage Rate Paid at Berkeley, CA 94704-9998, and at additional mailing offices. POSTMASTER: Send address changes to Journal Publications, 311 U.C. Berkeley School of Law, University of California, Berkeley, CA 94720-7200.

Correspondence. Address all correspondence regarding subscriptions, address changes, claims for non-receipt, single copies, advertising, and permission to reprint to Journal Publications, 2850 Telegraph Avenue, Suite 561 #7220 Berkeley, CA 94705-7220; (510) 643-6600; JournalPublications@law.berkeley.edu. Authors: see section entitled Information for Authors.

Subscriptions. Annual subscriptions are $65.00 for individuals and $85.00 for organizations. Single issues are $27.00. Please allow two months for receipt of the first issue. Payment may be made by check, international money order, or credit card (MasterCard/Visa). Domestic claims for non-receipt of issues should be made within 90 days of the month of publication; overseas claims should be made within 180 days. Thereafter, the regular back issue rate ($27.00) will be charged for replacement. Overseas delivery is not guaranteed.

Form. The text and citations in the *Journal* conform generally to the THE CHICAGO MANUAL OF STYLE (16th ed. 2010) and to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 19th ed. 2010). Please cite this issue of the *Berkeley Technology Law Journal* as 26 BERKELEY TECH. L.J. ___ (2011).

BTLJ ONLINE

The full text and abstracts of many previously published *Berkeley Technology Law Journal* articles can be found at http://www.btlj.org. Our site also contains a cumulative index, general information about the *Journal*, and the Bolt, a collection of short comments and updates about new developments in law and technology written by members of BTLJ.

INFORMATION FOR AUTHORS

The Editorial Board of the *Berkeley Technology Law Journal* invites the submission of unsolicited manuscripts. Submissions may include previously unpublished articles, essays, book reviews, case notes, or comments concerning any aspect of the relationship between technology and the law. If any portion of a manuscript has been previously published, the author should so indicate.

Format. Submissions are accepted in electronic format through the Express*O* online submission system. Authors should include a curriculum vitae and resume when submitting articles. The Express*O* submission website can be found at http://law.bepress.com/expresso.

Citations. All citations should conform to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 19th ed. 2010). In addition, the author should include his or her credentials, including full name, degrees earned, academic or professional affiliations, and citations to all previously published legal articles.

Copyrighted Material. If a manuscript contains any copyrighted table, chart, graph, illustration, photograph, or more than eight lines of text, the author must obtain written permission from the copyright holder for use of the material.

DONORS

The *Berkeley Technology Law Journal* and the Berkeley Center for Law & Technology acknowledge the following generous donors to Berkeley Law's Law and Technology Program:

Benefactors

CHADBOURNE & PARKE LLP

COOLEY LLP

COVINGTON & BURLING LLP

FENWICK & WEST LLP

FISH & RICHARDSON P.C.

KIRKLAND & ELLIS LLP

LATHAM & WATKINS LLP

MCDERMOTT WILL & EMERY

MORRISON & FOERSTER LLP

ORRICK, HERRINGTON & SUTCLIFFE LLP

SKADDEN, ARPS, SLATE, MEAGHER & FLOM LLP & AFFILIATES

WEIL, GOTSHAL & MANGES LLP

WHITE & CASE LLP

WILMER HALE

WILSON SONSINI GOODRICH & ROSATI

WINSTON & STRAWN LLP

Members

ALSTON + BIRD LLP

BAKER BOTTS LLP

BINGHAM MCCUTCHEN LLP

DLA PIPER

FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP

GOODWIN PROCTER LLP

GUNDERSON DETTMER STOUGH VILLENEUVE FRANKLIN & HACHIGIAN, LLP

HAYNES AND BOONE, LLP

HICKMAN PALERMO TRUONG BECKER, LLP

KEKER & VAN NEST LLP

KNOBBE MARTENS OLSON & BEAR LLP

MORGAN, LEWIS & BOCKIUS LLP

MUNGER, TOLLES & OLSON LLP

ROPES & GRAY LLP

SCHWEGMAN LUNDBERG WOESSNER

SIDLEY AUSTIN LLP

KILPATRICK TOWNSEND & STOCKTON LLP

WEAVER AUSTIN VILLENEUVE & SAMPSON, LLP

VAN PELT, YI & JAMES LLP

Patrons

BAKER & MCKENZIE

DURIE TANGRI

ADVISORY BOARD

BOARD OF EDITORS 2010–2011

Executive Committee

Editor-in-Chief
ELIZABETH OFFEN-BROWN

Managing Editor
APRIL ELLIOTT

Senior Articles Editors
ALEX BAXTER
JONAS HERRELL
KRISTIN KEMNITZER

Senior Executive Editor
MORGAN HAGUE

Senior Annual Review Editors
ELIZABETH ERAKER
DAVID STARK

Editorial Board

Submissions Editors
PARKER KUHL
REBECCA NEIPRIS
ALEXANDER REICHER

Production Editors
JOSEPH ROSE
LAUREN SIMS

Bluebook Editors
TAYLOR BURRAS
JILLIAN FEINBERG
ADAM MCNEILE

External Relations Editors
HEATHER HANEY

Notes & Comments Editors
MICHELLE MA
KELLY YANG

Symposium Editors
WYATT GLYNN
JANA MOSER

Web Content Editor
WILL MOSELEY

Publishing Editor
JAMES PERRY

Web Editor
ANDREW FONG

Assistant Managing Editor
JESSICA LYON

Annual Review Editors
REZA DOKHANCHY
SARA GIARDINA

Member Relations Editor
TINA SALADINO

Articles Editors

EBBY ABRAHAM
AMIT AGARWAL
CHARLES CIACCIO
TARAS CZEBINIAK
AMY HAYDEN
RYAN IWAHASHI
RUBINA KWON
BRIAN LAHTI
BRITTANY LOVEJOY
NIKHIL MATANI
AARON MACKEY
AYLIN ONCEL
MILES PALLEY
MICHAEL SHEEN
ARIELLE SINGH
MICHAEL SOBOLEV

BERKELEY CENTER FOR LAW & TECHNOLOGY 2010–2011

Executive Director
ROBERT BARR

Faculty Directors

AMY KAPCZYNSKI
DEIRDRE MULLIGAN
SUZANNE SCOTCHMER
PETER MENELL
PAMELA SAMUELSON
ROBERT MERGES
PAUL SCHWARTZ
MOLLY VAN HOUWELING

Assistant Director
LOUISE LEE

Assistant Director
DAVID GRADY

Affiliated Faculty and Scholars

AARON EDLIN
JOSEPH FARRELL
RICHARD GILBERT
BRONWYN HALL
THOMAS JORDE
MICHAEL KATZ
DAVID MOWERY
DAVID NIMMER
DANIEL RUBINFELD
ANNALEE SAXENIAN
JASON SCHULTZ
HOWARD SHELANSKI
CARL SHAPIRO
MARJORIE SHULTZ
LON SOBEL
TALHA SYED
DAVID TEECE
JENNIFER M. URBAN
HAL R. VARIAN
DAVID WINICKOFF

MEMBERSHIP SPRING 2011

Associate Editors

KEVIN BENDIX
COURTNEY BOWMAN
JARAD BROWN
BENEDIKT BURGER
CHRIS CIVIL
LAUREN ESCHER
SAMANTAK GHOSH
DAVID GOETZ
JORDAN GONZALES
MARGARET GRAZZINI
ARIANA GREEN
AMIR HASSANABADI
HOLLY HOCH
WINNIE HUNG
NATHANIEL JOHNSON
GARY JUSKOWIAK
RYAN KLIMCZAK
JULIA KOLIBACHUK
NICHOLAS LEEFER
JANE LEVICH
JESSICA MARTINEZ
NATALIE MARTIROSSIAN
SONYA PASSI
WILL PIEROG
DAVID ROSEN
JOE SEXTON
NIRAJAN SINGH
JENNIFER SPENCER
KRISTAL SWOPE
PRISCILLA TAYLOR
CAROLINE THUFASON
HARUKO UNO
WEI WANG
JOSE DE WIT
STEVEN WONG
JACKIE WOO
ANDREA YANKOVSKY
ALBERT YEH
ROBERT YEH
LUIS ZAMBRANO

Members

LILY ACKERMAN
CHRISTINE BAE
MEAGHAN BANKS
ROSS BARBASH
ZACH BARON
ERIK BAUMAN
ACHIM BRINKER
JOSEPH BUI
BOBBY CARROLL
THOMAS CHIA
NOAM COHEN
KRISTEN CORPION
KRISTA CORREA
NATHAN DAMWEBER
KARAN DHADIALLA
SPIRO DHAPI
LEAH DURANTI
MUSETTA DURKEE
RACHEL FISCHETTI
LALITHA GADEPALLY
INDRANEEL GHOSH
DANIELLE GILLILAND
CONRAD GOSEN
MICAH GRUBER
JURABEK HOMIDOV
JAMES HUANG
GWYNNE HUNTER
VIDWATH KASHYAP
DANIEL KAZHDAN
PUNEET KOHLI
VERONIKA KRIZOVA
CLAUDIA LANGER
TIFFANY LEE
YVONNE LEE
HSIWEN LO
ANGELA MAKABALI
TAYLOR MARGOT
ZACHARY MARKARIAN
NEHA MATANI
HANNAH MINKEVITCH
CIARA MITTAN
ERIN MORGAN
JOHN OWEN
MINU RAMANI
HILARY RICHARDSON
JUSTIN TEIXEIRA
EDWARD TOROUS
JULIA VAN DE WALLE
DAVID VERNON
NIKOLAUS WOLOSZCZUK
KILEY WONG
KUO-LIANG YEH

SECRET INVENTIONS

J. Jonas Anderson[†]

TABLE OF CONTENTS

---

† Assistant Professor, American University, Washington College of Law. The author would like to thank Robert Barr, Robert Bartlett, Tim Holbrook, Mark Lemley, Peter Menell, Rob Merges, Jason Rantanen, Amelia Rinehart, Pam Samuelson, Sharon Sandeen, Jason Schultz, Sean Seymore, Molly Van Houweling, and Allen Yu for their comments on earlier versions of this Article. The author is also indebted to the Berkeley Center for Law & Technology for its generous support of this Article.

I. INTRODUCTION

In 1953, Norm Larsen, a self-taught chemist and founder of the Rocket Chemical Company, was attempting to develop a chemical that would fortify metal against rust. On his fortieth attempt, he created a chemical that displaced the standing water that slowly corrodes metal. He named his invention "Water Displacement, Fortieth Attempt." Larsen immediately commercialized his invention, selling it to the U.S. government to protect the outer skin of the Atlas missile from rust and corrosion. Five years after Larsen developed his chemical product he began offering it to the American consumer.[1] Water Displacement, Fortieth Attempt was a resounding commercial success due to the product's affordability and wide range of common household uses. Today, over eighty percent of American households own Larsen's product, now known as WD-40.

Larsen never patented WD-40. Instead, Larsen's company, the Rocket Chemical Company (later renamed the WD-40 Company) relied on trade secrecy to protect its intellectual property.[2] Although other companies have since created water-displacement chemicals that are similar, if not identical,

1. U.S. Patent No. 6,315,152 (filed Nov. 13, 2001) (retelling the Norm Larsen story as background for a patent application for a tube storage device).

2. Douglas Martin, *John S. Barry, Main Force Behind WD-40, Dies at 84*, N.Y. TIMES, July 22, 2009, at B12.

to the chemical Larsen created, WD-40 continues to enjoy commercial success across the world.[3]

The success story of the unpatented WD-40 formula runs counter to traditional conceptualizations of patent law's role in promoting innovation. Patents are often conceptualized as a means of luring secret inventions out of the dark, shadowy cave of trade secrecy, and into the bright, public sunlight of the patent system.[4] Courts tend to characterize the preference for patents over trade secrets as a matter of sound public policy, but this understanding is both incomplete and under-theorized.[5]

Reliance upon trade secrecy, it is thought, leaves the know-how surrounding valuable inventions in the hands of a select few. The traditional quid pro quo view of the patent system imagines the patent grant as the carrot used to entice inventors to reveal their valuable secrets to the public. Secrecy, as conceptualized by the traditional patent quid pro quo viewpoint, is antithetical to the purposes animating the patent system.[6]

The rhetoric used by courts to describe the patent system as discouraging secrecy pervades certain patent doctrines as well. Various patent doctrines attempt to persuade inventors to forego secrecy by favoring patentees over trade secret holders in various potential disputes. For example, a first inventor who relies on trade secrecy risks losing the right to practice her own invention if a subsequent inventor chooses to patent the invention.[7] Other legal doctrines close the proverbial doors of the patent office to inventors

3. Gwendolyn Bounds, *Boss Talk: No More Squeaking By—WD-40 CEO Garry Ridge Repackages a Core Product*, WALL ST. J., May 23, 2006, at B1 (claiming that the WD-40 company sells over one million cans per week).

4. *See generally* Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470 (1974) (describing the patent "quid pro quo" which seeks to encourage inventors to reveal their discoveries via the patent system).

5. *See, e.g.*, Carl Shapiro, *Prior User Rights*, 96 AM. ECON. REV., no. 2, 2006, at 92, 95 ("The effects of encouraging inventors to adopt trade secret versus patent protection are not well understood. Further work is needed to compare the . . . costs that result from inducing some inventors to seek trade secret rather than patent protection.").

6. Gordon L. Doerfer, *The Limits of Trade Secret Law Imposed by Federal Patent and Antitrust Supremacy*, 80 HARV. L. REV. 1432, 1441 (1967); Mark A. Lemley, *The Surprising Virtues of Treating Trade Secrets as IP Rights*, 61 STAN. L. REV. 311, 314 (2008) ("[T]he law operates in various ways to encourage inventors to choose patent over trade secret protection where both are possible."); Jason Mazzone & Matthew Moore, *The Secret Life of Patents*, 48 WASHBURN L.J. 33, 35 (2008) ("Federal law . . . expresses a clear preference for the inventor who discloses an invention to the public and obtains a patent over the inventor who keeps the invention a secret."); Shapiro, *supra* note 5, at 95 ("[T]he current patent system rewards applicants who are most aggressive in seeking patents over those who simply use their own inventions internally as trade secrets.").

7. *See, e.g.*, Gillman v. Stern, 114 F.2d 28, 30 (2d Cir. 1940).

who file for a patent more than one year after commercialization.[8] These doctrines are designed to convince inventors to seek patent protection for their inventions at an early stage in the inventive process, rather than to continue working in secret.

Despite the traditional distaste for secrecy displayed by patent law, secrecy offers several underappreciated benefits. First, secret inventions reduce the administrative and judicial burdens associated with patenting. The U.S. Patent & Trademark Office (PTO) cannot keep up with the over 500,000 patent applications filed each year.[9] Similarly, patent litigation has become an enormously expensive and time-consuming affair, resulting in the creation of an entirely new circuit court of appeals to handle patent appeals.[10] Reliance upon trade secrecy does not involve the expensive administrative and judicial procedures that patent protection entails.

Second, the use of trade secrets does not reduce competition for innovation, as the use of patents does. Unlike a patented invention, a secret invention does not limit competitors from independently discovering or reverse engineering the invention. The unfettered competition that trade secrecy permits attracts competitors to the most successful and profitable inventive spaces. Patents, on the other hand, can discourage inventors from entering into well-researched areas. While the increased competition for innovation may result in duplicative research, the social benefits that come from innovative competition may outweigh the costs of duplication, particularly when research costs are small.[11]

Trade secrecy's competitive benefits extend to the realm of commercialization and development as well. Trade secret exclusivity has an uncertain duration. The potential loss of exclusivity can motivate inventors to rapidly commercialize, develop, and improve their invention. Patentees, on the other hand, may not be as diligent in commercializing due to the patent's relative security from competitors.[12]

Lastly, the availability of secrecy can increase the ex ante incentive to invent in certain cases. Some inventions (such as inventive manufacturing

8. 35 U.S.C. § 102(b), (g) (2007).

9. U.S. Patent & Trademark Office, U.S. Dep't of Commerce, *U.S. Patent Statistics, Calendar Years 1963–2010* (Mar. 2011), http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.pdf [hereinafter PTO, *U.S. Patent Statistics*] (recording 520,277 U.S. patent applications filed in 2010).

10. *See* Rochelle Cooper Dreyfuss, *The Federal Circuit: A Case Study in Specialized Courts*, 64 N.Y.U. L. REV. 1, 6 (1989) (stating that the Federal Circuit was created partially as a result of the caseload crisis at the federal courts).

11. *See infra* Part III.

12. Ted Sichelman, *Commercializing Patents*, 62 STAN. L. REV. 341, 358 (2010).

and chemical methods) are more valuable to their inventors as secrets than those inventions would be as patents. This increased value is due to the avoided costs of patenting (the cost of obtaining and enforcing patent rights) as well as the potential private benefits of secrecy (primarily the potential duration of exclusivity and the ability to conceal their invention from competitors). Secrecy can thus increase an invention's private value as compared to the same invention if patented; increased private value increases the ex ante incentives to create.[13]

In light of the social benefits of secret inventions, this Article argues that trade secrets and patents should be viewed not as opposing systems of invention protection, but rather as complementary tools for policy makers. Along those lines, the Article constructs a framework for determining when policy makers should prefer patents to secrets, and vice versa. The framework is modeled upon the patent reward theory, which explains the existence of the patent system as a means of overcoming the public goods problem of economics. By employing the reward theory model and introducing the concept of differing inventive value for patented and secret inventions, the framework suggests situations in which an inventor's protection preference differs from society's preference.

The framework concludes that policy makers ought not to discourage secrecy, as the law currently does. Because secrecy is a market inefficiency that (at times) permits inventors to amortize their investment costs, there is no risk of reduced innovation when inventors choose trade secret protection: inventor choice of protection scheme is the best means of eliminating the free rider problem. Furthermore, in light of the societal costs of patenting, the law should actually encourage secrecy over patenting in certain circumstances. By encouraging secrecy in certain cases, policy makers can better balance the innovation incentives of the patent system.

This Article proposes two categories of doctrinal changes in light of the constructed framework. First, the Article urges the elimination of patent doctrines that discourage the use of secrecy. The adoption of prior user rights is recommended, along with changes to the standards used in patent priority disputes. Second, this Article begins to examine potential ways in which the law can actively encourage the use of secrecy. A potential limitation of patent subject matter is examined. Additionally, this Article

13. *See* Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 TEX. L. REV. 1031, 1054 (2005) [hereinafter Lemley, *Free Riding*] ("In a private market economy, individuals will not generally invest in invention or creation unless the expected return from doing so exceeds the cost of doing so—that is, unless they can reasonably expect to make a profit from the endeavor.").

proposes the creation of a secret invention registry. The registry would encourage the use of secrecy by lowering the private cost of enforcing trade secret rights.

II. PATENT LAW AND SECRECY

Innovators encounter a diverse array of legal means to appropriate their innovations.[14] Patents, copyrights, trademarks, and trade secrets offer different legal protection mechanisms for different types of inventions, works, or commercial marks. Some innovations can only be protected by one form of intellectual property.[15] The Nike "swoosh," for example, is protectable only as a trademark. An artistic work, such as *Gone with the Wind*, is protected under copyright law but cannot be patented.[16] Some innovative subject matter falls within the ambit of multiple protection regimes. For example, software can often be protected with a utility patent, a copyright, or both. Ornamental designs, such as designer candle holders, can be patented (by a design patent), copyrighted, and/or trademarked.[17]

While the overlap between copyright and patent at times permits two forms of protection, there is one intellectual property overlap that requires an

14. Of course, intellectual property protection is not required when innovation occurs. Alternatively, an innovator may decide to forgo legal protection for her invention and instead disclose her innovation to the public free of charge. Such disclosure may occur for numerous reasons, including enhanced professional reputation, conformity to common industry or community norms, or because of lack of earning potential associated with the innovation. The social and private benefits of voluntary disclosure are outside the scope of this article. For more on open disclosure's benefits for peer production, see generally Yochai Benkler, *Coase's Penguin, or Linux and the Nature of the Firm*, 112 YALE L.J. 369 (2002) (arguing that open source production models are superior to hierarchical models of production because of open source's more efficient acquisition and processing of human capital availability).

15. This assumes that trade secrets are considered intellectual property, a proposition that has received substantial attention and criticism. For more on trade secrecy's place in the intellectual property universe, see generally Lemley, *supra* note 6 (arguing that trade secrets should be considered forms of intellectual property); Michael Risch, *Why Do We Have Trade Secrets?*, 11 MARQ. INTELL. PROP. L. REV. 1 (2007) (arguing that trade secrets are not intellectual property because the justification for trade secret law is not based on incentives to invent).

16. Although there are now patent applications for movie plot lines pending at the PTO, the Supreme Court's recent decision in *Bilski v. Kappos* likely dooms those applications as unpatentable "abstract ideas." Bilski v. Kappos, 130 S. Ct. 3218 (2010) (holding a method of hedging risk unpatentable as an abstract idea).

17. U.S. PATENT & TRADEMARK OFFICE, U.S. DEP'T OF COMMERCE, MANUAL OF PATENT EXAMINING PROCEDURE (MPEP) § 1512.I (8th ed. Rev. 8, July 2010) ("There is an area of overlap between copyright and design patent statutes where the author/inventor can secure both a copyright and a design patent.").

innovator to choose between protection schemes: that of trade secret and patent law.[18] An innovator who chooses to patent cannot simultaneously enjoy trade secrecy because the patent application reveals her secret to the world.[19] Similarly, the choice to maintain an invention as a long-term secret precludes patenting that invention.[20] Thus, since the law precludes inventors from receiving simultaneous patent and trade secret protection, an inventor must select the regime that provides the best protection for the particularities of her invention.[21]

A. PATENTS AND TRADE SECRETS: LEGAL DIFFERENCES

An invention is eligible for patenting at the moment it is "reduced to practice" or when an inventor produces descriptions of the invention that enable a skilled artisan to practice the invention.[22] "Reduction to practice" can occur in one of two ways: constructive (which occurs upon filing a patent application) or actual.[23] Actual reduction to practice requires that an invention work for its intended purpose.[24] When an invention is ready for patenting, an inventor can choose to patent or to continue working in secret. The window for patenting ends one year after the innovation has been commercialized or used publicly.[25]

---

18. *Compare* Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 487–88 (1974) (discussing selection of patenting or secrecy), *with In re* Yardley, 493 F.2d 1389, 1395–96 (C.C.P.A. 1974) (holding that the constitutional provision distinguishing copyrights and patents does not "require[] an election" of one form of protection over the other).

19. There is a delay between the filing of a patent application and the publication thereof during which an invention may be considered both patented and secret. Secrecy in this case expires upon publication, which typically occurs eighteen months after filing. 35 U.S.C. § 122(a)–(b)(1)(A) (2006).

20. *Id.* § 102(b) (2006).

21. It should be noted that the choice presented so far is somewhat stylized. The literature reveals that the choice between patent and secrecy is often not an all-or-nothing choice. *See* Karl F. Jorda, *Patent and Trade Secret Complementariness: An Unsuspected Synergy*, 48 WASHBURN L.J. 1, 31 (2009). Inventors often will employ a hybrid strategy in protecting their invention in which they patent some aspects of their invention and maintain other aspects as trade secrets. This strategy provides some of the benefits of both protection schemes. For instance, patent infringement suits are available as a remedy, yet the risk of free riding from the patent document itself is reduced due to the presence of the trade secret. Part IV, *infra*, deals with this hybrid strategy in more detail.

22. Pfaff v. Wells Elecs., Inc., 525 U.S. 55, 67 (1998).

23. Hybritech Inc. v. Monoclonal Antibodies, Inc., 802 F.2d 1367, 1376 (Fed. Cir. 1986).

24. *Id.*

25. 35 U.S.C. § 102(b). Similarly, an invention is ineligible for patent protection one year after the invention appears in a qualifying publication.

In making the patent/trade secret election, inventors must consider the different scope and strength of protection offered by the two regimes.[26] The legal protection offered by a patent differs from that offered by trade secrecy in four fundamental ways. First, a trade secret has a potentially limitless lifespan, while a patent is constitutionally time-limited. The Constitution requires that patents be granted "for a limited term."[27] Currently, the right to exclude that a patent provides can extend for up to twenty years. Trade secrets, on the other hand, allow inventors to exclude others for as long as secrecy continues.

The most famous example of trade secrecy's duration is the formula for Coca-Cola syrup. The Coca-Cola Company protects the formula for Coke as a trade secret, and it has been doing so for over a century.[28] Had the company instead chosen to patent the formula it would have been forced to disclose the secret to the world and been unable to exclude others from copying that formula over the past eighty years.[29]

However, the potentially limitless life of a trade secret comes with a risk. The second fundamental difference between patent and trade secret protection is that a secret invention has a narrower exclusionary scope. A patent permits a patentee to exclude any unauthorized use of the invention, even if the invention was independently developed.[30] A trade secret, on the other hand, only provides a legal remedy against misappropriation of the secret.[31] Secret inventions risk discovery through independent invention or

26. This Article is concerned only with those innovations that are potentially patentable. Trade secret law's subject matter is more inclusive than patent law. A trade secret can encompass information that is neither new nor non-obvious and is therefore ineligible for patent protection. *See, e.g.*, Risch, *supra* note 15, at 11, 12 (citing 35 U.S.C. §§ 101–103 (2000)). Similarly, unoriginal information—such as names and phone numbers—can be protected as a trade secret but is not eligible for patent protection. *Id.* at 12. This Article is concerned with only the subset of potential trade secrets that would also be eligible for patent protection.

27. U.S. CONST. art. I, § 8, cl. 8.

28. *See* Coca-Cola Bottling Co. v. Coca-Cola Co., 107 F.R.D. 288 (D. Del. 1985) (mentioning that the trade secret 7X formula was kept in bank vault that could only be accessed by a board resolution).

29. Other famous culinary trade secrets include Colonel Sanders' original recipe for fried chicken and McDonald's original special sauce. Many of these trade secrets would not be eligible for patent or copyright protection because recipes are ineligible for any sort of intellectual property protection. *See, e.g.*, Emily Cunningham, *Protecting Cuisine Under the Rubric of Intellectual Property Law: Should the Law Play a Bigger Role in the Kitchen?*, 9 J. HIGH TECH. L. 21 (2009) (noting that any intellectual property protection for methods of food preparation is unlikely to be used, although patent law is arguably available).

30. *See, e.g.*, Samson Vermont, *Independent Invention as a Defense to Patent Infringement*, 105 MICH. L. REV. 475, 480 (2006) (arguing for an independent inventor defense in patent law).

31. *Id.*

reverse engineering.[32] Thus, if one were to independently stumble upon the secret Coke formula, or if one were to reverse engineer the formula via chemical testing, nothing prevents the discoverer from commercializing that formula.

The third fundamental difference in the protection regimes is that a secret invention requires no legal formalities to obtain exclusionary rights.[33] Filing for a patent involves a lengthy, expensive process. Patent attorneys draft stylized legal documents that are required to describe the metes and bounds of the patent, disclose the invention in a way that permits a "person skilled in the art to make and use the invention without undue experimentation,"[34] and demonstrate that the invention is novel, non-obvious, and useful.[35] The attorney then files the document with the U.S. Patent and Trademark Office. The initial filings are often rejected, necessitating amendments to the original filings in the form of a continuation, further filings demonstrating the patentability of the invention, or other negotiations with the patent office.[36] This cycle can continue ad infinitum. Patents cost upwards of $10,000 to file and are expensive to maintain.[37]

Trade secrets, on the other hand, require no formal registration with the government.[38] Instead, trade secret protection requires owners to invest in "reasonable measures" to keep the secret.[39] Thus, secret inventions require a measure of self-help in order to exclude. By choosing secrecy, inventors avoid the cost of obtaining a patent and the risky, costly business of patent enforcement. Patent litigation is an extremely costly undertaking and requires patent holders to monitor competitors for infringement, which can be quite costly and difficult depending on the visibility of the invention.[40] However, trade secrets carry their own set of costs: negotiations and relationships must

32. UNIF. TRADE SECRETS ACT § 1 cmts. 1–2 (amended 1985).

33. *Id.* § 1 (listing the requirements for a trade secret).

34. *See In re* Wands, 858 F.2d 731, 735 (Fed. Cir. 1988).

35. 35 U.S.C. §§ 101–103.

36. *See* CRAIG A. NARD, THE LAW OF PATENTS, 41–42 (2d ed. 2011) (explaining the typical process of obtaining patent rights).

37. Mark A. Lemley, *Rational Ignorance at the Patent Office*, 95 NW. U. L. REV. 1495, 1498–1500 (2001) (estimating 2001 costs).

38. *See* UNIF. TRADE SECRET ACT § 1 (listing the requirements for a trade secret).

39. *See* ROGER M. MILGRIM, MILGRIM ON TRADE SECRETS §§ 1.03–.04 (1996).

40. Upwards of $7 billion was spent on legal fees surrounding patent litigation and patent prosecution in 2001. Lemley, *supra* note 37, at 1498–1503.

be closely monitored and controlled through non-disclosure agreements, employee confidentiality agreements, and physical protection.[41]

Lastly, and perhaps most importantly for innovation policy, trade secrets differ from patented inventions in the amount of disclosure that is legally required (or permitted) to protect an invention. A secret, by its nature, cannot be broadly disclosed. Once a trade secret is widely known, it no longer qualifies for legal protection.[42] Conversely, a patented invention must be fully disclosed to the public. The patent document itself must "enable" a skilled artisan to practice the invention.[43]

Oftentimes an inventor will desire disclosure, either through a patent or a published article. But at other times an inventor may wish to keep her invention secret, either to maximize profit, minimize competition, or to conduct further research and development before choosing whether to disclose. However, such secrecy is not always feasible.[44] Often, disclosure of an invention is tied to commercialization. For example, an improved pop-top soda can is effectively disclosed once on the market; secrecy of the commercialized product is virtually impossible.

Secrecy is often used by inventors as a means of appropriating an invention.[45] In fact, in certain circumstances it is preferred to patenting because it is a more effective means of securing profits from an innovative idea.[46] Numerous surveys demonstrate that inventors in fields in which secrecy is feasible view secrecy as the more effective method of appropriating their inventions. For example, using historical data from the Crystal Palace World's Fair, Petra Moser has shown that in fields where secrecy is feasible,

41. *See, e.g.*, Elizabeth Rowe, *When Trade Secrets Become Shackles: Fairness and the Inevitable Disclosure Doctrine*, 7 TULANE J. TECH. & INTELL. PROP. 167, 201–03 (2005).

42. *See* MILGRIM, *supra* note 39, § 1.05.

43. 35 U.S.C. § 112. In *Ariad v. Lilly*, the Federal Circuit held that Section 112 requires that a patentee both teach one skilled in the art the manner of practicing the invention (enablement) and demonstrate that the inventor possesses the claimed invention (written description). Ariad Pharm., Inc. v. Eli Lilly & Co., 598 F.3d 1336, 1344 (Fed. Cir. 2010).

44. *See* Timothy R. Holbrook, *Possession in Patent Law*, 59 SMU L. REV. 123, 133–34 (2006).

45. *See* Anthony Arundel & Isabelle Kabla, *What Percentage of Inventions Are Patented? Empirical Estimates of European Firms*, 27 RES. POL'Y 127 (1998); Wesley Cohen et al., *Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)* 17 (Nat'l Bureau of Econ. Research, Working Paper No. 7552, 2000) (estimating the patent propensity rate to be 54% for product innovations and 27% for process innovations based on a survey of 1,478 R&D labs in the United States).

46. *See* Cohen et al., *supra* note 45, at 17; Richard C. Levin et al., *Appropriating Returns from Industrial R&D*, 3 BROOKINGS PAPERS ON ECON. ACTIVITY 783, 795 (1987) (reporting results from a survey of high-level R&D executives finding that secrecy was "considered more effective than patents in protecting processes").

inventors typically rely on trade secret protection.[47] However, over time, as reverse engineering becomes less costly, inventors increasingly turn to patent protection.[48]

Thus the difference in protection preference is largely attributable to the feasibility of secrecy. Such feasibility differs among invention types and inventive industries. Based on a 1994 survey of research labs at 1,478 U.S. manufacturing companies, Wesley Cohen et al. found that a wide range of industries considered secrecy to be the most effective method of appropriating the value of an invention. Those industries included food, textiles, paper, petroleum, all chemical industries, rubber, plastics, mineral products, metals, machine tools, electrical equipment, motors, generators, semiconductors, and navigation instruments.[49] Furthermore, secrecy was "clearly the most effective" method of securing an invention for process innovators, while patents were less effective for process than for product innovations.[50] A considerable number of similar surveys confirm Cohen et al.'s results.[51]

For example, the 2008 Berkeley Patent Survey found that certain industries (primarily the software, internet, manufacturing, and chemical processing industries) perceive patenting to be among the least important means of capturing a competitive advantage.[52] This stratification of value across industry and innovation type indicates that secrecy is more valuable in industries in which it is available (including software, manufacturing, chemicals) and certain invention types that are less revealing (methods and processes) while patents provide more private value for other industries (pharmaceuticals, consumer products) and invention types (product innovations).

47. Petra Moser, Innovation Without Patents—Evidence from World Fairs (July 16, 2010) (unpublished manuscript), http://www.ssrn.com/abstract=930241.

48. *Id.*

49. Cohen et al., *supra* note 45, at 10 n.21.

50. *Id.* at 10.

51. *See, e.g.*, Stuart J.H. Graham et al., *High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey*, 24 BERKELEY TECH. L.J. 1255 (2009); Levin et al., *supra* note 46, at 799 (finding that secrecy was preferred to patenting in the process innovation industry, whereas patents were more valuable to product innovators).

52. *See* Graham et al., *supra* note 51, at 1285–87. The other strategies in the Berkeley Patent Survey were secrecy, first-mover advantage, copyrights, trademarks, reverse engineering, and complementary assets. *Id.* at 1289.

B. DISCOURAGING SECRECY

With few exceptions, the existence of the United States patent system is justified on utilitarian grounds.[53] The patent system's goal of stimulating innovation is manifested in the U.S. Constitution's articulation of Congress's power to "promote the Progress of Science and the useful Arts by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."[54] The reward theory is the predominant theoretical explanation for how the patent system promotes the progress of the useful arts.

Reward theorists justify the patent system as a means of inducing the creation and disclosure of new and useful inventions. Without the supranormal profits obtainable with a patent, the theory predicts that many inventions would remain undiscovered or shrouded in secret.[55] Reward theory has thus been characterized by courts as a quid pro quo between society and the inventor. In exchange for society's offer of patent protection, the inventor must disclose her invention to the public. These dual incentives acting upon an inventor—the incentive to invest and the incentive to disclose—form the basis for the reward theory. The goal is that in exchange for a twenty-year period of exclusivity, inventors will be incentivized to create new and useful inventions and then reveal those innovations to the public. This Section briefly discusses the quid pro quo view of the patent system and the rhetoric employed by the courts to discourage secrecy. Then, moving beyond the legal rhetoric, this Section describes how patent doctrine discourages inventors from relying on trade secrecy.

1. The Rhetorical Distaste for Secrecy

Because it requires disclosure, patent law precludes simultaneous protection of an invention as both a patent and a trade secret. Beyond the structural rejection of secrecy, patent law has traditionally been conceptualized by courts as a means of prying secret inventions from their

53. *See, e.g.*, Lemley, *supra* note 6, at 329; Katherine J. Strandburg, *Experimental Use and the Patent Bargain*, 2004 WIS. L. REV. 81, 90–91 (2004). Non-utilitarian theories proffered for the existence of the patent system include those grounded in moral rights. *See, e.g.*, Lawrence C. Becker, *Deserving To Own Intellectual Property*, 68 CHI.-KENT L. REV. 609, 619–29 (1993) (evaluating the patent system through a desert-for-labor argument).

54. U.S. CONST. art. I, § 8, cl. 8.

55. Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEX. L. REV. 989, 993–1000 (1997); Sichelman, *supra* note 12, at 358.

inventors and disclosing those inventions to the public.[56] The patent quid pro quo has provided courts with a rhetorical narrative to account for the existence of the patent system.[57]

The traditional view envisions the possibility of a patent as the carrot that is offered to inventors operating in secret.[58] Viewed from this perspective, the patent system's primary goal is to offer a reward that will incentivize inventors to disclose their secrets to the public. The price that the public pays for the revelation of secret inventions is the property-like exclusive rights of patent protection. The bargain is viewed as beneficial to society because society pays for secrets by giving up what it otherwise would not possess—a description of the invention and the right to eventually practice that invention once the patent expires.[59]

The rhetoric surrounding the quid pro quo generally emphasizes the social benefits of patent disclosure.[60] On this view, secret inventions are unlikely to be revealed or disclosed unless a reward (in the form of a patent) is offered to possessors of such inventions.[61] For this rhetorical conceptualization to make sense, one must assume that inventions benefit society more as revealed patents than as concealed trade secrets. This view is understandable: society obviously benefits from having valuable knowledge disclosed. However, the assumption that patents are always preferable to secrets fails to account for the societal costs that accompany the patent grant.

Perhaps the best example of the degree to which the rhetoric of the patent quid pro quo has influenced the courts' view of the value of secrecy is in the 1974 case *Kewanee Oil Co. v. Bicron Corp.*[62] In *Kewanee*, the Supreme Court analyzed whether state trade secret law (Ohio law in this case) was pre-empted by the operation of federal patent law.[63] Kewanee Oil Co. had developed, after significant investment, "many processes, procedures, and

56. *See generally* Mason v. Hepburn, 13 App. D.C. 86, 96 (D.C. Cir. 1898) (stating that the patent law's concealment doctrine is designed to favor patentees over trade secret holders).

57. *Id.*

58. *See* Benjamin Roin, Note, *The Disclosure Function of the Patent System (or Lack Thereof)*, 118 HARV. L. REV. 2007, 2010 (2005).

59. *See, e.g.*, Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 535 U.S. 722, 736 (2002) ("[P]atent rights are given in exchange for disclosing the invention to the public."); W.L. Gore & Assocs. v. Garlock, Inc., 721 F.2d 1540, 1550 (Fed. Cir. 1983) ("Early public disclosure is a linchpin of the patent system.").

60. *See* Roin, *supra* note 58, at 2012.

61. Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 487–88 (1974).

62. *Id.*

63. *Id.* at 472.

manufacturing techniques" in the growth of crystals.[64] One such technique enabled the company to grow a seventeen-inch crystal that proved useful for the detection of ionizing radiation, which the company maintained as a trade secret.[65] Kewanee sued some of its former employees for misappropriation of that trade secret.[66] In reversing the Sixth Circuit Court of Appeals, the Supreme Court held that the federal patent system did not preempt the states' ability to protect trade secrets.[67]

With respect to secret inventions, the Court viewed the patent system as specifically designed to draw trade secrets into the public sphere. "[T]he federal interest in disclosure is at its peak" with patentable secret inventions, the Court concluded.[68] In fact, the Court stated that "[t]he interest of the public is that the bargain of 17 years of exclusive use in return for disclosure be accepted."[69] Thus, the Court, without citation or much evaluation, elevated the patent quid pro quo to a status beyond a mere bargain or contract. In the Court's view, the Patent Act operates in large part as a secrecy disclosing mechanism. This expansive view of the traditional patent bargain favors disclosure of secrets without regard to the justification for preferring patents. Indeed, the Court immediately follows its description of the patent quid pro quo with a statement that denies, albeit in dicta, secrecy's ability to coexist with patent law: "If a State, through a system of protection, were to cause a substantial risk that holders of patentable inventions would not seek patents, but rather would rely on the state protection, we would be compelled to hold that such a system could not constitutionally continue to exist."[70] In sum, the Court understood the quid pro quo view of the patent system as potentially in conflict with the practice of trade secrecy.

Kewanee held that trade secrecy was not preempted by the patent system because it did not pose a reasonable risk of deterring the filing of patent applications for reasons which have been roundly criticized by commentators.[71] But the view that the patent system is designed to reduce

64. *Id.* at 473.

65. *Id.*

66. *Id.*

67. *Id.* at 474.

68. *Id.* at 489. Indeed, Katherine Strandburg has shown that it is this class of inventions (those that are (1) eligible for patent protection, (2) more valuable to their owners as trade secrets, and (3) promise profits greater than their development costs), and this class only, that even concern the patent quid pro quo. Strandburg, *supra* note 53, at 110–11.

69. *Kewanee*, 416 U.S. at 489.

70. *Id.*

71. *Id.* at 474, 490–91. For a strong critique of the *Kewanee* Court's reasoning on this point, see Sharon K. Sandeen, Kewanee *Revisited: Returning to First Principles of Intellectual*

the number of secret inventions is a view that courts have embraced, both before and after *Kewanee*.[72]

The traditional view of the patent bargain emphasizes patents as preferable to secrets. Trade secrets are seen as potential targets of the patent system, rather than as a potentially complementary form of intellectual property. The rhetorical heft of the patent quid pro quo tends to cloud the tradeoffs inherent in patent protection. Furthermore, reliance on this view of the patent quid pro quo *may* obscure any potential benefits that might inure to the public by encouraging secrecy. A more complete theoretical understanding of the costs and benefits of secrecy should be employed when attempting to craft the proper incentives for innovation policy.

To be sure, some courts and a majority of patent scholars have framed the quid pro quo as a choice rather than as a policy lever. Scholars recognize that the patent system benefits society not merely because of the increased disclosure that results from patenting, but also (and primarily) because of the incentive to invent that the patent system creates.[73] Any disclosure benefit from patenting, on this view, is secondary to the benefit of increased amounts of innovation that result from the patent bargain. This view, widely adopted by commentators, has been the minority viewpoint for courts.

2. The Doctrinal Distaste for Secrecy

The conceptual distaste for secrecy in patent jurisprudence is also reflected in patent doctrine. Carl Shapiro noted that "the current patent system rewards applicants who are most aggressive in seeking patents over

Property Law To Determine the Issue of Federal Preemption, 12 MARQ. INTELL. PROP. L. REV. 299, 345–47 (2008).

72. *See, e.g.*, Russo v. Ballard Med. Prods., 550 F.3d 1004, 1012 (10th Cir. 2008) ("Federal law expresses a strong interest in seeing that patentable innovations do not stay bottled up in secret but are instead shared with the public in order to promote social progress. This interest is most obviously embodied in patent law's bargain of providing inventors with many years of monopoly rents in return for the public's opportunity to use and enjoy their ideas."); Mason v. Hepburn, 13 App. D.C. 86, 96 (D.C. Cir. 1898) ("The true ground of the [concealment] doctrine, we apprehend, lies in the policy and spirit of the patent laws and in the nature of the equity that arises in favor of him who gives the public the benefit of the knowledge of his invention, who expends his time, labor, and money in discovering, perfecting, and patenting, in perfect good faith, that which he and all others have been led to believe has never been discovered, by reason of the indifference, supineness, or wilful act of one who may, in fact, have discovered it long before."); *see also* Brief for the United States as Amicus Curiae at 13, Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470 (1974) (No. 73-187), 1974 WL 185610, at *13 ("Since election by inventors to rely on trade secret law reduces disclosure by diverting inventions away from the patent system, technological progress is slowed, contrary to the goal of federal patent policy.").

73. Roin, *supra* note 58, at 2012.

those who simply use their own inventions internally as trade secrets."[74] Patent law encourages inventors to patent secret inventions in three ways. First, statutory bars to patentability require the prompt filing of a patent application.[75] Failure to file within the time period allotted by statute will result in the loss of rights in the invention. Second, during patent priority disputes (interferences), a first inventor can lose the rights in her invention if she is found to have "abandoned, concealed, or suppressed" the invention.[76] These priority rules favor an inventor who does not employ secrecy, even if that inventor was second in time. Third, a first inventor who chooses to maintain her invention as a secret can be liable for infringement to a second inventor who decides to patent.[77] The relative lack of "prior user rights" for secret inventors is unique to the American patent system.[78]

a) Statutory Bars to Patentability

An inventor loses all rights to patent an invention if she does not file a patent application within one year of the invention being "in public use or on sale" in the United States.[79] Courts have set a very low standard for what constitutes public use. The Federal Circuit has declared that public use includes "any use of that invention by a person other than the inventor who is under no limitation, restriction or obligation of secrecy to the inventor."[80] Under this standard, public use can encompass some unrevealed uses of an invention.[81] The use of an invention in secret—if connected to commercial exploitation of the invention—may be considered public use by the courts and the PTO.[82] Public use cases largely turn on factual issues, and courts often focus their inquiry in an evaluation of the purposes of the public use bar.[83] Among these purposes is the encouragement of prompt patent filing.[84]

74. Shapiro, *supra* note 5, at 95; *see also* Margo A. Bagley, *The Need for Speed (and Grace): Issues in a First-Inventor-To-File World*, 23 BERKELEY TECH. L.J. 1035, 1049 (2008) (discussing the U.S. first-to-invent system's "default preference" for patents over trade secrets).

75. 35 U.S.C. § 102(b) (2006).

76. § 102(g).

77. *See infra* Section II.B.2.c.

78. For information on exceptions to this general rule, see *infra* note 96.

79. § 102(b).

80. *In re* Smith, 714 F.2d 1127, 1134 (Fed. Cir. 1983).

81. MPEP, *supra* note 17, § 2133.03(a).II.A.1; *see also* Egbert v. Lippmann, 104 U.S. 333, 336 (1882) (stating that public use may occur "even though the use and knowledge of the use may be confined to one person").

82. MPEP, *supra* note 17, § 2133.03(a).II.A.1; *see also* TP Labs, Inc. v. Prof'l Positioners, Inc., 724 F.2d 965, 972 (Fed. Cir. 1983) (noting that secret but commercial use of an invention could constitute "public use" under the statute).

83. *See, e.g.*, *Egbert*, 104 U.S. at 338 (holding that a corset patent was invalid because of the public-use bar); Motionless Keyboard Co. v. Microsoft Corp., 486 F.3d 1376, 1385 (Fed.

Courts have also lowered the threshold requirements for what constitutes "on sale." In *Pfaff v. Wells Electronics, Inc.*, the Supreme Court affirmed a decision that the on-sale bar had been triggered by an offer to sell, even though the invention was not reduced to practice at the time of the offer.[85] Thus, a sale of an as-yet-unreduced-to-practice invention can begin the one-year time period in which an inventor has to patent an invention.[86] For inventive methods, the sale of a product tied to the inventive method may trigger the on-sale bar even if the method remains concealed.[87] The on-sale and public use bars encourage inventors to file for patents at an early stage of an invention's development. *Pfaff*'s holding discourages inventors from maintaining inventions as trade secrets while conducting initial commercial activities because the time period for filing a patent begins once an offer for sale has been made.

b) Patent Priority Rules

When two inventors apply for a patent on the same invention, a complex set of priority rules governs who will receive the patent. Priority is determined at the filing stage in a proceeding called an interference. A similar set of rules governs whether a prior invention is considered prior art in determining validity at trial. Under American law, the first inventor to file is presumed to be the first to invent.[88] But that presumption is rebuttable upon a showing that another inventor actually reduced the invention to practice prior to the earliest filing date,[89] with several caveats, one of which is described immediately below.[90]

Although the American system generally rewards the first inventor, a second inventor can receive patent protection over a first inventor if she can demonstrate that the first inventor "abandoned, suppressed, or concealed" the invention at any time after the second inventor successfully created the

Cir. 2007) (reversing a finding of invalidity due to the trial court's misapplication of the concept of public use).

84. *See, e.g.*, Nancy S. Paik, *Implied Professional Obligation of Confidentiality Sufficient To Overcome Public Use Defense to a Claim of Patent Infringement?* Bernhardt v. Collezione—*The Federal Circuit Court of Appeals' Surprising Recent Announcement on the Public Use Bar*, 4 CHI.-KENT J. INTELL. PROP. 332, 333–34 (2005).

85. 525 U.S. 55 (1998).

86. *Id.*

87. *See In re* Caveney, 761 F.2d 671, 675–76 (Fed. Cir. 1985) (stating that sale of the product of a secret method triggers the on-sale bar).

88. *See, e.g.*, Charles L. Gholz, *First To File or First To Invent?*, 82 J. PAT. & TRADEMARK OFF. SOC'Y 891 (2000).

89. 1 DONALD S. CHISUM, CHISUM ON PATENTS § 10.03[1][C][1] (2006).

90. *See infra* Section II.B.2.c.

invention.[91] All three statutory terms reflect a single concept of an inventor who fails to patent an invention or commercialize it, or both.[92] For the PTO to consider an invention concealed, it need only be shown that an inventor did not take active steps to make an invention publicly known. Courts have interpreted concealment to mean that within a reasonable amount of time, no steps have been taken by the inventor to make the invention publicly known.[93] Public knowledge may occur through a patent application, a public announcement, or public use.[94] Not only can an inventor who both invented and filed for a patent first lose all rights to her invention, she also may not be able to use her prior invention to invalidate the second inventor's patent.[95]

c) Prior User Rights

The third aspect of the patent system that discourages secrecy is U.S. patent law's general lack of prior user rights. Unlike trade secret holders in a majority of countries,[96] U.S. inventors do not generally possess prior user rights in their inventions.[97] That is, a first inventor that practices her invention in secret cannot use her prior use and invention as a defense

91. 35 U.S.C. § 102(g) (2006).

92. *See* 1 CHISUM, *supra* note 89, § 10.08[1].

93. Correge v. Murphy, 705 F.2d 1326, 1330 (Fed. Cir. 1983).

94. *Id.*

95. *In re* Suska, 589 F.2d 527, 529 (C.C.P.A. 1979) ("The result of applying the suppression and concealment doctrine is that the inventor who did not conceal (but was the 'De facto' last inventor) is treated legally as the first to invent, while the 'De facto' first inventor who suppressed or concealed is treated as a later inventor."). This surprising result affects both who is entitled to the patent and whether a first invention constitutes prior art and invalidates another's later-issued patent. *Id. But see* Dunlop v. Ram, 524 F.2d 33 (7th Cir. 1975) (holding that "secret" use of a machine or process is "public" if the details of the machine or process are ascertainable by inspection or analysis of the product that is sold or publicly displayed). Robert Merges explains the difference in the case law by focusing on the inventor's actions rather than the nature of the invention. When inventors intentionally conceal (as opposed to merely possess non-revealing technology) courts tend to consider the invention suppressed. ROBERT P. MERGES, PATENT LAW & POLICY 461 (4th ed. 2007).

96. *The Patent Prior User Rights Act and the Patent Reexamination Reform Act: Hearing on S. 2272 and S. 2341 Before the Subcomm. on Patents, Copyrights and Trademarks of the S. Comm. on the Judiciary*, 103d Cong. 24 (1994) (statement of Roger S. Smith, President, Intellectual Property Owners, and Assistant General Counsel for Intellectual Property Affairs, IBM) (noting that prior user rights are "common" in foreign countries and that a WIPO study found that the "vast majority" have such rights).

97. A prominent exception to the lack of prior user rights in the United States was adopted by Congress in 1999. 35 U.S.C. § 273 (2006). The 1999 "safe-harbor" was granted to protect companies with business method trade secrets that feared a rush of patenting after the Federal Circuit's decision in *State Street Bank & Trust Co. v. Signature Financial Group, Inc.*, 149 F.3d 1368 (Fed. Cir. 1998). Robert C. Haldiman, *Prior User Rights for Business Method Patents*, 20 ST. LOUIS U. PUB. L. REV. 245, 246 (2001).

against a subsequent patentee.[98] For example, consider the case of Inventor 1 inventing an improved method of manufacturing widgets. If she chooses to maintain her invention as a secret, she loses her right to patent that invention one year after she sells the invention or puts it into public use.[99] Those priority rules limit Inventor 1's ability to claim exclusive rights in her invention after certain activities occur.

The lack of prior user rights, however, does more than merely limit Inventor 1's ability to patent her invention; it potentially limits her ability to practice her own invention. If Inventor 2 discovers and patents the method of manufacturing widgets, she can sue Inventor 1 for infringement.[100] Inventor 1's earlier invention and use is not a valid defense to patent infringement. Therefore the possibility of infringement liability hangs over the head of a first inventor if she chooses to practice her invention in secret.[101] The threat of infringement liability to a subsequent inventor can be a powerful deterrent against keeping an invention secret.[102]

III. EXAMINING SECRECY

Patent law's doctrinal discouragement of secrecy attempts to influence inventors at the margins to patent rather than maintain trade secrets. Courts routinely reference the patent quid pro quo without examination of the theoretical relationship between patents and trade secrets.[103] Reliance upon the accepted wisdom of patents as preferable to secrets obscures the benefits that secrecy might provide. This Part fills the theoretical lacuna of the patent/trade secret trade-off from a societal perspective. In doing so, it examines the potential benefits and drawbacks of secrecy when compared to patenting.

---

98. *See* JAMES BESSEN & MICHAEL J. MEURER, PATENT FAILURE: HOW JUDGES, BUREAUCRATS, AND LAWYERS PUT INNOVATORS AT RISK 32–34 (2008).

99. *See supra* Section II.B.2.a.

100. This assumes, of course, that Inventor 2 did not misappropriate the invention. If so, Inventor 1 may have a claim for misappropriation.

101. *See, e.g.*, Gillman v. Stern, 114 F.2d 28, 30 (2d Cir. 1940).

102. *See* Vincenzo Denicolò & Luigi A. Franzoni, *Patents, Secrets, and the First Inventor Defense*, 13 J. ECON. MGMT. & STRATEGY 517, 518–19 (2004).

103. *See, e.g.*, Bonito Boats, Inc. v. Thunder Craft Boats, Inc., 489 U.S. 141, 148 (1989) ("Thus, from the outset, federal patent law has been about the difficult business 'of drawing a line between the things which are worth to the public the embarrassment of an exclusive patent, and those which are not.' ") (quoting 13 THE WRITINGS OF THOMAS JEFFERSON 335 (definitive ed. 1907)); United States v. Dubilier Condenser Corp., 289 U.S. 178, 186–87 (1933) ("Thus a monopoly takes something from the people. An inventor deprives the public of nothing which it enjoyed before his discovery, but gives something of value to the community by adding to the sum of human knowledge.").

This Part proceeds in three Sections. Section III.A examines the theoretical economic justification for patenting (and intellectual property more generally) and the relationship that justification has with secrecy. It concludes that the economic concern motivating the patent system—the public goods market failure—is inapplicable when secrecy is an available appropriation means. Thus, as a theoretical matter, the economic justification for conferring patent protection on secrecy-appropriable inventions is lacking. Section III.B considers the traditional support offered by courts and theorists for discouraging secrecy, namely the benefits of disclosure and coordination. In doing so, this Section examines the ability of trade secrecy to perform the same beneficial functions often attributed to the patent system. Lastly, Section III.C describes the overlooked potential benefits of secrecy.

A. SECRET INVENTIONS AND PUBLIC GOODS

Unlike real property, intellectual property law (and patent law more specifically) is concerned with intangible assets. Unlike land, which can be depleted by overuse, information can be reused infinitely with no depletion to the resource. Thus, economists often refer to information as a pure "public good."[104] Public goods are those goods which are nonexcludable (cannot easily be excluded from others' use) and nonrivalrous (consumption by one person does not deplete the resource). Information is analogized to a public good because it can be easily copied (nonexcludable) and used by an infinite number of individuals (nonrival).[105]

Lighthouses are the classic example of a public good.[106] The light from the lighthouse is essentially nonrivalrous: the use of the light by one ship does not diminish the value of the light to another ship. The light is also nonexcludable: a lighthouse owner cannot easily exclude users from using the lighthouse's services—those that refuse to pay for a lighthouse's guidance (free riders) enjoy the same benefits as those that pay. Thus, the nonexcludability of the light drives the price down to a point where a

104. *See* 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, 614–16 (Nat'l Bureau of Econ. Research ed., 1962).

105. *Id.*; Mark A. Lemley, *Ex Ante Versus Ex Post Justifications for Intellectual Property*, 71 U. CHI. L. REV. 129 (2004) ("Ideas are public goods: they can be copied freely and used by anyone who is aware of them without depriving others of their use."); Lemley, *Free Riding*, *supra* note 13, at 1050–51.

106. Christopher Yoo, *Copyright and Public Good Economics: A Misunderstood Relationship*, 155 U. PA. L. REV. 635, 644 & n.25 (2007).

lighthouse owner cannot profitably provide the good, even though there is demand.

The economic literature predicts that public goods such as lighthouses and innovative information will be under-produced in an unregulated market because of the difficulty, if not impossibility, of capturing the positive externalities of such goods.[107] For example, because she cannot capture the social benefit that her light creates, a lighthouse owner will operate her lighthouse at less than the socially optimal level. Analogizing lighthouses to information predicts the same result: inventive information will be suboptimally produced because competitors will be able to copy the information at low cost and drive the market for innovative products towards marginal cost. Rational inventors will not invest heavily in producing new knowledge when they know that they will be unable to recoup their investment costs.

The economic justification for the patent system depends upon the public goods rationale[108]: because inventions are nonexcludable, new inventions will be suboptimally produced absent patent protection. Economists see the patent system as a means of transforming nonexcludable public goods (inventive ideas) into private goods (patented inventions).[109] Thus, the innovation market failure is overcome with the promise of supranormal profits via the excludability offered by a patent.[110]

However, the public goods analogy fails to justify granting patents when secrecy is available. There are two reasons for this imperfect fit. First, inventions that are appropriable as trade secrets do not suffer from the excludability problems associated with pure public goods; such inventions are excludable via secrecy itself.[111] Inventions that can be appropriated as trade secrets do not, by definition, require non-market incentives—such as the

107. *See* A.C. PIGOU, THE ECONOMICS OF WELFARE 331 (4th ed. 1938) (describing how externalities can result in suboptimal production of goods); *see also* Francis M. Bator, *The Anatomy of Market Failure*, 72 Q.J. ECON. 351, 370 (1958) (describing market failures based upon nonappropriability).

108. *See, e.g.*, Holbrook, *supra* note 44, at 132; Lemley, *Free Riding*, *supra* note 13, at 1053 ("[T]he basic economic justification for intellectual property law comes from . . . the risk that creators will not make enough money in a market economy to cover their costs.").

109. *See* Arnold Plant, *The Economic Theory Concerning Patents for Inventions*, *reprinted in* SELECTED ECONOMIC ESSAYS AND ADDRESSES 36 (Inst. of Econ. Affairs ed., 1974).

110. Lemley, *supra* note 55, at 993–1000 (stating the traditional economic argument for the patent system's ability to overcome the public goods market failure).

111. *See* Lemley, *Free Riding*, *supra* note 13, at 1057 ("Economic theory offers no justification for rewarding creators anything beyond what is necessary to recover their average total costs.").

patent system—to attract investment ex ante.[112] An inventor can amortize her investment via secrecy rather than the patent system. This is true precisely because secret inventions enjoy the natural market inefficiency of secrecy. That is, because the information contained in a secret invention is not widely known, holders of such knowledge can often obtain greater profits than would be possible if the information were public.

Mark Lemley argues that the presence of market imperfections, such as secrecy, does not change the public good nature of information.[113] I take no quarrel with that position. However, to the extent that patent theory is concerned with overcoming free riding, the question is not whether information is a public good, but rather whether the problems that public goods create (nonexcludability in this case) can be overcome. When secrecy is selected by an inventor to protect an innovation, we can presume the absence of a public goods market failure. In fact, by rewarding innovators the market is operating as desired. The choice to maintain an invention as a trade secret indicates an inventor's belief that her up-front research costs can be recouped outside of the patent system. In such a case, secrecy circumvents the public goods problem without resorting to patenting.

Certainly, there are cases in which inventions will be under-produced in the absence of the patent system's legally-sanctioned excludability. In some instances, the patent system may be the only means of profitably excluding free riders. But the patent system is not always necessary to foster innovation. Indeed, in many industries, secrecy provides greater incentives to invent, as evidenced by innovator's preference for secrecy.[114]

The second reason that the public goods analogy is inapplicable for cases in which secrecy is feasible is that patents are often a poor means of creating excludability in those cases. Even if we accept the public goods rationale, the patent system must demonstrate that it can overcome the problems associated with public goods, primarily nonexcludability.

However, unlike with real property, detection of infringement of intellectual property boundaries requires knowledge that is often unobtainable to a trade secret holder. Just as secrecy is a market inefficiency

112. F. M. SCHERER, INDUSTRIAL MARKET STRUCTURE AND ECONOMIC PERFORMANCE 444–45 (2d ed. 1980) (noting that natural market imperfections reduce the need for intellectual property protection).

113. Lemley, *Free Riding*, *supra* note 13, at 1052 n.87.

114. In a 1994 survey, secrecy was considered the most effective method of appropriability in a wide, diverse range of industries, including food, textiles, paper, petroleum, all chemical industries, rubber, plastics, mineral products, metals, machine tools, electrical equipment, motors, generators, semiconductors, and navigation instruments. Cohen et al., *supra* note 45, at 10 n.21.

that inventors can leverage to their benefit, it can be a detriment in detecting infringement. Whenever secrecy enables an inventor to hide her inventions from the public, secrecy is also likely to prevent the inventor from detecting infringing use of her invention.[115] If maintaining an invention in secret is realistic, competitors may feel that the risk of being caught infringing a patented invention is negligible. Patents provide little solace for the owner of such an invention. Patent law requires full disclosure but does not guarantee full compliance by competitors. If detection of infringement is difficult or impossible, the patentee has little ability to enforce her rights. Thus, patents are likely to be a poor means of excludability when inventions are truly maintainable as trade secrets.

As shown, inventions that can be carried out profitably in secret do not justify legal interventions into the marketplace to overcome the public goods problem.[116] Because secret inventions are excludable, they do not suffer from the problems associated with public goods. Because of the natural excludability provided by secrecy, utilitarian patent theorists should refuse patents in cases in which secrecy is a viable option for inventors.[117]

The lack of justification for granting patents when secrecy is available might not be problematic if patents were costless. If patents were costless, society might be agnostic to the use of and justification for patents. However, patents are not costless. They result in deadweight losses to consumers because patent holders can charge supranormal rates for patented inventions to the extent that the patent offers the ability to exclude competitors.[118] Ideally, these losses are justified by the value of the innovation that is incentivized by the patent reward. However, in cases where the patent system is not required to encourage research into an invention, such as when secrecy is a viable option, the patent system does not

115. A. Samuel Oddi, *Un-unified Economic Theories of Patents—The Not-Quite-Holy Grail*, 71 NOTRE DAME L. REV. 267, 285 n.126 (1996) (citing Robert P. Merges, *Rent Control in the Patent Districts: Observations on the Grady-Alexander Thesis*, 78 VA. L. REV. 359, 376–77 (1992)) (noting that trade secret protection is preferable to patent protection for processes because it is difficult to detect infringement of a process).

116. *See* SCHERER, *supra* note 112, at 444–46 (noting that natural market imperfections such as imitation lags, first-mover advantage, and nonpatent barriers reduce the economic need for patent protection).

117. *See* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 379 (2003) (stating that intellectual property rights should be granted only when "making intellectual property rights excludable creates value").

118. Peter S. Menell & Suzanne Scotchmer, *Intellectual Property Law*, *in* 2 HANDBOOK OF LAW AND ECONOMICS 1473 (A. Mitchell Polinsky & Steven Shavell eds., 2007).

encourage innovation, it merely rewards it.[119] To the extent that the patent system is unnecessary to promote innovation, analysis of the benefits and drawbacks of permitting patenting must be examined in order to determine the most socially beneficial means of encouraging innovation.

B. TRADITIONAL JUSTIFICATIONS FOR DISCOURAGING THE USE OF SECRECY

1. Disclosure

Courts have consistently cited the second incentive of patent law's reward theory—the disclosure incentive—as the principal benefit that the public receives from the patent system.[120] Legal theorists have similarly emphasized the harm resulting from reduced disclosure when inventions are maintained as secrets.[121] Indeed, disclosure and its accompanying benefits are fundamental to traditional notions of the patent quid pro quo. Katherine Strandburg has noted that when secret inventions are lured into the patent system the only benefit to the public is the resulting disclosure.[122] Furthermore, fears regarding the increased use of trade secrecy have centered around the harm to the public of decreased disclosure.[123] In sum, it is presumed that patents are better at promoting innovation than secrets due to patent law's disclosure doctrine.

However, there are three reasons to suspect that the effectiveness of patent disclosure in encouraging innovation is somewhat limited. First, patents generally do a poor job of promoting innovation through teaching. Patents do not perform much of a teaching function because of some paradoxical elements of patent law that discourage would-be innovators from

119. Strandburg, *supra* note 53, at 111 (finding that the reward theory cannot justify granting patents for inventions that can profitably be maintained in secret).

120. *See, e.g.*, Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 481, 489 (1974); Pfaff v. Wells Elecs., Inc., 525 U.S. 55, 63 (1998) (stating that the patent system should be thought of as "a carefully crafted bargain that encourages both the creation and the public disclosure of new and useful advances in technology, in return for an exclusive monopoly for a limited period of time"). For a more thorough discussion of the courts' treatment of the disclosure requirement, see Roin, *supra* note 58, at 2011–13.

121. *See* F. SCOTT KIEFF ET AL., PRINCIPLES OF PATENT LAW 68 (4th ed. 2008) (describing the basis of patent law's incentive to disclose and stating that "secrecy would deprive the public of the new knowledge").

122. Strandburg, *supra* note 53, at 111.

123. *See* Brief of Respondents, Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470 (1974) (No. 73-187), 1973 WL 172412; Brief of Amicus Curiae National Patent Council, Inc. for Itself and Representing National Small Business Association, Inc., Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470 (1974) (No. 73-187), 1973 WL 173805; Doerfer, *supra* note 6, at 1441.

consulting patents for technical knowledge. Second, disclosure requirements are often insufficient to promote disclosure that would enable follow-on innovation. Patentees in some industries are able to obtain patent protection while retaining essential know-how. Thus, some inventors are able to patent an invention while still maintaining enough of the invention in secret to prevent valuable knowledge from being transferred to the public. Third, concerns about the loss of public disclosure that would result from increased use of trade secrecy may be overblown. Trade secret law permits a limited form of disclosure that may replicate many of the beneficial effects of patent disclosure.

a) The Ineffective Teaching Function of Patent Disclosure

The courts' view of the patent quid pro quo often focuses on the innovation benefits that result from disclosing secret inventions.[124] The portrayal of disclosure as an innovation-promoting mechanism is not without theoretical support. If disclosure disseminates information that encourages follow-on innovation, then the social cost of the patent grant may be less than the social benefit of the follow-on innovation.[125] While patents exclude others from producing products that are covered by the patent's claims, perhaps the disclosure of the invention to the public will teach others the novel method or product and encourage improvements or tangential innovations that would not have occurred without the lesson contained within the patent. As an added benefit, after patent expiration the disclosed invention enters the public domain, free for all to use. In contrast, entrance into the public domain does not occur automatically with trade secrecy.

But, there are reasons to doubt the extent of the patent system's ability to teach follow-on innovators. Much of the doubt in the literature concerning patent law's teaching ability stems from aspects of the patent system that discourage innovators from consulting patents. Timothy Holbrook pointed to four aspects of the patent system that undermine the disclosure function of patents: the limited experimental use exception, the risk of willful infringement faced by those that do examine prior patents, the eighteen-month delay for publication of patent applications, and the moribund reverse doctrine of equivalents.[126] Holbrook concluded that "[n]ot only are the disclosure obligations inconsistent with the theoretical justifications of patent

124. For more on potential reasons for courts' and scholars' differing explanations for the patent system, see Roin, *supra* note 58, at 2012 (speculating on the reasons for courts' preference for the disclosure rationale).

125. Holbrook, *supra* note 44, at 134 n.56.

126. *Id.* at 139–45.

law, but the current structure of the patent system undermines the ability of patents to actually perform this function."[127]

The Federal Circuit effectively eliminated any experimental use exception for practicing an invention by follow-on innovators.[128] Follow-on innovators can read a patent, but they cannot make or use the patent in order to study the invention's properties or the manner in which it functions.[129] Without such a right, the ability of a patent to teach anything is severely limited. Furthermore, innovators who examine patents risk triggering a duty to investigate infringement and liability for willful infringement with accompanying treble damages.[130] Lastly, patent applications are not published for at least eighteen months, which reduces their value as teaching aids to follow-on innovators.[131] Many industries have such rapid innovation rates that eighteen-month-old innovations are relegated to the history books. Other commentators concur in Holbrook's findings that the patent system's disclosure ability is overstated.[132]

Jeanne Fromer has argued that disclosure promotes the progress of science and the useful arts not only in economic terms—that is, by conferring information to society that can be used for future innovations—but also by democratizing the process of innovation.[133] Fromer argued that disclosure levels the scientific playing field by permitting all interested parties to operate with the same basic information. According to Fromer, numerous

127. *Id.* at 146.

128. *See* Madey v. Duke Univ., 307 F.3d 1351, 1362–63 (Fed. Cir. 2002); Holbrook, *supra* note 44, at 140.

129. *Madey*, 307 F.3d at 1362–63.

130. *See In re* Seagate Tech., 497 F.3d 1360, 1366–71 (Fed. Cir. 2007) (en banc); Mark A. Lemley & Ragesh K. Tangri, *Ending Patent Law's Willfulness Game*, 18 BERKELEY TECH. L.J. 1085, 1090 (2003).

131. Patent applicants filing only in the United States can opt out of the eighteen-month publication requirement. 35 U.S.C. § 122(b)(2)(B) (2006). A 2002 study suggests that eleven percent of U.S. applications are not published at eighteen months. NAT'L RESEARCH COUNCIL, A PATENT SYSTEM FOR THE 21ST CENTURY 64 (2004).

132. *See* Doerfer, *supra* note 6, at 1444 (observing, wryly, that "the method in which the [patent] statute is administered seems to be quite compatible with the nondisclosure aims of trade secret law"); Roin, *supra* note 58, at 2027–28; Strandburg, *supra* note 53, at 113–18. Doerfer also notes that patent law's claim to promote disclosure is undermined by the secret nature of patent applications. Doerfer, *supra* note 6, at 1445. Since Doerfer's article, most patent applications are now published after eighteen months. While this change clearly reflects a desire for more disclosure, a system that truly valued disclosure over all other considerations would publish applications at filing.

133. Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539, 551 (2009).

minds are more capable of effecting technological progress than centralized control.[134]

While justifying the theoretical basis for disclosure as a democratizing discovery, Fromer agrees with Holbrook that patents do not, in general, fulfill their teaching role.[135] She points to evidence that most inventors spend little to no time reading others' patents.[136] This lack of relevance for technologists may stem from fear of willful infringement,[137] the inability to comprehend the legal jargon of a patent document,[138] or the lack of meaningful information that patents convey.[139] Patentees themselves tend not to consult others' patents,[140] and they learn of the patents that are eventually cited in their own applications only after invention.[141] Furthermore, inventors rank patents last among sources of inspiration for their inventions.[142]

Part of the explanation for innovators' apparent disinterest in patent documents is that patents are a poor medium for communicating technical information.[143] Although patent specifications are meant to be written for those skilled in the art, most scientists and engineers find patents to be repetitive and often incomprehensible.[144] Rules of claim construction

134. *Id.*

135. *Id.* at 560–62 ("[T]he evidence tends to show that potential inventors are not turning to patent disclosures to inspire their research.").

136. *Id.* at 560.

137. Lemley, *supra* note 37, at 1510 n.63.

138. Robert Barr, Speech at the Federal Trade Commission and Department of Justice Antitrust Division Roundtable on Competition, Economic, and Business Perspectives on Substantive Patent Law Issues: Non-Obviousness and Other Patentability Criteria 79–80 (Oct. 30, 2002), *available at* http://www.ftc.gov/opp/intellect/021030trans.pdf.

139. Doerfer, *supra* note 6, at 1444 (noting the Patent Office's "reluctance to require detailed specifications").

140. *See* Adam B. Jaffe, *The Meaning of Patent Citations*, *in* ADAM B. JAFFE & MANUEL TRAJTENBERG, PATENTS, CITATIONS, AND INNOVATIONS: A WINDOW ON THE KNOWLEDGE ECONOMY 389–90 & fig.5 (2002) (finding that less than 20% of patentees "learn about" the patents eventually cited in their applications before working on their inventions).

141. *Id.*

142. Behind "awareness of a commercial or technological opportunity, word of mouth, personal interactions, viewing a presentation or demonstration, joint work with others, and technical literature." Fromer, *supra* note 133, at 562; *accord* James Bessen, *Patents and the Diffusion of Technical Information*, 86 ECON. LETTERS 121, 122 (2005); Asish Arora, Marco Ceccagnoli & Wesley M. Cohen, *R&D and the Patent Premium* 17 (Nat'l Bureau of Econ. Research, Working Paper No. 9431, 2003), *available at* http://www.nber.org/papers/w9431 (finding no measurable impact on information flows resulting from patent disclosure).

143. *See* Sean B. Seymore, *The Teaching Function of Patents*, 85 NOTRE DAME L. REV. 621, 633–41 (2010) (describing how the widespread use of "patentese" reduces the teaching function of patents); *see also* Roin, *supra* note 58, at 2025.

144. Fromer, *supra* note 133, at 560–62; Roin, *supra* note 58, at 2025.

encourage patent attorneys to draft their specifications broadly so as not to have the narrowness of their disclosure read into their claims.[145]

Patentees can avoid fully disclosing their inventions via a number of methods. First, inventors can delay publication of their patent application until their patent issues. Patent applications are usually published eighteen months after they are filed, but if a patentee agrees to file for patent protection only in the United States, publication is delayed until issuance.[146] Patent applications typically take much longer than eighteen months to issue as patents; thus innovators in industries with product cycles that are shorter than the period of patent pendency can delay disclosure until after their innovation is obsolete.[147]

Second, patentees in certain industries can disclose enough information to obtain a patent but less than enough to reveal how that innovation is practiced. This is especially common in the software, computer hardware, and business innovation industries.[148] This practice permits innovators to patent the central innovation of their invention yet retain certain trade secret know-how or show-how that is required to effectively practice the invention.[149] Although the Federal Circuit requires patent applications to "provide a disclosure sufficient to enable one skilled in the art to carry out the invention,"[150] oftentimes a follow-on innovator must enter into a license agreement with a patentee to obtain the knowledge withheld from the patent application. Without that information, it is often difficult and costly to practice (and thus improve upon) an invention.[151] Thus, any claim that disclosure, by itself, provides a benefit to the public greater than the cost of a patent grant is weakened.[152] If follow-on innovators must obtain information directly from patentees in order to practice a patented invention, then that

145. Roin, *supra* note 58, at 2026 (citing claim drafting advice encouraging practitioners to describe as many variations as possible).

146. 35 U.S.C. § 122(b)(2)(B) (2006).

147. *See* Roin, *supra* note 58, at 2024 (finding that innovators in industries with short product cycles are more likely to withhold disclosure until issuance).

148. *Id.* at 2024 n.102 (citing FED. TRADE COMM'N, TO PROMOTE INNOVATION: THE PROPER BALANCE OF COMPETITION AND PATENT LAW AND POLICY ch. 3, at 33 (2003), *available at* http://www.ftc.gov/os/2003/10/innovationrpt.pdf).

149. *See* Jorda, *supra* note 21, at 31; Gregory J. Maier, *Software Protection—Integrating Patent, Copyright and Trade Secret Law*, 69 J. PAT. & TRADEMARK OFF. SOC'Y 151, 163–65 (1987) (noting that software developers can obtain patent rights while not disclosing source code).

150. Amgen, Inc. v. Chugai Pharm. Co., 927 F.2d 1200, 1213 (Fed. Cir. 1991).

151. *See* Roin, *supra* note 58, at 2025 (noting that "[m]any patented inventions cannot be recreated" from the information contained in the patent; without that information licensing or reverse engineering is required to practice an invention).

152. *Id.* at 2025 (stating that the practice of not disclosing key know-how in a patent "calls into question the extent to which patent disclosures can produce R&D spillovers").

patent disclosure has provided, at most, a means of locating the information necessary to perform further research.

All of this is not to suggest that patent disclosure is without value. Even if the teaching function of patents is limited, there are occasions in which the disclosure of an invention can lead to innovation that is not measurable by the impact of the patent document alone. For instance, the existence of patent protection may permit the scientific publication of ground-breaking research that otherwise would have remained concealed.[153] That is, even if patents do not adequately disclose information, they may enable scientists to disclose research via other means that better serve follow-on innovators.

But while such instances of valuable disclosure undoubtedly exist, they do not, standing alone, justify the discouragement of secrecy. There are other types of secret information, outside of patentable subject matter, that would promote innovation if disclosed, such as business survey information, customer data, laboratory data, etc.[154] Society refuses to pay for this information precisely because there is no reason to do so—it will be produced privately to the extent it is valuable. The public does not demand a complete democratization of competition. Indeed, competition relies on companies retaining certain tangible or intangible advantages over their competitors. Typically policy makers rely upon market mechanisms and scientific norms to determine what information is valuable enough for private individuals to invest in its creation.

b) Trade Secret Disclosure

Trade secrecy does not enjoy a system of mandated public disclosure, as patent law does. Indeed, public disclosure of a secret destroys the legal protection of trade secret law. However, commentators have noted that trade secret law encourages disclosure, although of a more targeted nature than patent disclosure.[155] The Court in *Kewanee* relied on trade secrecy's limited

153. *See id.* at 2027 ("Even given the current structural limitations, however, the patent system still serves a limited disclosure function by allowing inventors to discuss and publicize their research freely.").

154. *See, e.g.*, Henry J. Silberberg & Eric G. Lardiere, *Eroding Protection of Customer Lists and Customer Information Under the Uniform Trade Secrets Act*, 42 BUS. LAW. 487, 487 (1987) (noting that unpatentable customer information is among many businesses' "most precious" trade secrets).

155. Lemley, *supra* note 6, at 314 ("[F]or certain types of inventions we may actually get more useful 'disclosure' at less cost from trade secret than from patent law."); Sandeen, *supra* note 71, at 344 (noting that while the qualitative and quantitative scope of disclosure is different with trade secrets and patents, "it is true that trade secret law helps to facilitate the sharing of secret information between those with a need to know").

promotion of disclosure in finding no preemption of trade secret laws.[156] By protecting against misappropriation, trade secrecy reduces the cost of protecting secrets and permits innovators to market their ideas, as long as owners engage in a minimum level of protection.[157] Thus, trade secrecy exhibits some elements of disclosure that patent law encourages.[158] Of course, patent holders can engage in the same targeted disclosure as trade secret holders, and with less risk of loss of protection. But it is likely that many of the benefits to future innovation that come from patent disclosure can also exist in the more limited world of trade secret disclosure.

Indeed, given the scope of trade secret protection, inventors who maintain inventions as trade secrets likely have more incentive to efficiently disclose their inventions to the proper individuals. As described more fully in Section III.C, *infra*, trade secrecy encourages competition because the exclusivity of trade secrecy can end at any time. Thus, unlike patents, which have a certain duration, trade secret owners likely feel time pressure to maximize the value of an innovation. Thus trade secrets may lead to earlier—albeit more limited—disclosure than patents.

The idea that targeted, limited, inventor-initiated disclosure is more beneficial than patent disclosure for promoting innovation is open to debate. However, given the limited value of broad-based patent disclosure, it is likely that the innovative benefits resulting from trade secret's targeted disclosure at least approach the innovative value of patent disclosure.

2. Coordination of Commercialization and Research

The prospect theory, a second major strain of patent theory, concerns itself with efficiently allocating scarce research dollars. Edward Kitch suggested that the patent system places the patent holder "in a position to coordinate the search for technological and market enhancement of the patent[]," thus "increas[ing] the efficiency with which investment in innovation can be managed."[159] Kitch viewed the patent holder as having the power to coordinate future investment in the prospect because "no one is likely to make significant investments in searching for ways to increase the

156. Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 485–86 (1974) (describing a disclosure function in trade secret law).

157. Additionally, trade secret law permits the disclosure of secrets "in connection . . . with information that is relevant to public health or safety, or to the commission of a crime or tort, or other matters of substantial concern." RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 40 illus. 1(c) (1993).

158. *See* Sandeen, *supra* note 71, at 344.

159. Edmund Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265, 276 (1977).

commercial value of a patent unless he has made previous arrangements with the owner of the patent" to share in the profits of that effort.[160]

A patent system focused on coordinating downstream research activity, like that envisioned by prospect theorists, would likely benefit from the disclosure of secret inventions: increased openness would lead to increased coordination. However, the prospect theory presupposes that patents lead to coordination of research and that markets for innovation coordination operate efficiently. Both of these assumptions are contentious. First, as demonstrated in Section III.B.1.a, *supra*, patent documents do not do a good job of facilitating research or commercialization coordination. Scientists rarely consult patent documents.[161] The sheer number of patents issued every year (over 200,000 in 2010) makes it impossible for an innovator to stay current with all of the issued patents in a particular field.[162] Owners of trade secrets are just as likely as patent holders to seek out those that can best commercialize and market an innovation, perhaps more so.[163] Patent protection and disclosure may have the perverse effect of placating the drive of patentees from seeking out commercialization partners; they can simply rely on the patent for protection against competition or independent invention.

Numerous scholars have cast doubt on the assumption that pioneering inventors will efficiently market their patented technology. Rebecca Eisenberg noted that the likelihood of efficient licensing is lowest when "subsequent researchers want to use prior inventions to make further progress in the same field in competition with the patent holder."[164] Robert Merges explained that bargaining breakdowns occur between holders of blocking patents as a result of mistaken assumptions and irrational choices.[165]

As a corollary to the prospect theory's emphasis on coordination of research and development of an invention, the theory also emphasizes the

160. *Id.*

161. Fromer, *supra* note 133, at 560–62. Fromer also notes that the Intellectual Property Owners Association Survey likely overestimates the number of technologists who consult patents prior to invention by lumping together the research stage of invention with latter stages of invention. *Id.* at 561 n.104.

162. *See* PTO, *U.S. Patent Statistics*, *supra* note 9.

163. *See generally* Sichelman, *supra* note 12 (arguing that the dominant theories of patent law do not incentivize commercialization of patented inventions).

164. Rebecca S. Eisenberg, *Patents and the Progress of Science: Exclusive Rights and Experimental Use*, 56 U. CHI. L. REV. 1017, 1072–73 (1989).

165. Robert P. Merges, *Intellectual Property Rights and Bargaining Breakdown: The Case of Blocking Patents*, 62 TENN. L. REV. 75, 89 (1994).

elimination, or at least reduction, of duplicative research.[166] Kitch noted that "a patent system enables firms to signal each other, thus reducing the amount of duplicative investment in innovation."[167] In Kitch's view, patents put "the patent owner in a position to coordinate the search for technological and market enhancements of the patent's value so that duplicative investments are not made and so that information is exchanged among the searchers."[168]

Secrecy, to Kitch, does not permit other researchers to determine the efficient level of search.[169] Indeed, Kitch viewed the fact that "technological information can be used without signaling that fact to another"[170] as a problem that the patent system intends to solve. Other firms are unlikely to know of the success of the original inventors and thus cannot redirect their research accordingly. As Kitch acknowledged, under a trade secrecy regime "the competitive firm might never learn of a new product until it is marketed."[171]

However, duplicative research can provide social benefits. First, multiple firms inventing in the same area can result in distinct and improved innovations.[172] Firms that are initially engaged in pursuing the same goal often end up inventing different means of achieving that goal. Society benefits from having varied innovative products. Furthermore, multiple inventive efforts can result in new uses for the same product. In the area of biochemistry, for instance, multiple firms investigating a similar chemical compound can develop different uses for the compound. This also leads to competition and lower prices for consumers.

Duplication of research and commercialization efforts is often indistinguishable from competition.[173] Competition entails duplication of

166. Many reward theorists take a skeptical view to limiting independent invention because competition is also reduced. *See* Lemley, *supra* note 6, at 336 n.103. Duplication of research and commercialization efforts is often indistinguishable from competition. *See* Vermont, *supra* note 30, at 495. Competition entails duplication of costs by competing firms (an inefficient cost), but it also results in increased production and lowered costs for consumers (a social benefit). *Id.* Rivalry among firms to develop and create new inventions is seen by many reward theorists as deserving of encouragement from the law. *See* Robert P. Merges & Richard R. Nelson, *On the Complex Economics of Patent Scope*, 90 COLUM. L. REV. 839, 908 (1988). While rivalry creates inefficient duplication of effort and resources among competitors, it also tends to generate rapid technological progress. *Id.*

167. Kitch, *supra* note 159, at 278.

168. *Id.* at 276.

169. *Id.* at 278.

170. *Id.* at 276.

171. *Id.* at 278.

172. *See* Lemley, *supra* note 6, at 336 n.103.

173. *See* Vermont, *supra* note 30, at 495.

costs by competing firms (an inefficient cost), but also results in increased production and lowered costs for consumers (a social benefit).[174] Rivalry among firms to create and develop new inventions is seen by many reward theorists as deserving of encouragement from the law.[175] While rivalry creates duplication of effort and resources among competitors, it also tends to generate rapid technological progress.[176]

A final potential benefit of invention races is that duplicative research may actually encourage disclosure. John Duffy has a different take on secrecy's effect on duplicative research than Kitch. Duffy asserts that secrecy is much more difficult to maintain in the initial stages of research than Kitch suggests.[177] He claims that in the initial stages of research, firms may be incentivized to communicate their results with competitors rather than maintain them in secret.[178] Thus, to the extent that research secrecy is difficult to maintain, duplicative research and secrecy can, in some instances, lead to increased disclosure among competitors.

C. THE OVERLOOKED BENEFITS OF SECRECY

1. Increased Competition

The innovative benefits of trade secrecy have often been overlooked by courts and commentators.[179] Perhaps the primary benefit of secrecy is that of increased competition for innovative ideas. Robert Merges and Richard Nelson stated that "multiple and competitive sources of invention are socially preferable to a structure where there is only one or a few sources. Public policy, including patent law, ought to encourage inventive rivalry, and not hinder it."[180] Patents grant inventors control over the positive externalities associated with their invention, as well as control over future improvements and new uses of their invention. Granting such broad rights

174. *Id.*

175. *See* Merges & Nelson, *supra* note 166, at 908 (arguing that for industries involving cumulative technologies, public policy ought to encourage a "rivalrous structure" rather than a "race to invent" structure in order to "generate rapid technological progress").

176. *Id.*

177. John F. Duffy, *Rethinking the Prospect Theory of Patents*, 71 U. CHI. L. REV. 439, 497–98 (2004).

178. *Id.*

179. *See* Vincent Chiappetta, *Myth, Chameleon, or Intellectual Property Olympian? A Normative Framework Supporting Trade Secret Law*, 8 GEO. MASON L. REV. 69, 88–90 (1999) (arguing that trade secret law does not encourage innovation); Risch, *supra* note 15, at 26–27 (arguing that "creating incentives to innovate is a very minor justification for trade secret law").

180. *See* Merges & Nelson, *supra* note 166, at 908.

to future innovation reduces competition and, in some cases, may reduce innovation generally.[181]

The monopoly control of a patent may reduce the incentive to improve upon the patented technology. Indeed, while patents in some cases induce invention, they may retard the commercialization and improvement of that invention.[182] Trade secret holders, on the other hand, do not control the future excludability of their innovation. If someone independently replicates a secret invention, the value of the invention to the original inventor plummets since the exclusive use of the invention is now gone. Similarly, reverse engineering of an invention destroys trade secret excludability.

There is substantial debate in the literature about whether monopoly power or competition provides stronger incentives to improve upon an invention.[183] However, it appears that many industries rely on competition to spur innovation.[184] It is possible that the security offered by a patent may deaden the incentive to improve the invention, whereas the incentive to improve upon a trade secret is enhanced by the indeterminate length of exclusivity.[185] The fear of independent invention and reverse engineering may motivate trade secret holders to commercialize and improve their invention before someone else comes up with the same idea.

John Duffy argues that patents better promote improvements than trade secrets. Duffy proposed that the patent system functions as a type of Demsetzian auction, in which the winner is the innovator who promises to let the patent expire earliest.[186] Duffy's version of the prospect theory accentuates and directs competition by encouraging early discovery. According to Duffy, the patent system does not discourage competition; it encourages competition at an extremely early stage in the innovative process. Duffy views the ability to obtain blocking patents as a check on a patentee's

181. *See, e.g.*, PIGOU, *supra* note 107.

182. *See generally* Sichelman, *supra* note 12 (arguing that the dominant theories of patent law do not incentivize commercialization of patented inventions).

183. For differing viewpoints, compare JOSEPH A. SCHUMPETER, CAPITALISM, SOCIALISM, AND DEMOCRACY 100–03 (Harpers & Row 3d ed. 1962) (1942) (arguing in favor of monopolies) with Merges & Nelson, *supra* note 166 (arguing for competition).

184. *See, e.g.*, Mark A. Lemley & Lawrence Lessig, *The End of End-to-End: Preserving the Architecture of the Internet in the Broadband Era*, 48 UCLA L. REV. 925, 960–62 (2001) (arguing the internet industry relies upon competition); Howard Shelanski, *Competition and Deployment of New Technology in U.S. Telecommunications*, 2000 U. CHI. LEGAL F. 85 (arguing that the telecommunications industry relies upon competition to spur innovation).

185. *See* Lemley, *Free Riding*, *supra* note 13, at 1060 (arguing that patents "may simply give less incentive to improve on first-generation technology than competition for the rights to improvements").

186. Duffy, *supra* note 177, at 445.

monopoly power. He argues that "[c]ompetition to obtain, and to maintain, a monopoly position can be harnessed to constrain the monopolist and to increase social welfare."[187]

However, Duffy's view that blocking patents act to encourage competition relies upon secrecy. In Duffy's view, the real competition occurs *before* a patent is granted, in the period when two competitors are competing for patent rights. In cases in which secrecy is not a viable appropriability mechanism, Duffy is undoubtedly correct that patents encourage competition and innovation. But in the case in which secrecy is available, it is not clear that the patent encourages more competition than secrecy offers organically: it likely does in some cases and does not in others. Those cases in which it does not are likely to be the cases in which secrecy provides more private value than patenting.

Duffy's second argument—that the existence of blocking patents encourages patentees to improve and develop their inventions[188]—rests on certain assumptions. First, that licensing markets for blocking patents are strong and functioning. As demonstrated in Section III.B.2, *supra*, such markets do not appear to be robust.

The second assumption underlying the argument that blocking patents encourage commercialization is that follow-on innovators are not discouraged by the existence of a foundational patent to such a degree as to take their research dollars elsewhere. Knowledge of an existing foundational patent has been shown to discourage follow-on researchers.[189] Even if unknown at the time research began, such a foundational patent reduces the profits available to a researcher since she must share her profits with the original patentee whose work was unknown (and therefore unhelpful) to the follow-on development. Mark Lemley also noted that the inherent uncertainty of patent boundaries may chill further improvement by leaving

187. *Id.* at 490.

188. *Id.* at 489–90.

189. *See, e.g.*, Mark F. Grady & Jay I. Alexander, *Patent Law and Rent Dissipation*, 78 VA. L. REV. 305, 316–21 (1992) (using rent dissipation theory to explain instances in which courts grant broad rights to discourage follow-on innovations); Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCIENCE 698 (1998) (describing potential downstream product development as an effect of widespread patenting in the biomedical sector); Lemley, *supra* note 55, at 997–98 (noting that "efficient creation of new works requires access to and use of old works"); Merges & Nelson, *supra* note 166, at 843–44 (surveying historical examples in various industries to assess the effect of patent scope on follow-on innovation).

innovators in a patented space unclear whether they are running afoul of the law.[190]

Trade secrets, on the other hand, do not diminish the incentive for competitors to attempt to innovate. If the competitor is successful, she will enjoy the rights to practice the invention with no fear of an infringement suit or forced license arrangement. Trade secrecy thus leverages competition to promote commercialization and improved innovation.

2. Reduced Administrative Burden

The second area in which secrecy can prove more beneficial to society than patenting is in the reduced costs that secrecy imposes on administrative agencies. The PTO is now inundated with over 500,000 patent applications each year.[191] This is in addition to the current backlog of over 700,000 patent applications waiting to be examined.[192] The enormous number of applications means that each application is examined for only eighteen hours on average.[193] The inability to properly examine patent applications results in the grant of numerous invalid patents each year. Critics have not missed the opportunity to mock the PTO for the dubious patents that it issues each year.[194]

Secrets have no administrative regime, and reliance on trade secrecy to protect an invention removes the PTO from the equation. Furthermore, patent applications are not cheap: they cost on average $10,000 to $30,000.[195] With secrets, the money saved by not filing a patent application can be invested towards future innovation. Thus, patent applications exact a sort of innovation tax that trade secrecy avoids.

Perhaps even more staggering than the administrative cost of the patent system is the cost of enforcing patent rights. Patent litigation costs upwards of $15 billion per year to patentees and accused infringers.[196] An average

190. Lemley, *Free Riding*, *supra* note 13, at 1061; Mark A. Lemley & Dan L. Burk, *Quantum Patent Mechanics*, 9 LEWIS & CLARK L. REV. 29, 52–56 (2005).

191. *See* PTO, *U.S. Patent Statistics*, *supra* note 9.

192. U.S. Patent & Trademark Office, U.S. Dep't of Commerce, *Patent Inventory Statistics—FY09* (2010), http://www.uspto.gov/patents/stats/appbacklog.jsp (reporting a backlog of 718,835 patent applications).

193. Lemley, *supra* note 37, at 1496 n.3.

194. Simson Garfinkel, *Patently Absurd*, WIRED, July 1994, at 104; James Gleick, *Patently Absurd*, N.Y. TIMES MAG., Mar. 12, 2000, at 44; Lawrence Lessig, *The Problem with Patents*, THE INDUSTRY STANDARD, Apr. 23, 1999, *available at* http://www.lessig.org/content/standard/0,1902,4296,00.html.

195. Lemley, *supra* note 37, at 1498–99 (estimating 2001 costs).

196. BESSEN & MEURER, *supra* note 98, at 139, fig.6.5. Note that Bessen and Meurer's numbers are likely understated. *Id.* at 140–41.

patent case costs upwards of $5 million.[197] The public strain from patent enforcement is quite tangible as well. Judicial resources are strained with the complexity and time consumption of patent cases. The burden of patent appeals was so great that in 1982 Congress created a new circuit court, the Court of Appeals for the Federal Circuit, to handle those appeals.[198] Enforcing trade secrets is a much more affordable process. Trade secret cases average around one-third the cost of a similar-sized patent case.[199]

Lastly, trade secrets eliminate the rent-seeking behavior that patents often attract. The large value associated with the exclusive rights in certain technologies has resulted in a booming secondary market for the exclusivity rights of patents. While the merits of the secondary market being used purely for litigation are hotly debated,[200] there is ample evidence that abuses of the patent system are widespread. The so-called "troll" phenomenon, in which parties acquire patents simply to use them as weapons in extracting licensing fees from established companies has received ample attention in the literature.[201] Patent trolls have been analogized to a large innovation tax imposed privately on certain industries.[202] Trade secrets do not provide the same opportunities for rent-seeking. Because trade secrets do not permit exclusion of independent inventors, they do not provide any rent targets.

3. Incentive Value

The first two categories of secrecy's potential benefits are familiar to patent scholars. They are the benefits that accrue by avoiding the costs of patenting. In essence, reduced administrative costs and increased competition accrue simply by avoiding the drawbacks associated with patenting. However,

197. A trade secret litigation case in which damages are over $25 million costs around $1–2 million to litigate. WOLF, GREENFIELD & SACKS, P.C., Q&A: INTELLECTUAL PROPERTY LITIGATION 10 (2009), http://www.wolfgreenfield.com/files/litigation_copy 1.pdf. A patent case with similar damages costs $5 million on average. *Id.* at 5.

198. *See* Dreyfuss, *supra* note 10, at 6.

199. *See* Lemley, *supra* note 6, at 331 n.81 (citing AM. INTELLECTUAL PROP. LAW ASS'N, REPORT OF THE ECONOMIC SURVEY 2007, at 25–26 (2007)).

200. *See, e.g.*, Spencer Hosie, *Patent Trolls and the New Tort Reform: A Practitioner's Perspective*, 4 I/S: J.L. & POL'Y FOR INFO. SOC'Y 75 (2008) (arguing in support of secondary patent markets); James F. McDonough III, *The Myth of the Patent Troll: An Alternative View of the Function of Patent Dealers in an Idea Economy*, 56 EMORY L.J. 189, 190 (2006) (same).

201. *See, e.g.*, Mark A. Lemley & Carl Shapiro, *Patent Holdup and Royalty Stacking*, 85 TEX. L. REV. 1991, 1993 (2007) (describing holdup and royalty stacking burdens that trolls impose on manufacturers). *But see* John M. Golden, *"Patent Trolls" and Patent Remedies*, 85 TEX. L. REV. 2111, 2116 (2007) (questioning conclusions of Lemley and Shapiro).

202. *See* Gerard N. Magliocca, *Blackberries and Barnyards: Patent Trolls and the Perils of Innovation*, 82 NOTRE DAME L. REV. 1809, 1810 & n.7 (2007) (noting the common claim of trolls serving as a tax on innovation).

secrecy's benefits are more than simply the residual benefits of foregoing patent protection. In certain cases, secrecy can encourage innovation where patenting cannot.

To understand the incentive to invent function of secrecy, we first must establish that secret inventions and patented inventions often have distinct values to their owners.[203] For example, a new drug effective against migraines is likely to be much more valuable to its owner as a patent than as a secret. As a patent, the owner can exclude non-licensed manufacturers from reproducing the drug and therefore charge higher prices. As a trade secret, the drug would be subject to reverse engineering, which would likely reveal the drug's formulation. Public knowledge of this formulation would allow others to copy and sell the drug at a lower price.

Conversely, an innovative process of manufacturing that same drug, involving heating certain compounds to specific temperatures before combination, may be more valuable as a secret than as a patent. As a trade secret, the owner will not have to pay for a patent application, nor will she have to monitor competitors to ensure noninfringement. If she has manufacturing capabilities, she can keep the secret confidential from all but the employees that work at her factory. As a patent, on the other hand, she is forced to disclose the invention to all (including her competitors) and may not be able to detect infringement if one of those competitors infringes her invented process.

Seen in this light, it is quite apparent that secrecy can encourage innovation. That is, the private value of a trade secret, S, can be greater than the private value of a patent, P. When $S > P$, economic theory predicts that secrecy will provide greater incentives to invent than patenting.[204] Of course, this assumes that inventors can know ex ante the value of their future invention. While an inventor can obviously not know such information with exactitude, it seems likely that the relative values of the two protection regimes is possible at the point of deciding between patent or trade secret.[205] Indeed, economic theories of patenting depend upon such an assumption.[206]

Empirical evidence suggests that when secrecy is feasible, trade secrecy provides greater incentives to innovate than patenting. Petra Moser, in

203. *See* Denicolò & Franzoni, *supra* note 102, at 519 (noting that inventors select the protection that affords them the greatest scope of protection).

204. *See, e.g.*, Vermont, *supra* note 30, at 489 (explaining the relationship between private incentives to innovate and the social costs of patenting).

205. *See id.* at 478 ("An inventor will not pursue an invention unless her expected revenue exceeds her expected costs of invention.").

206. Lemley, *Free Riding*, *supra* note 13, at 1054.

examining data from the nineteenth- and twentieth-century world fairs, concludes that only a fraction of innovations (about twelve to fifteen percent) were patented.[207] She found that patent rates vary by industry and that variation is predicated upon the ability to protect one's innovation through secrecy.[208] For innovations in industries that are able to maintain secrecy, patent rates are lower than the average (around five percent). In industries in which secrecy is less available, the rates are much higher (around fifty percent).[209]

Intriguingly, Moser discovered that as reverse engineering methods in an industry improved, patenting rates increased. Patenting rates in the chemical industry, which had been as low as five percent between 1851 and 1876, increased to nearly twenty percent between 1893 and 1915 as reverse engineering became more cost-effective.[210] At the same time, other industries in which secrecy was ineffective (such as machine manufacturing) maintained consistently high rates of patenting.[211] Moser's findings indicate that inventors can determine, with some efficiency, the protection regime that will maximize the private value of an invention.

The fact that secrecy can provide stronger incentives to invent than patenting in certain cases is based on the tension within the reward theory between disclosure and incentives to invent.[212] The tension between disclosure and investment incentives is greatest when the disclosure itself harms the private value of the invention. As in the case of the secret pharmaceutical process, disclosure reduces an inventor's ability to exclude because detection ability is inversely correlated to the ability to maintain the invention in secret.[213] The ability to rely on secrecy permits inventors whose inventions are more valuable when undisclosed to maximize the value of their invention.

In addition to secrecy's incentive function, secrecy can help balance distortions in innovative investment that patent rights may encourage. Scholars have long been concerned that the rents available from patent

207. Petra Moser, *Why Don't Inventors Patent?* 3 (Nat'l Bureau of Econ. Research, Working Paper No. 13294, 2007), *available at* http://www.nber.org/papers/w13294.

208. *Id.* at 4.

209. *Id.* at 37 tbl.3.

210. *Id.* at 3.

211. *Id.*

212. Holbrook, *supra* note 44, at 146; Strandburg, *supra* note 53, at 105.

213. OFFICE OF INT'L AFFAIRS, NAT'L RESEARCH COUNCIL, GLOBAL DIMENSIONS OF INTELLECTUAL PROPERTY RIGHTS IN SCIENCE AND TECHNOLOGY 10–12 (Mitchell B. Wallerstein et al. eds., 1993) (noting that infringement detection issues are particularly acute with inventions maintained in secret).

protection would inefficiently redistribute investment effort and dollars.[214] If the law could force all secret inventions to be disclosed, the investment in those technologies that are non-self-revealing would be reduced. This result is due to patent disclosure reducing the private value of non-self-revealing inventions: just as a system of pure trade secrecy would create special incentives for secret inventions,[215] a system of full disclosure creates disincentives to invest in such inventions. In this way, the law's disclosure requirement may induce firms to invest in technologies that are more valuable when disclosed (patented inventions) than those that are more valuable when kept as proprietary information (secret inventions). This has negative consequences for society as non-self-revealing technologies with social benefits (i.e., lower costs for goods, reduced environmental impact, etc.) will be under-developed.

IV. TOWARDS A FRAMEWORK FOR SECRECY POLICY

The patent system provides a means for individuals and firms to invest ex ante in innovative activity, while knowing that a means of recouping that investment is potentially available ex post in the form of a patent. Such an incentive is not required to encourage investment for all inventions, however. For some inventions, first-mover advantage, complementary assets, secrecy, or some other market imperfection serves as an alternate means of recouping initial investments. The preceding Part described the societal advantages and disadvantages involved with increased reliance on secrecy. This Part will begin to construct a framework from which policy makers can analyze when to prefer one type of protection scheme, and when intervention is required to encourage inventors to make that choice.

A. CONSTRUCTING THE FRAMEWORK

1. Private Valuation: Inventor Choice of Protection Regime

The same invention can have vastly different private values depending upon the mode of protection used to protect that invention. Inventors that

214. *See, e.g.*, JAMES W. HENDERSON, HEALTH ECONOMICS & RESEARCH POLICY 288 (4th ed. 2009) (including the distortion of research incentives among the patent system's potential drawbacks). Of course the patent system can also direct research in socially beneficial directions. *See, e.g.*, PIGOU, *supra* note 107, at 185 (arguing that patents redirect inventive activity into areas of general usefulness). Jonathan Barnett argues that repeat market players may overcome some of these inefficiencies by efficiently balancing the strength of intellectual property rights. Jonathan M. Barnett, *Property as Process: How Innovation Markets Select Innovation Regimes*, 119 YALE L.J. 384, 432–33 (2009).

215. Kitch, *supra* note 159, at 279.

choose to practice an invention in secret have determined that the invention is worth more as a trade secret than as a patent. That is, taking into account the private advantages (including protection from independent invention, signaling effects, etc.) and disadvantages (including disclosure, limited duration, cost, etc.) of a patent, the inventor has decided that she can capture more of the value of her invention through secrecy. Various factors influence this decision. Among the most prominent of these are the potential market life of the invention, the feasibility of secrecy, and the likely use of the invention.[216]

First, inventors must take into account the likely lifespan of the invention. Patents are limited to twenty years, whereas trade secrets are valid as long as the secret is kept. The commercial lifespan consideration typically favors choosing secrecy over patenting for both extremely long commercial life-cycles and extremely short ones. For extremely long product cycles measured in multiple decades, patents are less attractive because they expire after twenty years. If an innovation promises to be valuable for a period of time longer than twenty years, it may behoove the inventor to keep the innovation as a secret.[217]

For extremely short life cycles, patent protection is sometimes impractical because of the delay involved in obtaining a patent. On average a patent takes around five years to issue.[218] After that delay, an invention's value may be extinguished. In rapidly moving industries, patenting and its accompanying expenses are often not attractive to inventors; the invention's product life-cycle will end before the patent issues. Similarly, the financial returns available from being the first product to market often dwarf any increased returns that a patent could provide.

---

216. For a more complete view of the decision between patenting and secrecy, see Holly Amjad, *Patent vs. Trade Secret: Look at Costs, Industry, Returns*, BUS. J. KAN. CITY, Feb. 3, 2002, *available at* http://www.bizjournals.com/kansascity/stories/2002/02/04/smallb3.html; Ozzie A. Farres & Stephen T. Schreiner, *Patent or Trade Secret: Which Is Better?*, 124 BANKING L.J. 274 (2007); Daniel C. Munson, *The Patent-Trade Secret Decision: An Industrial Perspective*, 78 J. PAT. & TRADEMARK OFF. SOC'Y 689 (1996); Sharon K. Sandeen, *Checklist for Choosing Between Patent and Trade Secret Protection*, 479 PLI/PAT 725 (1997).

217. Eisenberg, *supra* note 164, at 1029.

218. The exact average pendency is impossible to know, but Hal Wegner estimates a sixty month pendency. Hal Wegner, Overall Patent Pendency (March 5, 2010) (unpublished manuscript) (on file with author). The official government estimate is a more respectable, albeit less realistic twenty-four months. *See* U.S. PATENT & TRADEMARK OFFICE, U.S. DEP'T OF COMMERCE, PERFORMANCE AND ACCOUNTABILITY REPORT: FISCAL YEAR 2000, at 38 (2000), *available at* http://www.uspto.gov/about/stratplan/ar/2000/00goals.pdf (latest available official statistics state average pendency for patent application is 25.0 months). *See generally* 35 U.S.C. § 154(b)(1)(B) (2006) (statutory guarantee that patents issue no more than three years after filing).

Second, inventors must analyze the feasibility of maintaining an invention in secret. Secrecy is possible only with a limited set of inventions. Inventions that are easily discerned via reverse engineering, or that are likely to be developed independently, are ideal candidates for patent protection.[219] Consumer products are very difficult to maintain in secret. Widespread distribution limits an inventor's ability to meaningfully control the downstream use of a product and prevent competitors from reverse engineering it.[220] Indeed, an entire "teardown" industry has sprung up that permits cost-effective reverse engineering of even the most sophisticated consumer products, such as iPhones.[221] Products that are generally available to competitors are unlikely to remain secret for long because once an invention is sold or marketed, it is easily replicated.[222]

The concealment of a trade secret can be threatened from within as well as from reverse engineers without. Unscrupulous employees, former business partners, and hackers all pose risks for inventors attempting to maintain secrets.[223] Because trade secrets must be closely guarded, the cost of protecting them can be very high. Innovators must weigh the cost of maintaining the secret when they decide whether or not to patent.[224] Physical security measures, employee agreements, and cyber-security can all be costly means of protecting an invention; a cost that may be greater than the cost of obtaining legal protection through the patent system.

Patents, on the other hand, require disclosure. Not only does a patent describe the manner of practicing an invention, it alerts competitors to the invention's existence.[225] Competitors alerted to a patented method are free to "design around" the invention and bring competing products or processes to

219. *See* Moser, *supra* note 207, at 1.

220. Andrew Beckerman-Rodau, *The Choice Between Patent Protection and Trade Secret Protection: A Legal and Business Decision*, 84 J. PAT. & TRADEMARK OFF. SOC'Y 371, 386 (2002).

221. *The Lowdown on Teardowns*, THE ECONOMIST, Jan. 21, 2010, at 62, 63.

222. The rise of a sophisticated reverse-engineering regime has led some commentators to claim that it is now virtually impossible to maintain inventions in secret. *See* Mazzone & Moore, *supra* note 6, at 35; *see also* Holbrook, *supra* note 44, at 134 (stating that the set of non-self-disclosing inventions is "small"). This argument, however, tends to focus on inventions that are contained within commercially available consumer products. While it is true that certain inventions, particularly products, are often impossible to conceal once they are sold, other types of inventions, such as chemical and industrial processes, are not disclosed to the public when the final product is sold. Many processes are not revealed in the products that they create; the very existence of the process may be undetectable.

223. This is not a new phenomenon. *See* Victor M. Harding, *Trade Secrets and the Mobile Employee*, 22 BUS. LAW. 395 (1967).

224. Beckerman-Rodau, *supra* note 220, at 382.

225. *Id.* at 384.

the market.[226] In contrast, if the invention is maintained as a secret, competitors may not know of the existence of the invention, let alone the manner of practicing the invention for themselves.

However, disclosure of an invention can be desirable to certain inventors. A patented invention can signal to competitors that a particular area of technology has been cornered. This may discourage other companies from investing in the same invention because the patent prevents any future developments in that space.[227] Similarly, Clarisa Long theorized that patents are often used as a signal of innovative activity at a firm.[228] Under Long's theory, firms may desire the disclosure of a patent because it enables them to attract investment from investors who rely upon patents as a signal of innovative strength.

Disclosure is also desirable when an inventor wants to widely market, sell, or license her innovation. The patent system's ability to overcome Arrow's paradox—one will not pay for an invention that isn't disclosed—has long been heralded by commentators.[229] Disclosure of trade secrets, while permitted, is more costly and limited than with patents.[230] Secrecy is impractical when exploitation of the invention requires impersonal communication to a large number of firms.[231]

Inventions that are easily maintained in secret are also likely to be infringed in secret.[232] Thus, the ability to maintain an invention in secret has two important roles in determining the proper means of appropriation: it allows an inventor to enjoy a competitive advantage for a potentially limitless time period, and it reduces the value of a patent on that invention because the cost of detecting infringement is increased.

Lastly, in deciding upon invention protection, inventors must take into account the likely use of any new invention. As detailed above, inventions embodied in consumer products are often poor candidates for secrecy. If the business model for an innovation calls for widespread licensing, rather than in-house use, patent protection may make more sense than secrecy.[233] On the

226. For more on the ability to design around, see S. Glazier, *Inventing Around Your Competitors' Patents*, MANAGING INTELL. PROP., July/August 1995, at 10.

227. This idea has been explored thoroughly in the literature on the prospect theory of patent law. *See generally* Duffy, *supra* note 177, at 476; Kitch, *supra* note 159, at 267–71.

228. Clarisa Long, *Patent Signals*, 69 U. CHI. L. REV. 625, 627–28 (2002).

229. Arrow, *supra* note 104, at 614–16; Eisenberg, *supra* note 164, at 1029.

230. Lemley, *supra* note 6, at 314.

231. WARD S. BOWMAN JR., PATENT AND ANTITRUST LAW 13 (1973).

232. Oddi, *supra* note 115, at 285 n.126 (stating that because process patent infringement is difficult to detect, processes are ideal candidates for trade secret protection).

233. Beckerman-Rodau, *supra* note 220, at 403.

other hand, products and processes which are to be used in the internal workings of a company are much more likely to have value as secrets.

2. Comparing Private and Public Preference

While the factors involved in any individual inventor's decision of whether to patent are quite complex, in general we can say that inventors will choose secrecy when the expected return from secrecy exceeds the expected return from patenting. If we define *S* as the value of a trade secret and *P* as the value of a patented invention, we expect inventors to choose secrecy when $S > P$. *S* represents the value to the inventor of the invention as a secret, taking into account the risks and costs of secrecy as detailed above. In other words, *S* equals the increased profit one can expect from using the invention if kept secret forever reduced by some function accounting for the potential discovery of the secret. *P*, on the other hand, represents the increased profit to be expected over the twenty-year life of the patent (including any licensing royalties), minus enforcement costs and patent fees.

Let us introduce a third variable, *R*, which represents the cost of researching, developing, and commercializing the invention. *R* represents the revenue that must be generated by the invention to allow the inventor to recoup her upfront costs. Generally, when the expected return from either secrecy or patenting exceeds *R*, the invention will be produced; conversely, when *S* and *P* are both less than *R*, the invention will not be produced. That is, when $R > P$ *or* S, a potential inventor will not expend the necessary effort to produce the invention because she will not recoup her upfront research costs. Four scenarios in which theory predicts that inventive effort will be expended deserve our attention. I label those scenarios Public Goods, Reverse Public Goods, Valuable Secret, and Valuable Patent. The scenarios are analyzed in more detail in Section IV.B, *infra*.

B. FRAMEWORK SUMMARY

The chart below summarizes the secrecy framework described in this section:

Scenario	Private Valuation	Inventor Preference	Societal Preference
Public Goods	$P > R > S$	Patent	Patent
Reverse Public Goods	$S > R > P$	Trade Secret	Trade Secret
Valuable Secret	$S > P > R$	Trade Secret	Trade Secret
Valuable Patent	$P > S > R$	Patent	Trade Secret

The framework revolves around two primary assumptions. First, private inventors will choose the intellectual property protection that offers the

greatest private value. Thus, if at the moment of selection, trade secrecy appears to offer the greater private reward, an inventor will choose to maintain her invention in secret. Conversely, if patent protection appears to offer the greatest private reward, patenting will be selected as the appropriation mechanism.

Second, societal preference is premised upon an innovation regime that primarily encourages innovation and secondarily reduces social costs. Thus, society prefers the appropriation regime that incentivizes creation of innovative devices and methods. When innovation is incentivized under both patent and trade secrecy, society prefers to offer the protection regime that carries the lowest social burden. For reasons described herein, I conclude that trade secrecy carries lower social costs than patenting.

The framework suggests that policy makers ought to be more concerned with encouraging the use of secrecy, rather than discouraging it. Policy makers need not concern themselves with influencing decisions in the scenarios in which inventor preference coincides with public preference.[234] Contrary to contemporary understanding, secrecy does not need to be discouraged by patent law. In fact, the one case where the socially optimal selection differs from the expected inventor selection suggests the need for a policy that creates incentive to keep an invention as a trade secret.

Thus, secrecy policy should be motivated by two primary concerns. First, the use of trade secrecy need not be discouraged. The public goods scenario is the only situation in which secrecy is not the socially preferred method of protection. Because of the existence of the patent system, inventors will seek patents on such inventions without any intervention from policy makers. Second, in a small set of cases—when $P > S > R$—secrecy should be encouraged. The following section outlines ways in which the two policy objectives suggested by the secrecy framework might be employed.

1. The Public Goods Scenario

The reward theory's incentive to invent is generally concerned with the public goods market failure.[235] The public goods scenario occurs when the expected return from secrecy is less than the cost of development and the expected return from patenting exceeds development costs, or $P > R > S$. In this case, we would not expect innovation to occur without the patent

234. Trusting individuals in valuation decisions instead of courts is the primary justification scholars have offered for injunctive relief in patent cases. *See* ROBERT P. MERGES ET AL., INTELLECTUAL PROPERTY IN THE NEW TECHNOLOGICAL AGE 297–99 (3d ed. 2003).

235. *See e.g.*, Holbrook, *supra* note 44, at 132; Lemley, *Free Riding*, *supra* note 13, at 1053.

system. Secrecy alone is insufficient to induce investment because the inventor will not be able to recoup her initial investment. The promise of a patent, however, is sufficient to induce investment. This scenario is the only one of the four that involves the production problems associated with public goods.

The public goods scenario occurs quite frequently. This scenario likely describes the majority of patented product inventions, which tend to be difficult to conceal and therefore easily copied. For example, Chester Carlson's invention of the Xerox machine required large investments in the then novel field of imaging technology.[236] Had Carlson maintained his invention as a trade secret, it would have been possible for a competitor to reverse engineer the xerography process once the machines were sold publicly. A copyist could have offered a lower priced alternative since she would have avoided the research costs incurred by Carlson. Of course, he avoided the free riding problem by obtaining a number of patents covering his technology.[237] Patenting enabled him to exclude others from practicing his invention and thus charge supranormal prices in order to recoup investment costs.[238]

The public goods scenario is the primary economic justification for the existence of the patent system.[239] This scenario describes the classic economic win-win: inventors benefit by being able to recoup investment costs and the public benefits by receiving new technologies that are disclosed for public consumption.

In this scenario, inventor preference is aligned with societal preference. When secrecy does not provide sufficient means of recouping investment, rational inventors will choose to patent their inventions. Society prefers patenting in the public goods scenario because reliance upon secrecy results in reduced investment in and production of innovation. Thus, the mere existence of the patent system will encourage investment in and disclosure of novel innovations. There is no justification for discouraging secrecy in this case because rational inventors will independently make the socially optimal choice—patenting.

---

236. *See generally* DAVID OWEN, COPIES IN SECONDS: CHESTER CARLSON AND THE BIRTH OF THE XEROX MACHINE (2004) (describing the large investments made in the invention of the Xerox machine).

237. *Id.* at 141 (noting that Carlson received forty total patents on xerography).

238. *See id.* at 288–89 (describing the potential competitors "lining up to sue" over Carlson's patent misuse).

239. *See, e.g.*, Holbrook, *supra* note 44, at 132.

2. The Reverse Public Goods Scenario

Certain inventions, particularly process inventions, are significantly more valuable to inventors as secrets than the same invention would be if patented. At times, the protection offered by secrecy can provide a means of appropriating an invention and recouping the invention's investment costs while the same invention, if disclosed as a patent, would not provide an inventor with sufficient incentive to innovate. I name this scenario the "reverse public goods scenario." In the reverse public goods scenario, mandated disclosure (if possible) would lead to the underproduction of certain inventions, namely those inventions in which infringement detection would be difficult and therefore trade secrecy more valuable.[240] Infringement detection is difficult for some of the same reasons that secrecy is appealing: the marketed product or service does not reveal the underlying technology. The reverse public goods scenario is defined as $S > R > P$.

In the reverse public goods case, as in the classic economic case of public goods, free riders would drive down the cost of an invention, reducing the ability of the innovator to recoup costs.[241] However, in this case it is the disclosure of the invention, not reverse engineering, that provides the free riding opportunity.

Chemical manufacturing methods are an example of the reverse public goods scenario. These methods are often undetectable to a potential reverse engineer.[242] Thus, if the innovator of a new type of process chooses to patent, competitors could use the process in secret with little fear of detection and subsequent infringement suits. Because of the difficulty in detecting infringement, companies tend to maintain such processes as secrets rather than disclose them through the patent office. If these inventions were somehow forcibly disclosed, inventors would choose to invest less in those technologies where private value is undermined by disclosure, resulting in less innovation and ultimately less disclosure.

The reverse public goods scenario does not occur under current law because of the appropriability that secrecy provides. Thus, because inventors are not forced to patent (and disclose) their inventions, they are free to maintain inventions as trade secrets and rational actors will do precisely that. Society has a preference for these inventions remaining as secrets for precisely the same reasons that patenting is preferred in the public goods

240. PIGOU, *supra* note 107, at 185.

241. *See, e.g.*, Holbrook, *supra* note 44, at 132; Lemley, *Free Riding*, *supra* note 13, at 1053.

242. Lemley, *supra* note 6, at 339 (noting that chemical processes are not transparent to the world).

scenario: in a world of full disclosure, certain inventions would be suboptimally produced, thereby reducing overall innovation. Again, this scenario does not present a justification for discouraging secrecy. In fact secrecy is the socially optimal choice in this scenario. Fortunately, rational inventors will make that choice as well.

3. The Valuable Secret Scenario

A third scenario presents a more difficult case in determining whether to incentivize trade secrets or patents. There are cases in which both patenting and secrecy promise returns greater than investment costs. In some of those cases the expected returns from a trade secret exceed the expected returns from a patent, or $S > P > R$. Patent doctrines designed to discourage secrecy are primarily concerned with this scenario.[243] Here, the patent system provides enough of an expected return to encourage innovation, just not as large of a return as secrecy. Secrecy will be the preferred method of appropriation for rational inventors in this scenario. Traditional treatments of this scenario have tended to prefer patenting, as described in Section IV.A, *supra*.

The *Kewanee* case is an example of the valuable trade secret scenario. The valuable innovation involved in Harshaw Chemical's 17-inch radiation-detecting crystal was the process used in manufacturing and growing the crystal.[244] Harshaw likely would have been able to profit from its invention if patented because discovering a 17-inch crystal in radiation-detection products would likely constitute prima facie evidence of infringement, assuming no other methods of growing such crystals were known. However, for reasons which are unclear from the published opinion, Harshaw's leadership felt that maintaining the method of crystal growth as a trade secret would provide more private value for the company. This may be because the market for such crystals was relatively small, detection of the use of such crystals would be prohibitively expensive, or for some other reason.

Maximizing public value in this case is more complicated than in the previous two scenarios because the primary concern of the patent system—stimulating innovation—is not a concern: both the patent system and trade secrecy promise a return on innovative investment. In this Section, I suggest that secrecy is socially optimal in the valuable secrets scenario. Secrets exhibit numerous benefits over patents in this scenario. First, reliance on trade

243. *See* Eisenberg, *supra* note 164, at 1072 (stating that the quid pro quo only concerns this scenario).

244. Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 473 (1974). Harshaw spent over $1 million in developing its crystal. *Id.*

secrecy reduces the administrative costs of innovation. Second, secrecy encourages competition in both innovation and commercialization. Third, secrecy provides a natural market mechanism of inducing inventive activity. Fourth, the use of trade secrecy when available leads to increased incentives to invent in the future. Furthermore, deadweight losses from secrecy will attract competitors whereas patents will discourage competition. The unencumbered ability to innovate in innovative areas that are protected as trade secrets encourages competitors to reduce the deadweight losses in circumstances in which those losses are excessive.

Trade secrecy avoids some of the costs associated with patenting, namely administrative costs and the costs to future innovators of navigating patent entitlements.[245] Trade secrecy often involves some sort of disclosure.[246] However, the deadweight losses from a secret invention in this scenario are likely to be higher than those from patenting. Deadweight losses refer to the losses that result from certain consumers who would purchase the invention at the marginal price being priced out of the market due to the exclusionary power of the producer.[247] Because higher private value signals the ability to charge higher prices, we should expect secrecy to result in greater deadweight losses than patenting in this scenario.[248]

However, the societal advantages of trade secrets over patents may overcome the greater deadweight losses on the margins. For those cases in which deadweight losses are much greater for secrets than patents, the large relative private value of the secret invention is unlikely to be overcome by small policy tweaks. Encouraging such privately valuable inventions to be patented would likely require eliminating secrecy completely in such cases, an obvious impossibility. Additionally, it should be noted that because trade secrecy does not restrict competition, we should expect market forces to counteract somewhat any deadweight losses that arise from secrecy. Whereas patents create deadweight losses and restrict competitors from attempting to lower those losses, trade secrecy likely attracts competitors due to the lack of competitive restrictions.

Furthermore, there is little economic justification for granting patents in this scenario, let alone preferring them to trade secrets. Patent law is designed

245. *See* Vermont, *supra* note 30, at 490–92 (classifying the costs associated with granting patents as monopoly losses, rent dissipation, and miscellaneous costs).

246. *See* Sandeen, *supra* note 71, at 344.

247. *See, e.g.*, WILLIAM NORDHAUS, INVENTION, GROWTH AND WELFARE: A THEORETICAL TREATMENT OF TECHNOLOGICAL CHANGE (1969) (recognizing that the creation of intellectual property rights in innovation will lead to deadweight loss as a result of monopoly pricing).

248. *See* Denicolò & Franzoni, *supra* note 102, at 530–34.

to overcome a particular market failure—free riding on information.[249] The valuable secret scenario does not present such a market failure.

In addition to secrecy's incentive and competitive benefits, permitting inventors to select secrecy in this case has innovative benefits. Economists have developed economic models that demonstrate that overall investment is reduced when patent strength is increased for inventions in which secrecy is a viable option. Illoong Kwon showed that when the patent propensity (the ratio of innovations for which patent protection is sought) is less than one, "strengthening patent protection always *decreases* research investment."[250] This result holds for both models involving single innovations[251] as well as cumulative innovations.[252] Kwon's models support the theoretical model of this Article and also have an intuitive explanation. As patent protection becomes stronger, firms will increasingly prefer patents even when ex ante expected profits would be greater under a secrecy regime because they will have a "strong incentive to apply for patents in order to exclude the other firms from the product market."[253] Thus, widespread patenting in an industry incentivizes inventors to increasingly rely on patent protection, thereby reducing the value of those inventions that can be maintained in secret. The reduced value in turn reduces the ex ante incentive to invest in innovation.

4. The Valuable Patent Scenario

The final scenario describes situations in which both patenting and trade secrecy provide sufficient incentives to innovate, but a patented invention has more private value than a secret. In other words, when $P > S > R$. There are a number of real-life scenarios in which secrecy alone is sufficient to propel an innovator to create, but patenting provides a higher potential return on the initial investment.

Many patented inventions likely fall within this scenario. For instance, several methods of financial investing would be appropriable as either trade secrets or as patents.[254] For years such financial methods were largely

249. *See, e.g.*, Holbrook, *supra* note 44, at 132; Lemley, *Free Riding*, *supra* note 13, at 1053.

250. Illoong Kwon, Patent Portfolio Races and Secrecy 3 (2009) (unpublished manuscript), http://www.albany.edu/~ik325357/Research/Portfolio.pdf.

251. Illoong Kwon, Secrecy and the Fallacy of Patent Protection 2–3 (2010) (unpublished manuscript), http://www.albany.edu/~ik325357/Research/Fallacy.pdf.

252. *Id.*

253. Kwon, *supra* note 250, at 23.

254. *See, e.g.*, John R. Allison & Emerson H. Tiller, *The Business Method Patent Myth*, 18 BERKELEY TECH. L.J. 987, 1015 (2003) (noting that the change to business method patentability has led to worries that methods maintained as trade secrets would now be patented).

maintained as trade secrets; companies that developed superior methods were rewarded when their investments earned higher profits than those of their competitors.[255] After *State Street Bank & Trust Co. v. Signature Financial Group, Inc.* established the patentability of such methods, many financial institutions began patenting some of their inventive method innovations that would have been maintained as secrets prior to *State Street*.[256] In other words, some financial innovations promised more private value for their creators as patents than as secrets. Much of this added value likely came from a patent's ability to exclude competitors from practicing the invention. There is little reason, however, to think that these methods were insufficiently incentivized prior to *State Street*. Rather, once offered the choice, the institutions felt that patenting held greater promise for earning profit than did trade secrecy.[257]

However, the social value of financial institutions rushing to patent methods that a decade earlier would have been maintained in secret is likely negative: patenting in this scenario involves greater deadweight losses,[258] higher administrative costs,[259] and potentially decreased incentives to commercialize.[260] All of these costs come with no increase in innovation, because secrecy alone would have provided (and once did provide) sufficient incentives to create.

Secrecy, on the other hand, provides the same innovation at a lower cost to society. Furthermore, it enables competition in innovation to flourish, likely resulting in improved products and better incentives to commercialize quickly. Indeed, prior to *State Street*, there was no shortage of inventive financial methods. Competition among rivals ensured that new and innovative financial methods would continue to be developed. Companies protected their investment in such methods through a variety of methods such as marketing, first-mover advantage, and secrecy. The introduction of wide-spread patenting into the field has created uncertainty as to rights clearance as well as a new competitor, the so-called "patent troll."

---

255. Kwon, *supra* note 250, at 4.

256. *Id.*; Josh Lerner, *Where Does State Street Lead? A First Look at Finance Patents, 1971 to 2000*, 57 J. FIN. 901, 906–07 (2002) (noting the increasing acceleration of finance patents over time); *see also* State St. Bank & Trust Co. v. Signature Fin. Group, Inc., 149 F.3d 1368 (Fed. Cir. 1998).

257. *See* Lerner, *supra* note 256, at 906–07.

258. F.M. SCHERER & DAVID ROSS, INDUSTRIAL MARKET STRUCTURE AND ECONOMIC PERFORMANCE 449–50 (3d ed. 1990) (documenting excess pricing of patented products).

259. Many of the costs that Vermont refers to as "miscellaneous costs." Vermont, *supra* note 30, at 492.

260. *See* Sichelman, *supra* note 12.

Trade secrecy avoids many of the social costs that wide-spread patenting creates. However, rational individuals and companies will continue to choose to patent their inventions in cases where patenting provides more private value than secrecy. Unlike the first three scenarios, public value is maximized by encouraging inventors to choose an intellectual property regime that reduces private value. Thus, encouraging inventors to rely on trade secrecy in the valuable patents scenario should be the focus of policy makers.

Patent doctrine and rhetoric regarding secret inventions has been concerned with discouraging inventors from choosing secrecy. This focus, however, is misplaced. As shown above, when rational inventors prefer trade secrecy (when a trade secret provides more private value), that choice is socially optimal (it provides the largest social value).[261] Instead of discouraging secrecy, policy makers and courts should adopt policies that encourage secrecy in the limited set of circumstances when $P > S > R$.

5. Framework Caveats

The framework is subject to a number of caveats. First, the framework involves rough ex ante estimates of private invention value. These values are difficult to determine ex ante. They involve calculations of the odds of success as well as predictions of market demand—values which are very difficult to predict with precision. However, it is likely that many inventors can make at least some determination of relative future value.[262] Indeed, the reward theory generally presumes that inventors can determine the relative value of a potential invention and the cost of creation.[263] The framework constructed in this Article merely adds a third value, *S*, to the fundamental framework of the reward theory. Inventors who can determine, on some level, expected patented returns likely can make an estimation of expected trade secrecy returns as well.

Another potential drawback to the framework is that the economic rationale underlying the framework (and patent theory generally) assumes perfectly rational actors. While those assumptions have been questioned elsewhere, their support (or lack thereof) is beyond the scope of this

261. The social value of a patented invention, of course, is dependent on the social value derived from the disclosure of that invention. For purposes of this Article, it is assumed that the social value of disclosure, standing alone, is less than the social cost of a patent. If one places more value on the social benefit of patent disclosure, this conclusion may change.

262. Lemley, *Free Riding*, *supra* note 13, at 1064; *see also,* Vermont, *supra* note 30, at 499–500 (assuming inventors can value future invention value relative to research costs in both a world with and a world without an independent invention defense).

263. Vermont, *supra* note 30, at 499–500.

Article.[264] Such assumptions of future value and rational actors lie at the heart of patent theory.

Lastly, and perhaps most importantly from a policy perspective, the determinations depend upon one's valuation of patent disclosure. I have detailed the reasons for doubting the societal value of patent disclosure alone in Section III.B.1, *supra*. My view on the limited public value of patent disclosure is not unique.[265] However, reasonable minds can, and do, differ on this subject.[266] The value one places on patent disclosure may affect the societal value associated with the differing modes of protection. If one believes that patent disclosure, standing alone, has a large societal value, the public value discussion in this Article's valuable patent scenario may be altered. A view of strong patent disclosure value may suggest encouraging patents when patented inventions have a larger private value than a secret invention.

C. EMPLOYING THE FRAMEWORK

Moving towards a more theoretical framework for encouraging the use of secrecy requires certain changes to the doctrines of patent law. Before delving into the changes suggested by this Article's framework, it is important to note the rhetoric employed by courts. In intellectual property law, the choice of rhetoric employed to embody legal concepts has consequences.[267] Courts have consistently elevated the patent quid pro quo beyond merely an option for inventors; it has become a de facto social policy. Little effort has been made to theoretically support the discouragement of secrecy and, as detailed above, policy makers should in fact prefer precisely the opposite result in certain cases. Courts would be better served by couching their examination of the patent bargain in language suggesting patents as a means of encouraging innovation rather than eliminating secret inventions.

---

264. For more on the problem of assuming perfectly rational actors in law, see Robert C. Ellickson, *Bringing Culture and Human Frailty to Rational Actors: A Critique of Classical Law and Economics*, 65 CHI.-KENT L. REV. 23 (1989).

265. *See generally* Fromer, *supra* note 133 (suggesting ways to improve the disclosure system, while acknowledging the limited current value of disclosure); Holbrook, *supra* note 44 (noting the paradoxical elements of patent doctrine that discourage disclosure).

266. *See generally* Kitch, *supra* note 159 (arguing for a new theory of patent law based upon the signaling value of patent disclosure).

267. *See, e.g.*, Adam Mossoff, *Who Cares What Thomas Jefferson Thought About Patents? Reevaluating the Patent "Privilege" in Historical Context*, 92 CORNELL L. REV. 953 (2007) (investigating the role of courts referencing Thomas Jefferson's views of patent law).

1. Reversing the Doctrines Against Secrecy

Current patent doctrine attempts to influence innovator choice by discouraging secrecy. In a broad sense, one could consider any doctrine that strengthens patent protection to discourage secrecy. For example, if patent duration were to be increased from twenty to forty years, some inventors at the margins would undoubtedly be enticed to consider patenting over trade secrecy. This interplay between patent strength and the appeal of trade secrecy is inevitable and unavoidable. However, there are two groups of doctrines whose sole function is to discourage secrecy. This Section suggests changes to those doctrines that would better align patent doctrine with the policies outlined above.

a) Prior User Rights

Prior user rights are rights for first inventors to practice their invention regardless of whether the invention has been subsequently patented by another.[268] The lack of prior user rights encourages patenting by placing trade secret holders at risk of losing the right to practice their own invention. Indeed, commentators view the discouragement of trade secrecy as the strongest argument for denying prior user rights.[269] For example, some commentators have argued that the law correctly permits a patentee to exclude a first inventor from practicing her invention because such a risk of exclusion will encourage patenting.[270]

However, as shown above, discouraging secrecy does not have theoretical support as a policy objective. Refusing to grant prior user rights creates a sense of fear among trade secret holders and an inevitable push towards the patent system. Even critics of prior user rights have noted that while prior user rights decrease the incentive to patent, they likely increase the overall incentive to innovate.[271]

268. *See, e.g.*, James R. Barney, *The Prior User Defense: A Reprieve for Trade Secret Owners or a Disaster for Patent Law?*, 82 J. PAT. & TRADEMARK OFF. SOC'Y. 261 (2000) (describing the debate around prior user rights); Lisa M. Brownlee, *Trade Secret Use of Patentable Inventions, Prior User Rights and Patent Law Harmonization: An Analysis and Proposal*, 72 J. PAT. & TRADEMARK OFF. SOC'Y 523 (1990) (detailing the existence of prior user rights in foreign jurisdictions).

269. *See* Barney, *supra* note 268 (concluding that prior user rights would harm the public in part due to diminished disclosure); Shapiro, *supra* note 5, at 95.

270. Denicolò & Franzoni, *supra* note 102, at 517.

271. *Id.* at 529–30 (finding that in a system with prior user rights, the incentive to innovate is strengthened, although the incentive to patent is reduced).

Carl Shapiro has argued for the establishment of prior user rights.[272] He has posited that prior user rights enhance competition, reward innovation, and can partially correct problems caused by patents of questionable validity.[273] While I agree with Shapiro, this Article suggests an additional ground which supports prior user rights: granting prior user rights would more closely harmonize the law with the economic rationale underlying patent theory.

On the other hand, Vincenzo Denicolò and Luigi Franzoni have argued that prior user rights should be denied because society prefers patenting over secrecy in cases in which an inventor would prefer trade secrecy.[274] They argue that because greater deadweight losses occur when secrecy provides a greater benefit to the inventor, society should prefer patenting. However, Denicolò and Franzoni admit that denying prior user rights "reduces the incentives to innovate."[275] This is true because discouraging secrecy can decrease overall incentives to invest.

Since 1999, U.S. law has provided for some prior user rights for patents on business methods.[276] Congress is considering further legislation that would greatly expand prior user rights.[277] The business method exception for prior user rights, while desirable, does not go far enough to align patent doctrine with the reward theory. A better solution would be to grant blanket prior user rights to first inventors. Doing so would place holders of patented inventions and trade secrets on equal footing: patentees would be able to exclude others from using a patented invention, except for those inventors that invented prior to the patent application. Similarly, trade secret holders could operate knowing that later-filed patents would not subject them to infringement liability or the inability to practice their own invention.

b) Priority Rules and the One-Year Statutory Bar to Patentability

A subsequent inventor can obtain patent rights over a first inventor if the second inventor can show that the first inventor "abandoned, suppressed, or concealed" the invention at any time after the second inventor successfully

272. The law's lack of prior user rights has been criticized by commentators on the grounds of being unfair and for economic reasons. *See* Shapiro, *supra* note 5, at 95 (finding that prior user rights enhance competition, reward innovation with relatively little dead-weight loss, and more properly align the private and social incentives of innovation).

273. Shapiro, *supra* note 5, at 93.

274. Denicolò & Franzoni, *supra* note 102, at 530–34.

275. *Id.*

276. 35 U.S.C. § 273 (2006).

277. Patent Reform Act, H.R. 1260, 111th Cong. (2009).

reduced the invention to practice.[278] An invention is considered abandoned, suppressed, or concealed if an inventor fails to patent the invention within a reasonable period of time.[279] Courts weigh various factors when determining whether a delay in patenting is reasonable, but commercialization activities are not valid reasons for delay.[280] Courts have determined that commercialization means any commercial use, including use that is not observable by the public.[281]

The priority rule of Section 102(g) favors inventors who aggressively seek patent protection over those who practice inventions secretly for a time. An inventor who conceals her invention faces the risk of losing the patent rights in her invention to a subsequent patentee and is thus encouraged to patent, even if the information available at the time of selection indicates that secrecy would provide sufficient return on investment.[282]

Along with the lack of prior user rights, the priority doctrine is a threat to inventors who practice in secret. Whereas priority rules threaten the potential exclusive rights, the lack of prior user rights threatens the complete ability to practice an invention. The latter threat is potentially more worrisome for both an inventor and from an equity perspective,[283] but under current law the two go hand-in-hand. The loss of a priority battle means that one loses both the right to patent and the ability to practice the invention.

---

278. 35 U.S.C. § 102(g) (2006).

279. *See* 1 CHISUM, *supra* note 89, § 10.08[1].

280. *Id.*

281. *E.g.*, Lutzker v. Plet, 843 F.2d 1364, 1367 (Fed. Cir. 1988) ("[W]hen there is an unreasonable delay between the actual reduction to practice and the filing of a patent application, there is a basis for inferring abandonment, suppression or concealment. . . . The inventor's activities during the delay period may excuse the delay (e.g., he may have worked during that period to improve or perfect the invention disclosed in the patent application). . . . When, however, the delay is caused by working on refinements and improvements which are not reflected in the final patent application, the delay will not be excused. . . . Further, when the activities which cause the delay go to commercialization of the invention, the delay will not be excused.").

282. Priority rules also serve to limit an inventor's ability to extend the exclusivity period of an invention by tacking on a twenty-year patent term just as trade secrecy is expiring. *See* Pencock v. Dialogue, 27 U.S. (2 Pet.) 1, 19 (1829) (stating that allowing such term extension would "materially retard the progress of science and the useful arts"). One potential way of altering priority rules to both reward first invention and reduce "double-dipping" would be to reduce the patent term by the term of trade secret usage. Additionally, by granting broad prior user rights, a system in which patent priority is lost is less burdensome because patent rights would not restrict an original inventor from practicing her invention.

283. *See* Dunlop Holdings, Ltd. v. Ram Golf Corp., 524 F.2d 33, 36 n.11 (7th Cir. 1975) ("For it is less serious to hold that the first inventor has forfeited his right to a patent monopoly than it is to hold that he has forfeited any right to use his own invention without the permission of a subsequent inventor.").

Rather than focus on the public or private nature of an innovation, priority disputes ought to turn on the issue for which they were created: who invented first. First reduction to practice should be the primary concern of priority disputes. Rewarding the first inventor rather than the first non-secret inventor has both an intuitive equitable appeal,[284] as well as an economic appeal, as a means of encouraging commercialization of trade secrets. Currently, commercialization activities which delay the filing of a patent may be considered suppression of an invention.[285] The law should encourage such commercializing activities by awarding priority to the de facto first inventor.[286]

2. Encouraging Secrecy: Potential Steps

The law does not currently encourage inventors to maintain inventions as trade secrets. As described in this Part, however, economic theory suggests that at least in certain situations, it should. In general, we can assume that inventor choice will mirror the socially optimal choice. However, when a patented invention is more valuable than a trade secret and secrecy promises a sufficient return on investment, society should encourage secrecy. This final Section will begin to describe different ways in which secrecy can be actively encouraged as well as potential drawbacks from employing these changes.

a) Encouraging Secrecy Through Patent Law

When research costs are low, it can be assumed that *S* and *P* will both be greater than *R*. That is, as *R* approaches zero, both secrecy and patenting provide sufficient return on investment to induce invention. When that is the case, this Article's framework suggests that policy makers should encourage the use of trade secrecy. Often, inventors will prefer patenting in such cases because the risk of independent invention or reverse engineering is very high when research costs are minimal.

James Anton and Dennis Yao demonstrated that inventors of low-cost innovations will often rely on patent protection. They developed models that

284. *See id.*

285. *See In re* Caveney, 761 F.2d 671, 675–76 (Fed. Cir. 1985) (stating that sale of the product of a secret method triggers the on-sale bar).

286. The law regarding priority (and novelty in general) was recently modified via legislation. Leahy-Smith America Invents Act, Pub. L. No. 112-29, sec. 3, §§ 101, 102 (enacted Sept. 16, 2011). Beginning eighteen months after the enactment of the America Invents Act, priority to invention will be judged based upon the filing date of the patent application rather than the date of invention. *Id.* sec. 3(n). The move to a first-to-file system will harmonize U.S. patent law with the rest of the world and alleviate some of the concerns discussed above for applications filed after that date. For inventions filed before March 2013, however, the old priority rules still apply.

demonstrate that innovators will tend to seek patents for smaller, incremental, less valuable developments.[287] This behavior can be explained by the fact that these innovations are unlikely to be imitated, thus full disclosure does not harm the innovator.[288]

Unfortunately, it is difficult to determine ex ante which types of inventions require low investment and which do not. Much to the chagrin of economists, patent law does not concern itself with the amount of investment required to produce an innovation. Patent law's obviousness doctrine could potentially provide some help in this area. Further work on the obviousness doctrine's ability to weed out low-cost innovation could prove helpful in implementing the framework proposed in this Article.

Another potential means of encouraging the use of secrecy is to deny patents altogether to inventions that could have been maintained as trade secrets and thus do not require the patent system to encourage their creation. Section 101 of the Patent Act provides a means, albeit a heavy-handed one, of restricting the types of inventions that can be patented.[289] Any new and useful process, machine, manufacture, or composition of matter can be patented as long as it meets the other requirements of the Patent Act and does not fall under one of the non-patentable categories created by the Supreme Court, such as abstract ideas.[290]

David Olson proposed eliminating patent protection for business methods for utilitarian reasons. As Olson notes, the ability to accurately differentiate a specific class of inventions from another is requisite for effective use of Section 101 as a policy lever.[291] In the case of secret inventions, however, it is unlikely that subject matter categories will differentiate those inventions that could be profitably maintained in secret from those that cannot.

For example, consider one potential manner of distinguishing invention types: products and methods. Inventions appropriable through secrecy tend

287. James J. Anton & Dennis A. Yao, *Little Patents and Big Secrets: Managing Intellectual Property*, 35 RAND J. ECON. 1, 3 (2004).

288. *Id.* at 11–13.

289. *See* 35 U.S.C. § 101 (2006); David S. Olson, *Taking the Utilitarian Basis for Patent Law Seriously: The Case for Restricting Patentable Subject Matter*, 82 TEMP. L. REV. 181, 184 (2009) (arguing that the subject matter of inventions eligible for patenting "has developed with little explicit consideration of the utilitarian question").

290. *See* 35 U.S.C. § 101 (2006). *See generally* Bilski v. Kappos, 130 S. Ct. 3218 (2010) (holding a patent on a method of hedging risk invalid as an "abstract idea").

291. *See* Olson, *supra* note 289, at 184.

to be methods or processes and not products.[292] Products are, in general, subject to easier reverse engineering than are methods that can be performed in secret. Particular types of methods, such as manufacturing methods, chemical methods, and certain types of software are examples of inventions that are generally appropriable through secrecy.[293] However, the method/product distinction is an imperfect measure of the viability of secrecy. Certain types of method innovations are difficult to appropriate through secrecy; inventors of such methods depend on the patent system in order to obtain a return on their investment in the innovation. The blunt use of Section 101 to eliminate patentability on all, or a specific class, of methods is likely to result in reduced innovation in areas that would be socially beneficial.

b) Encouraging Secrecy Through Trade Secret Law: Secret Invention Registry

Another potential manner of encouraging secrecy involves modifying the existing system of trade secrecy. Trade secret law poses certain difficulties for inventors. Among the drawbacks of trade secret law from an inventor's perspective are the uncertainty of trade secret scope[294] and the potential loss of the right to practice one's invention if later patented.[295] One potential means of encouraging the use of secrecy is to reduce the uncertainty of those aspects of trade secret law.

As an initial step towards encouraging secrecy, a trade secret registry could be created. A trade secret registry would overcome one of the major drawbacks of litigating the misappropriation of trade secrecy: proving that a secret exists in the first place.[296] In general it can be said that proving the existence of a trade secret requires three elements: proof that the subject matter is not generally known, proof that reasonable efforts were taken by the owner to protect the secret, and proof that secrecy confers an economic advantage on the holder.[297]

---

292. *See* Oddi, *supra* note 115, at 285 n.126 (noting that processes are particularly good candidates for trade secret protection); *see also* Robert P. Merges, *Rent Control in the Patent Districts: Observations on the Grady-Alexander Thesis*, 78 VA. L. REV. 359, 376–77 (1992).

293. *See* Merges, *supra* note 292.

294. *See* James Pooley, *The Top Ten Issues in Trade Secret Law*, 70 TEMP. L. REV. 1181, 1181–82 (1997) (stating that the definitional problem of trade secret litigation is one of the most pressing issues of the law).

295. *See supra* Section IV.C.1.a.

296. *See, e.g.*, Pooley, *supra* note 294, at 1181–85 (noting that much litigation centers around whether a secret exists).

297. *See* MILGRIM, *supra* note 39, §§ 1.03–1.04.

Providing a registry in which trade secret holders can secretly catalogue valuable secrets would assist trade secret holders in demonstrating the first two elements of a trade secret. Registration would assist courts and other decision makers in analyzing the contours of the secret that the owner considered valuable enough to protect. Registration would also provide prima facie evidence of intent to protect a trade secret. One accused of misappropriation could then bring forth proof that either (1) the registered secret was well-known, or (2) the registrant had not demonstrated reasonable efforts to protect her secret. Absent such a showing, courts would assume that the registered secret was a reasonably protected secret invention.

The second benefit of a trade secret registry would be increased protection against independent patenting of the invention. As described in Section IV.C.1, *supra*, inventors who elect to commercialize their inventions in secret risk the loss of two rights: the right to eventually patent the invention and the right to practice the invention.[298] Both right losses potentially occur when a second party patents the invention.

The existence of a trade secret registry along with the establishment of legal protection for prior users would assist trade secret users in protecting against these losses. Adopting prior user rights, as urged in Section IV.C.1.a, *supra*, could lead to protracted legal battles about whether an invention was invented by a first inventor prior to patenting by a second. Registration would alleviate some of the difficulties in proving prior use and protect against false allegations of prior use. In fact, prior user rights could be extended only to those inventions that have been registered, including equivalents and obvious extensions.

Creating this registration regime would not be prohibitively expensive because registration does not require the same level of examination as patenting. Furthermore the PTO already has a statutorily authorized registration system. Section 157 of Title 35 authorizes the PTO to establish a statutory invention registration that requires applicants to give up future rights to a patent on the invention after publication of the application.[299] The statutory invention registration allows inventors to publish inventions that they do not intend to patent in a manner that precludes others from patenting the invention. The registration is used by inventors who do not desire a patent but want to prevent others from patenting.

Establishing prior user rights may disincline inventors from using the current registration system because they would no longer be at risk of losing

298. *See supra* Section IV.C.1.

299. 35 U.S.C. § 157(a)(3), (b) (2006).

the right to practice to a second inventor patentee. However, they may desire some means of demonstrating their prior use before an infringement action arises. Such non-public registration would not serve as prior art as the current registration does, because submissions would not be made public. However, registration would serve as prima facie evidence of prior invention and prior user rights.[300]

V. CONCLUSION

Patent law has long relied on the perceived wisdom that patenting is preferable to secrecy. This Article turns that logic on its head. Reliance on secrecy can have numerous underappreciated social benefits, including targeted disclosure, more rapid commercialization, and increased incentives to invent. Many of these benefits are the result of the competition-enhancing aspects of trade secrets and the lack of barriers to innovative entry.

The framework developed in this Article leads to two primary policy results. First, the legal system should not discourage the use of trade secrecy; rational inventors will select trade secret protection only in instances when it is also socially optimal without interference from policy makers. To eliminate the disincentives to rely on secrecy that exist in current law, this Article suggests establishing prior user rights and altering the standards for patent priority disputes.

Second, the secrecy framework that this Article has developed suggests that in certain circumstances secrecy should be encouraged. Without intervention from policy makers, patents will be the preferred method of protection more often than is socially desirable. The use of trade secrets should be encouraged when both secrecy and patenting provide sufficient incentives to invent. In such situations, the use of trade secrecy is socially preferable because identical amounts of innovation are produced but with fewer legal restrictions on competition. Furthermore, the fundamental economic concern of patent law's reward theory—free riders—is inapplicable in such cases.

Trade secrets can promote innovation. They do so in a limited set of cases, but they do so without many of the drawbacks associated with patents. Trade secrets have traditionally not been viewed as a means of incentivizing

300. The current publication requirement in the registration system would dissuade trade secret holders from registration. *See* § 157(b), (c). In order to alleviate this, the statute authorizing the invention registration would need to be altered to grant registrants the option of non-publication.

innovation or a means of encouraging inventors to refrain from patenting. Courts and policy makers should, however, view secrecy as a low-cost means of encouraging competition and innovation.

INTRUSIVE MONITORING: EMPLOYEE PRIVACY EXPECTATIONS ARE REASONABLE IN EUROPE, DESTROYED IN THE UNITED STATES

Lothar Determann[†] *& Robert Sprague*[††]

TABLE OF CONTENTS

---

† Dr. iur habil, Privatdozent, Freie Universität Berlin; Adjunct Professor, University of California, Berkeley School of Law and Hastings College of the Law, and Stanford Law School; Partner, Baker & McKenzie, San Francisco, California.

†† J.D., M.B.A. Associate Professor, University of Wyoming College of Business Management & Marketing.

The authors thank Aaron J. Lyttle, J.D. 2010, University of Wyoming College of Law, for his excellent research assistance for this Article, and for contributions from Charles W. Weinroth, J.D. Candidate 2011, University of California, Hastings College of the Law, and Benjamin Bäuerle, Associate, Baker & McKenzie, Munich, Germany.

I. INTRODUCTION

An increasingly global workforce communicates, collaborates, and connects in multinational enterprises and worldwide marketplaces with web- and cloud-based technologies across geographies and territorial borders. Globalization has leveled many historic differences, in the workplace and elsewhere. But, the law on workplace privacy could hardly be more different in the United States and the European Union. This difference raises significant challenges for the global employer who manages and monitors worldwide human resources with global processes and technologies. Additionally, this difference raises fundamental questions as to its origins in workplace privacy standards and why these differences resist convergence so stubbornly.

This Article examines the contrasting policy and legal frameworks relating to data privacy in the United States and the European Union, with a particular focus on workplace privacy and intrusive surveillance technologies and practices. Part II of this Article examines the U.S. perspective on modern work-related employer monitoring practices, the laws giving rise to possible employee privacy rights, and specific types of employer monitoring that may lead to actionable invasions of employee privacy rights. Part III then addresses the issue of employee privacy from the EU perspective, beginning with an overview of the formation of authority to protect individual privacy rights, followed by an analysis of the principal areas of protection and their application. Part IV then provides comparison and conclusions regarding the fundamental differences between the United States and the European Union in employee privacy protection.

II. EMPLOYER MONITORING AND EMPLOYEE PRIVACY—U.S. PERSPECTIVE

U.S. employers engage in a variety of work-related monitoring practices for a range of legitimate business purposes. In general, the right to privacy in the United States is conditioned on a "reasonable expectation of privacy," which is determined by the surrounding circumstances and society's[1] or a "reasonable person['s]"[2] views. Employees in the United States tend to have minimal expectations of privacy in the workplace at the outset. Employers usually destroy any remaining limited expectations via notices and warnings regarding monitoring in employee handbooks, computer log-on splash screens, electronic systems use policies, and privacy statements. Yet, there are new and ever-evolving types of monitoring that can catch employees by surprise and challenge employers' efforts in keeping their workforce aware of advances in technology. This challenge threatens employers' efforts to prevent any development of privacy expectations that could lead to privacy rights and their violation through intrusive surveillance.

A. WORK-RELATED EMPLOYER MONITORING

In the modern office, internet access and e-mail have become ubiquitous.[3] Wireless communications, global positioning systems (GPS), and radio frequency identification (RFID) chips are now common business tools.[4] Along with increased use of computers and communications systems at work comes increased computer and communications monitoring. Typical

1. Katz v. United States, 389 U.S. 347, 360–61 (1967) (Harlan, J., concurring). *See also* TBG Ins. Servs. Corp. v. Superior Court, 117 Cal. Rptr. 2d 155, 160 (Ct. App. 2002) ("When affirmative relief is sought to prevent a constitutionally prohibited invasion of privacy, the plaintiff must establish (1) a legally protected privacy interest; (2) a reasonable expectation of privacy in the circumstances; and (3) conduct by defendant constituting a serious invasion of privacy.") (citation and internal quotation marks omitted) (applying CAL. CONST. art. I, § 1).

2. *Katz*, 389 U.S. at 363 (White, J., concurring).

3. Qinyu Liao et al., *Workplace Management and Employee Misuse: Does Punishment Matter?*, 50 J. COMPUTER INFO. SYS. 49, 49 (2009). According to a 2008 Pew Internet & American Life Project survey, nearly one-third of American adults use e-mail or the Internet in their work. *See* MARY MADDEN & SYDNEY JONES, PEW/INTERNET, NETWORKED WORKERS, at i (2008), http://www.pewinternet.org/~/media//Files/Reports/2008/PIP_Networked_Workers_FINAL.pdf.pdf (reporting that of the 53% of American adults employed full- or part-time, 62% use e-mail or the Internet at work).

4. *See, e.g.*, NAT'L TELECOMM. & INFO. ADMIN. (NTIA) AND ECON. & STATISTICS ADMIN. (ESA), U.S. DEP'T OF COMMERCE, A NATION ONLINE: HOW AMERICANS ARE EXPANDING THEIR USE OF THE INTERNET 57–64 (2002) [hereinafter A NATION ONLINE]; Marisa Anne Pagnattaro, *Getting Under Your Skin—Literally: RFID in the Employment Context*, 2008 J.L. TECH. & POL'Y 237, 238 (2008); William P. Smith & Filiz Tabak, *Monitoring Employee E-mails: Is There Any Room for Privacy?*, 23 ACAD. MGMT. PERSP. 33, 33 (2009).

work-related monitoring includes scanning of sent and received e-mails by anti-virus and anti-spam software. This software monitors websites accessed by employees, as well as scans messages and attachments to block code that is considered harmful and content that is presumed inappropriate. Some employers use more intrusive methods: tracking a worker's every keystroke and mouse click; capturing screen shots to monitor communications via remote computing platforms outside the control of the employer's networks (such as webmail and blogging); storing copies of e-mail messages sent and received on servers where individual workers cannot access or delete the messages; logging information on actions performed by workers, including the applications used and the files accessed and printed; monitoring internet access, online sessions, and electronic chat conversations; and remotely viewing what the worker is doing in real time.[5] This monitoring is not restricted to the "workplace" per se, as a substantial number of people use computers in their homes and on the road to perform work on company-owned devices or even privately-owned devices which can be scanned while they are connected to the company network.[6] The latest widely-cited survey of workplace monitoring reveals that significant percentages of employers monitor employee internet usage (66%), e-mail (43%), and time spent on the phone and numbers called (45%), while 16% of employers record phone calls and 9% record voice mail messages.[7]

Employers in the United States monitor employees for three primary reasons: protecting information and other intellectual property assets; increasing productivity; and avoiding liability, including exposure associated with copyright infringement by employees, other improper uses of

5. *See* H. Joseph Wen, Dana Schwieger & Pam Gershuny, *Internet Usage Monitoring in the Workplace: Its Legal Challenges and Implementation Strategies*, 24 INFO. SYS. MGMT. 185, 186 (2007).

6. *See, e.g.*, A NATION ONLINE, *supra* note 4, at 62 ("[A]pproximately 24 million of the 65 million employed adults who use a computer at work also do work on a computer at home . . . ."); MADDEN & JONES, *supra* note 3, at v (reporting that 50% of employed e-mail users check their work e-mail on weekends); Laura Merritt, *Factor Gadgets into Remote-Access Policies*, N.Y. L.J., Apr. 27, 2010, at 5, *available at* http://www.law.com/jsp/nylj/PubArticle NY.jsp?id=1202453204420&The_Mobile_Workforce (noting that employers make available remote access and mobile devices to both high and low level employees who must then respond to e-mails and make calls on cell phones outside the workplace); Smith & Tabak, *supra* note 4, at 33 (noting that new communications devices are at least partially responsible for the blurring of work-life boundaries).

7. AM. MGMT. ASS'N (AMA) & EPOLICY INST., 2007 ELECTRONIC MONITORING AND SURVEILLANCE SURVEY 1–3 (2007), *available at* http://www.plattgroupllc.com/jun08/2007 ElectronicMonitoringSurveillanceSurvey.pdf.

computers by employees, or hostile work environments.[8] All employers want to ensure that confidential and proprietary information is not purposely or inadvertently disclosed by employees, or improperly accessed by individuals outside the firm.[9] Employers are also concerned about "junk computing"[10] and "cyberloafing."[11] Various surveys reveal that employees spend a significant amount of time at work surfing the Internet for non-work-related purposes and sending and reading personal e-mail messages.[12] One recent

8. Employers also justify monitoring and surveillance based on the argument that the organization owns the computers and equipment that employees use to do their jobs, so the organization has "both a right and an interest in policing the use of those facilities." JEFFREY M. STANTON & KATHRYN R. STAM, THE VISIBLE EMPLOYEE 116 (2006).

9. *See, e.g.*, William G. Porter II & Michael C. Griffaton, *Between the Devil and the Deep Blue Sea: Monitoring the Electronic Workplace*, 70 DEF. COUNS. J. 65, 66 (2003); Marian K. Riedy & Joseph H. Wen, *Electronic Surveillance of Internet Access in the American Workplace: Implications for Management*, 19 INFO. & COMM. TECH. L. 87, 91 (2010); Smith & Tabak, *supra* note 4, at 34; *see also* United States v. Martin, 228 F.3d 1 (1st Cir. 2000) (upholding conviction of theft of trade secrets based on defendant's e-mail correspondence with competitor's employee); PROOFPOINT, OUTBOUND EMAIL AND DATA LOSS PREVENTION IN TODAY'S ENTERPRISE, 2010, at 4 (2010), http://www.proofpoint.com/id/outbound/index.php (reporting survey data revealing percentages of surveyed firms reporting it is common or very common for outbound e-mail messages to contain valuable intellectual property or trade secrets which should not leave the organization (23%) or confidential or proprietary business information about the organization (21%)). *See generally* Michael L. Rustad & Sandra R. Paulsson, *Monitoring Employee E-mail and Internet Usage: Avoiding the Omniscient Electronic Sweatshop: Insights from Europe*, 7 U. PA. J. LAB. & EMP. L. 829, 838 (2005) (discussing the need for employers to monitor employee communications to prevent the loss of intellectual property rights).

10. *See, e.g.*, Ruth Guthrie & Paul Gray, *Junk Computing: Is It Bad for an Organization?*, 13 INFO. SYS. MGMT. 23 (1996) (defining "junk computing" as "the use of information systems in a way that does not directly advance organizational goals" and including examples of unnecessarily sending e-mail messages to multiple recipients and playing computer games).

11. *See, e.g.*, Vivien K. G. Lim, *The IT Way of Loafing on the Job: Cyberloafing, Neutralizing and Organizational Justice*, 23 J. ORG. BEHAV. 675, 677 (2002) (defining "cyberloafing" as "any voluntary act of employees using their companies' internet access during office hours to surf non-job-related Web sites for personal purposes and to check (including receiving and sending) personal e-mail"); *see also* Murugan Anandarajan, *Internet Abuse in the Workplace*, 45 COMM. ACM 53, 53 (2002) (characterizing the world wide web as providing "employees access to the world's biggest playground").

12. *See* Lim, *supra* note 11, at 676 (summarizing several surveys revealing various degrees to which employees use employers' computers and communications systems for personal uses). *But see* Riedy & Wen, *supra* note 9, at 90 (arguing personal use of the Internet and e-mail could make employees more productive and asserting that there is no direct evidence of decreased employee productivity by their sending an e-mail message rather than chatting with a colleague in the break room). *See also* Weber v. Univs. Research Ass'n, 621 F.3d 589, 590–92 (7th Cir. 2010) (affirming the granting of the defendant employer's motion for summary judgment in a case arising from the dismissal of an employee where monitoring revealed that the plaintiff employee had spent some sixteen hours in one week accessing non-work-related websites and frequently accessed personal e-mail accounts associated with the plaintiff's outside business).

survey indicates that over one-quarter of employers have fired employees for internet and e-mail abuse.[13] In fact, public companies in the United States are required to implement whistleblower hotlines and investigate inappropriate conduct as part of their overall obligation to avoid material weaknesses in their processes to ensure compliance with applicable law.[14]

Employers engage in work-related monitoring also in an effort to limit potential liability. There is concern some employees may be downloading music, movies, and other materials in violation of copyright laws, which could result in the employer facing vicarious liability through the doctrine of respondeat superior.[15] There are other improper uses of company computers that can possibly put employers at risk. For example, in one case, an employer was found potentially liable to the wife of an employee who had published nude pictures of the wife's daughter on the Internet using the employer's computer system.[16] Additionally, employers have concerns regarding the content of e-mail messages revealed in litigation-related discovery.[17]

---

13. AMA & EPOLICY INST., *supra* note 7, at 1.

14. *See* Cynthia Jackson, *A Global Whistle-Stop Tour*, DAILY J., Feb. 19, 2009, at 7, *available at* http://www.bakermckenzie.com/files/Publication/b3442009-d314-4585-a396-f1ec419acc6e/Presentation/PublicationAttachment/4840fa55-490c-448a-983f-fbdb28b9f7f5/ar_sfpa_DJ8GlobalWhistleStopTour_feb09.pdf.

15. Smith & Tabak, *supra* note 4, at 34; *see RIAA Collects $1 Million from Company Running Internal Server Offering Thousands of Songs*, RIAA (Apr. 9, 2002), http://www.riaa.com/newsitem.php?news_month_filter=4&news_year_filter=2002&resultpage=2&id=E9996E0C-D33C-CA18-851A-19690EE763FA (announcing settlement of copyright infringement claims against a company that allegedly permitted its employees to access and distribute thousands of infringing music files over its computer network).

16. Doe v. XYC, Corp., 887 A.2d 1156 (N.J. Super. Ct. App. Div. 2005). In *Doe*, computer technicians and supervisors were aware the employee was using the employer's computer system to visit pornographic websites while at work, but no action was taken due to the employer's policy to not monitor the internet activities of its employees. *Id.* at 1158–60. The court held:

> [A]n employer who is on notice that one of its employees is using a workplace computer to access pornography, possibly child pornography, has a duty to investigate the employee's activities and to take prompt and effective action to stop the unauthorized activity, lest it result in harm to innocent third-parties.

Id. at 1158.

17. *See* AMA & EPOLICY INST., *supra* note 7, at 2 ("Workers' e-mail and other electronically stored information create written business records that are the electronic equivalent of DNA evidence. As a result, 24% of employers have had e-mail subpoenaed by courts and regulators and another 15% have battled workplace lawsuits triggered by employee e-mail . . . ."); Linda Sandler, *"Stupid" Lehman E-Mails Didn't Stay "Just Between Us,"* BLOOMBERG (June 11, 2010, 7:06 AM), http://www.bloomberg.com/news/2010-06-11/lehman-probe-lesson-avoid-big-trouble-by-shunning-stupid-e-mail-terms.html

Employers are also concerned inappropriate e-mail and text messages and internet use could spur hostile work environment complaints.[18] In *Burlington Industries, Inc. v. Ellerth*, and its companion case *Faragher v. City of Boca Raton*, the U.S. Supreme Court held that an employer is subject to vicarious liability to a victimized employee for an actionable hostile environment created by a supervisor.[19] However, when no tangible employment action is taken, an employer may raise as a defense that it exercised reasonable care to prevent and correct promptly any sexually harassing behavior.[20] As a result of this defense's requirements, employers are under greater pressure to take steps to prevent their computer and communications systems from being used to create a hostile work

(describing techniques investigators used to search thirty-four million pages of Lehman Brothers Holdings Inc. e-mails and reports).

18. A hostile work environment is created when "[u]nwelcome sexual advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature . . . unreasonably interfere[s] with an individual's work performance or creat[es] an intimidating, hostile, or offensive working environment." 29 C.F.R. § 1604.11(a) (2010) (cited with approval in Meritor Sav. Bank, FSB v. Vinson, 477 U.S. 57, 65 (1986)). Additional protected classes, particularly race, are also protected from hostile work environments. *See, e.g.*, Curtis v. DiMaio, 46 F. Supp. 2d 206, 212–14 (E.D.N.Y. 1999), *aff'd*, 205 F.3d 1322 (2d Cir. 2000); Daniels v. WorldCom, Inc., No. CIV.A.3:97-CV-0721-P, 1998 WL 91261 (N.D. Tex. Feb. 23, 1998) (holding distribution of four racist e-mail messages within the company's e-mail system did not create an actionable hostile environment where the employer had taken prompt remedial action); Owens v. Morgan Stanley & Co., No. 96 CIV. 9747(DLC), 1997 WL 793004 (S.D.N.Y. Dec. 24, 1997) (addressing a hostile work environment claim based on race and holding that a single racist e-mail message does not create an actionable hostile environment).

19. Burlington Indus., Inc. v. Ellerth, 524 U.S. 742, 765 (1998); Faragher v. City of Boca Raton, 524 U.S. 775, 807 (1998).

20. Specifically, when no tangible employment action is taken, a defending employer may raise an affirmative defense to liability or damages comprised of two necessary elements: "(a) that the employer exercised reasonable care to prevent and correct promptly any sexually harassing behavior, and (b) that the plaintiff employee unreasonably failed to take advantage of any preventive or corrective opportunities provided by the employer or to avoid harm otherwise." *Burlington*, 524 U.S. at 765; *Faragher*, 524 U.S. at 807.

environment.[21] Today, employers cite efforts to prevent hostile work environments as a primary motivation for workplace surveillance.[22]

B. WORK-RELATED EMPLOYEE PRIVACY

There are three primary sources of privacy protection in the United States: the Constitution, common law, and statutes. While constitutional and common law rights to privacy have different origins and apply to different actors, they share many commonalities. As shown in Section II.B.1, *infra*, constitutional requirements for a recognized right to privacy often lay the foundation for common law privacy rights. While there are a variety of privacy-related statutes in the United States, they offer only marginal protections for employees.

1. Work-Related Rights to Privacy Under the Constitution

The U.S. Constitution provides for civil rights of individuals against actions of state actors, i.e., state and federal governments, including government employers, but not against actions of private employers.[23] The Constitution does not mention privacy expressly, but a right to privacy has

21. *See, e.g.*, *Burlington*, 524 U.S. at 770 (Thomas, J., dissenting) ("Sexual harassment is simply not something that employers can wholly prevent without taking extraordinary measures—constant video and audio surveillance, for example—that would revolutionize the workplace in a manner incompatible with a free society.") (citation omitted); Ellerth v. Burlington Indus., Inc., 123 F.3d 490, 513 (7th Cir. 1997) (Posner, C.J., dissenting), *aff'd*, 524 U.S. 742 (1998) ("It is facile to suggest that employers are quite capable of monitoring a supervisor's actions affecting the work environment. Large companies have thousands of supervisory employees. Are they all to be put under video surveillance?").

22. *See* Marc A. Sherman, *Webmail at Work: The Case for Protecting Against Employer Monitoring*, 23 TOURO L. REV. 647, 657 (2007). In a recent workplace surveillance survey, of the twenty-eight percent of employers reporting they had fired an employee for misuse of e-mail, sixty-two percent did so because of offensive or inappropriate language or content; and of the thirty percent of employers reporting they had fired an employee for misuse of the Internet, eighty-four percent did so because of viewing, downloading, or uploading inappropriate or offensive content. AMA & EPOLICY INST., *supra* note 7, at 8–9; *see* Forrester v. Rauland-Borg Corp., 556 F. Supp. 2d 850, 851 (N.D. Ill. 2005), *aff'd*, 453 F.3d 416 (7th Cir. 2006) (upholding dismissal of employee for engaging in sexual harassment based, in part, on sending obscene e-mail messages to harassment victims); *see also* PROOFPOINT, *supra* note 9, at 3 (reporting survey data revealing that 18% of surveyed firms report it is common or very common for outbound e-mail messages to contain adult, obscene, or potentially offensive content).

23. In contrast, the California Constitution protects a right to privacy expressly and expands this protection also to relations between individuals, California private-sector employers, and their employees. CAL. CONST. art. I, § 1 ("All people are by nature free and independent and have inalienable rights. Among these are enjoying and defending life and liberty, acquiring, possessing, and protecting property, and pursuing and obtaining safety, happiness, and privacy.").

been inferred relative to searches and seizures permissible under the Fourth Amendment.[24] In circumstances in which a person has a reasonable expectation of privacy, an invasion of that area of privacy by a government entity is presumptively unreasonable in the absence of a search warrant.[25] This Fourth Amendment implied right to privacy in limited circumstances lays the foundation for potential privacy rights for public-sector employees.[26]

As most recently reaffirmed by the U.S. Supreme Court in *City of Ontario, California v. Quon*, the starting point for determining work-related privacy for public-sector employees is found in the plurality opinion in *O'Connor v. Ortega*.[27] "Individuals do not lose Fourth Amendment rights merely because they work for the government instead of a private employer."[28] "Searches and seizures by government employers or supervisors of the private property of their employees, therefore, are subject to the restraints of the Fourth Amendment."[29]

The subject of a warrantless search must first have a reasonable expectation of privacy in the item or area searched before the search can be deemed unconstitutional.[30] In the public workplace, however, even if the employee has a reasonable expectation of privacy, a warrantless search may

---

24. The Fourth Amendment states:

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

U.S. CONST. amend. IV; *see* Katz v. United States, 389 U.S. 347, 351 (1967) (holding that what a person seeks to preserve as private may be protected under the Fourth Amendment).

25. *Katz*, 389 U.S. at 360–61 (Harlan, J., concurring).

26. As of December 2008, the federal government employed approximately 2.5 million people. *Federal Government Civilian Employment by Function: December 2008*, U.S. CENSUS BUREAU (2009), http://www2.census.gov/govs/apes/08fedfun.pdf (representing approximately 1.5% of the employed U.S. workforce). As of March 2008, the states employed nearly 4.4 million people on a full-time equivalent basis. *State Government Employment Data: March 2008*, U.S. CENSUS BUREAU (2009), http://www2.census.gov/govs/apes/08stus.txt (representing approximately 3% of the employed U.S. workforce); *see also Table 588. Employed Civilians and Weekly Hours: 1980 to 2008*, U.S. CENSUS BUREAU (2009), http://www.census.gov/compendia/statab/2010/tables/10s0588.pdf (reporting just over 145 million civilian employees in 2008).

27. City of Ontario, Cal. v. Quon, 130 S. Ct. 2619 (2010); O'Connor v. Ortega, 480 U.S. 709 (1987) (involving the search of a medical doctor's office by hospital administrators for disputed purposes).

28. *O'Connor*, 480 U.S. at 717.

29. *Id.* at 715.

30. *Id.*; *Katz*, 389 U.S. at 360–61 (Harlan, J., concurring).

still be reasonable and not in violation of the Fourth Amendment.[31] As such, the Supreme Court believes public employers must be given wide latitude when conducting work-related, non-investigatory searches, as well as for investigations of employee misconduct.[32] Whether a warrantless search by a government-employer that violates the reasonable expectations of privacy of an employee is permissible depends on the reasonableness of the intrusion,[33] which is determined by a two-step process: first, whether the action was justified at its inception; and second, whether the search as actually conducted was reasonably related in scope to the circumstances which justified the intrusion in the first place.[34]

As noted above, the Supreme Court reaffirmed the *O'Connor v. Ortega* plurality in *City of Ontario, California v. Quon*.[35] In *Quon*, the City of Ontario, California issued pagers to its SWAT officers, including Jeff Quon, to help them mobilize and respond to emergency situations.[36] The City had a "Computer Usage, Internet and E-Mail Policy," which stated employees should not have an expectation of privacy in e-mail messages. Although the Policy did not explicitly mention text messages,[37] the record indicates Quon was informed that the City considered text messages to be just like e-mail messages.[38]

---

31. *See O'Connor*, 480 U.S. at 719–20 ("In the case of searches conducted by a public employer, [courts] must balance the invasion of the employees' legitimate expectations of privacy against the government's need for supervision, control, and the efficient operation of the workplace.").

32. *Id.* at 723–24. The primary justification for this approach is that "in contrast to law enforcement officials . . . public employers are not enforcers of the criminal law; instead, public employers have a direct and overriding interest in ensuring that the work of the agency is conducted in a proper and efficient manner." *Id.* at 724. *But cf.* U.S. v. Warshak, 631 F.3d 266, 274, 283–89 (6th Cir. 2010) (holding that warrantless government seizure of defendant's e-mail messages during criminal investigation violated his Fourth Amendment rights).

33. *O'Connor*, 480 U.S. at 726.

34. *Id.* ("Ordinarily, a search of an employee's office by a supervisor will be 'justified at its inception' when there are reasonable grounds for suspecting that the search will turn up evidence that the employee is guilty of work-related misconduct, or that the search is necessary for a noninvestigatory work-related purpose such as to retrieve a needed file. . . . The search will be permissible in its scope when the measures adopted are reasonably related to the objectives of the search and not excessively intrusive in light of the nature of the misconduct.") (citation, internal quotation marks, and alterations omitted).

35. 130 S. Ct. 2619 (2010).

36. *Id.* at 2625.

37. The type of electronic messages that could be sent and received through the City-provided pagers.

38. *Id.* at 2625.

The City provided the pagers through an outside vendor, Arch Wireless, which charged the City a monthly base rate, plus additional fees for usage in excess of a set number of alphanumeric characters.[39] Quon, along with a few additional officers, quickly exceeded the monthly base usage allotment for their individual pagers.[40] The SWAT team's supervisor, Lieutenant Duke, told Quon and the other SWAT officers that as long as they paid the overage fees, he would not audit their pager text messages.[41] Over the next few months, Quon and other SWAT officers exceeded their monthly base allotment of characters sent and received and paid the overage charges for their individual pagers.[42] Over time Duke grew tired, as he put it, of "being a bill collector."[43] The Chief of Police then decided to audit the pager messages, ostensibly "to determine whether the existing character limit was too low—that is, whether officers such as Quon were having to pay fees for sending work-related messages—or if the overages were for personal messages."[44] An audit of text messages sent and received by Quon revealed that in one month alone, Quon sent or received 456 messages during work hours, of which no more than fifty-seven were work related.[45] As a result, Quon was "allegedly" disciplined.[46]

Quon, along with some of those with whom he communicated via his City-provided pager, sued the City and Arch Wireless for, inter alia, violation of their Fourth Amendment rights. Although the District Court agreed Quon had a reasonable expectation of privacy in the pager messages, it ruled that if a jury found the purpose of the audit was to determine the efficacy of the pager text limits, the City did not violate Quon's Fourth Amendment rights.[47]

---

39. *Id.*

40. *Id.* at 2625–26.

41. *Id.* at 2625.

42. *Id.* at 2625–26.

43. *Id.* at 2626.

44. *Id.*

45. *Id.*

46. *Id.* None of the court opinions specify whether Quon was directly disciplined as a result of his personal pager messages, though his personal use of the pagers resulted in an internal affairs investigation. *See, e.g.*, Quon v. Arch Wireless Operating Co., 445 F. Supp. 2d 1116, 1127 (C.D. Cal. 2006), *aff'd in part, rev'd in part*, 529 F.3d 892, 898 (9th Cir. 2008), *rev'd sub nom.* City of Ontario, Cal. v. Quon, 130 S. Ct. 2619 (2010). The District Court did explain one possible indirect negative impact for Quon as a result of the review of his pager messages: many of the messages were between Quon and his then-mistress, who had earlier been dismissed as a dispatcher for the City of Ontario due to improper conduct; Quon's then-wife believed she was denied a job with a different police force when Quon's messages with his mistress came to light. *Arch Wireless*, 445 F. Supp. 2d at 1127–28.

47. *Id.* at 1146. The District Court granted Arch Wireless's motion for summary judgment. *Id.* at 1138.

The Ninth Circuit Court of Appeals reversed, in part, agreeing Quon had a reasonable expectation of privacy, but ruling the search was unreasonable because it was not conducted in the least intrusive manner possible.[48]

In a nearly unanimous decision, the Supreme Court reversed the Ninth Circuit.[49] The Court held that even assuming Quon had a reasonable expectation of privacy in his text messages and that the City's review of those messages constituted a search within the meaning of the Fourth Amendment,[50] because the search was motivated by a legitimate work-related purpose, and because the measures were not excessive in scope given this purpose, it was reasonable under the *O'Connor* plurality.[51]

As such, although the U.S. Supreme Court recognizes that public employees may have limited, reasonable expectations of privacy in the workplace, the public employer is still free to search employees' offices, communications, and private property as long as the search is not overly intrusive—i.e., as long as it has a rational work-related justification and is limited in scope to that work-related justification. As discussed more fully in Section II.C, *infra*, the Supreme Court did not provide much helpful guidance for when an employee has a reasonable expectation of privacy in the workplace, as well as what constitutes an overly intrusive search that can impermissibly invade an employee's right to privacy.

2. Work-Related Rights to Privacy Under the Common Law

Private-sector employees do not enjoy any Fourth Amendment rights vis-à-vis searches or surveillance by their employers under the U.S. Constitution.[52] Any work-related privacy rights that private employees may have are derived from a common law right to privacy developed among the states during the twentieth century. These common law rights to privacy consist of four categories: (1) intrusion upon seclusion; (2) public disclosure of embarrassing private facts; (3) publicity which places a person in a false

48. Quon v. Arch Wireless Operating Co., 529 F.3d 892, 908–09 (9th Cir. 2008), *reh'g denied en banc*, 554 F.3d 769 (2009), *rev'd sub nom.* City of Ontario, Cal. v. Quon, 130 S. Ct. 2619 (2010). The Ninth Circuit also reversed the granting of Arch Wireless's motion for summary judgment. *Id.* at 903.

49. Justice Scalia joined in all but Section III.A of the majority opinion. *Quon*, 130 S. Ct. at 2624.

50. *Id.* at 2630.

51. *Id.* at 2632.

52. Georgia v. Randolph, 547 U.S. 103, 146 (2006) (Thomas, J., dissenting) ("[O]nly the action of an agent of the government can constitute a search within the meaning of the Fourth Amendment . . . .") (citation omitted); Skinner v. Ry. Labor Execs. Ass'n, 489 U.S. 602, 614 (1989) ("[T]he Fourth Amendment does not apply to a search or seizure, even an arbitrary one, effected by a private party on his own initiative . . . .").

light in the public eye; and (4) commercial appropriation of a person's name or likeness.[53] Of these four types of common law rights to privacy, intrusion upon seclusion is the most common tort that private-sector employees allege when they believe their privacy has been invaded by their employer.[54] As with public-sector employees, a private-sector employee must also first have a reasonable expectation of privacy relative to the intrusion.[55] In addition, the

53. William L. Prosser, *Privacy*, 48 CALIF. L. REV. 383, 389 (1960). These four invasions were later more formally codified in the Restatement. The first being: "Intrusion upon Seclusion[:] One who intentionally intrudes, physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns, is subject to liability to the other for invasion of his privacy, if the intrusion would be highly offensive to a reasonable person." RESTATEMENT (SECOND) OF TORTS § 652B (1997). Second: "Appropriation of Name or Likeness[:] One who appropriates to his own use or benefit the name or likeness of another is subject to liability to the other for invasion of his privacy." *Id.* § 652C. The third being:

> Publicity Given to Private Life[:] One who gives publicity to a matter concerning the private life of another is subject to liability to the other for invasion of his privacy, if the matter publicized is of a kind that (a) would be highly offensive to a reasonable person, and (b) is not of legitimate concern to the public.

Id. § 652D. And the fourth:

> Publicity Placing Person in False Light[:] One who gives publicity to a matter concerning another that places the other before the public in a false light is subject to liability to the other for invasion of his privacy, if (a) the false light in which the other was placed would be highly offensive to a reasonable person, and (b) the actor had knowledge of or acted in reckless disregard as to the falsity of the publicized matter and the false light in which the other would be placed.

Id. § 652E. Although New York was the first state to enact a specific right to privacy statute, it was limited solely to the commercial appropriation of a person's name or likeness without permission. N.Y. C.P.L.R. § 50 (McKinney 2008). This New York statute was held constitutional in *Rhodes v. Sperry & Hutchinson Co.*, 85 N.E. 1097 (N.Y. 1908), *aff'd*, 220 U.S. 502 (1911); *see also* Robert E. Mensel, *"Kodakers Lying in Wait": Amateur Photography and the Right of Privacy in New York, 1885–1915*, 43 AM. Q. 24, 25 (1991) (noting New York was the first state to enact a privacy statute). New York does not recognize a common law right to privacy. *See* Chimarev v. TD Waterhouse Investor Servs., Inc., 280 F. Supp. 2d 208, 216 (S.D.N.Y. 2003) (citing Howell v. N.Y. Post Co., 612 N.E.2d 699, 703 (N.Y. 1993)).

54. *See, e.g.*, Thygeson v. U.S. Bancorp, No. CV-03-467-ST, 2004 WL 2066746 (D. Or. Sept. 15, 2004); McLaren v. Microsoft Corp., No. 05-97-00824-CV, 1999 WL 339015 (Tex. App. May 28, 1999); K-Mart Corp. Store No. 7411 v. Trotti, 677 S.W.2d 632 (Tex. App. 1984).

55. *See, e.g.*, *Thygeson*, 2004 WL 2066746, at *18–21 (holding that the employee did not have a reasonable expectation of privacy in e-mail messages accessed from work although stored in a personal e-mail account, where the employer prohibited such conduct); *McLaren*, 1999 WL 339015, at *4 (concluding that the employee did not have a reasonable expectation of privacy in e-mail messages sent across and stored on the employer's computer system; distinguishing *K-Mart Corp.*); *K-Mart Corp.*, 677 S.W.2d at 640 (holding that the employer intruded upon an area where the employee had a "legitimate expectation of privacy" by searching the employee's locker which was secured by the employee's personal lock).

common law tort of intrusion upon seclusion requires that the intrusion be highly offensive to be actionable.[56]

Although public and private sector employees' work-related rights to privacy derive from different sources, all employees working in the United States face a common constraint on these rights: the employee must have a reasonable expectation of privacy, i.e., an actual expectation "that society is prepared to recognize as 'reasonable.' "[57] Employers can destroy actual expectations through the use of notices and consent forms. However, as the following review of cases and surveillance methods shows, the level of detail and specificity of such notices must increase when the intrusiveness of the surveillance program increases.[58] As a result, it is possible that courts may raise the bar for sufficient detail in notices so high that it cannot practically be met with respect to overly intrusive technologies. After all, employers are subject to operational limitations; updating monitoring notices to capture every new type of technology and monitoring measure provides a practical challenge. Additionally, employers compete for talent, and disclosing overly intrusive monitoring practices would deter candidates and drive away talent.[59] But, courts have stopped at raising the bar for notices and thus far have not clearly acknowledged an absolute core of privacy expectations that is protected against notices and consent altogether.[60] With no such common

56. *See* RESTATEMENT (SECOND) OF TORTS § 652B; *see also McLaren*, 1999 WL 339015, at *5 (holding that even if the employee could establish a reasonable expectation of privacy in e-mail messages sent across and stored on the employer's computer system, the employer's interception of those messages was not highly offensive).

57. Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., concurring).

58. *See* discussion *infra* Section II.C.

59. Barry A. Friedman & Lisa J. Reed, *Workplace Privacy: Employee Relations and Legal Implications of Monitoring Employee E-mail Use*, 19 EMP. RTS. & EMP. POL'Y J. 75, 81 (2007).

60. *See, e.g.*, Feminist Women's Health Ctr. v. Superior Court, 61 Cal. Rptr. 2d 187, 196 (Ct. App. 1997) (holding that a health center's requirement that female health workers perform vaginal and cervical self-examinations in front of co-workers and patients did not violate a health worker's right to privacy because the health workers were notified of the requirement in written policies). In *Feminist Women's Health Center*, the employee argued that the self-examination requirement, which mandated that the plaintiff "disrobe and insert a speculum in [her] vagina in front of a group of health workers," was an egregious breach of her right to privacy as protected by the California Constitution. 61 Cal. Rptr. 2d at 195. The court held that "[t]he Center was not obligated to hire plaintiff, and consent remains a viable defense even in cases of serious privacy invasions." *Id.* at 196 (citing Hill v. Nat'l Collegiate Athletic Ass'n, 865 P.2d 633, 657 (Cal. 1994)); *see also* Cramer v. Consol. Freightways, Inc., 209 F.3d 1122, 1131 (9th Cir. 2000) ("[P]rivacy rights can be altered or waived under California law and must be considered in context . . . ."), *amended en banc*, 255 F.3d 683 (9th Cir. 2001); Sporer v. UAL Corp., C 08-02835 JSW, 2009 WL 2761329, at *5 (N.D. Cal. Aug. 27, 2009) ("[H]aving advance notice that a company monitors computer use for compliance with the company's policies . . . and having an opportunity to consent to such monitoring,

law source, absolute core protections of employee privacy currently only arise from statutes, some of which establish very narrowly drafted prohibitions against monitoring that cannot be destroyed through unilateral notices.

3. Statutory Rights to Privacy

The United States has an amalgam of privacy statutes, enacted at different times, targeted for different purposes, and applicable to different entities.[61] Due to concerns arising from the growth of computer databases in the 1960s and 1970s,[62] Congress passed the Privacy Act of 1974, which regulates the collection and use of records by federal agencies.[63] The Act applies only to federal agencies, not to state or local agencies, nor to the private sector;[64] as such, its potential work-related application is limited to federal employers.[65] Most of the remaining federal privacy-related laws apply

further diminishes any reasonable expectation of privacy.") (citing TBG Ins. Servs. Corp. v. Superior Court, 117 Cal. Rptr. 2d 155, 163–64 (Ct. App. 2002)); Hernandez v. Hillsides, Inc., 211 P.3d 1063, 1077 (Cal. 2009) ("[N]otice of and consent to an impending intrusion can 'inhibit reasonable expectations of privacy . . . .' ") (quoting *Hill*, 865 P.2d at 655); *Hill*, 865 P.2d at 655 ("[E]ven when a legally cognizable privacy interest is present, other factors may affect a person's reasonable expectation of privacy[,] . . . [f]or example, advance notice of an impending action may serve to 'limit [an] intrusion upon personal dignity and security' that would otherwise be regarded as serious . . . ."); *TBG*, 117 Cal. Rptr. 2d at 160 ("Assuming the existence of a legally cognizable privacy interest, the extent of that interest is not independent of the circumstances, and other factors (including advance notice) may affect a person's reasonable expectation of privacy.") (citing *Hill*, 865 P.2d at 655).

61. *See, e.g.*, PAUL M. SCHWARTZ & DANIEL J. SOLOVE, INFORMATION PRIVACY: STATUTES AND REGULATIONS (2010–2011) (reproducing thirty-eight state and federal statutes addressing some element of information privacy).

62. Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137, 1164–65 (2002).

63. 5 U.S.C. § 552a (2006). Although the Privacy Act gives individuals the right to access and correct information about themselves held by federal agencies, *id.* § 552a(d), and restricts the use of information by federal agencies only for relevant and necessary purposes, *id.* § 552a(e), in reality, it provides only minimal privacy protection for individuals. For example, information held by federal agencies may be disclosed to law enforcement entities and consumer reporting agencies, *id.* § 552a(b)(7), (12), as well as for any routine use that is compatible with the purpose for which the agency collected the information, *id.* § 552a(b)(3). This "routine use exception" has been described as a significant loophole which has done little to prevent disclosure of personal information. Solove, *supra* note 62, at 1167–68.

64. SCHWARTZ & SOLOVE, *supra* note 61, at 133.

65. Application of the Privacy Act would be limited to employment-related records of federal employees and potentially employees of federal contractors. The Supreme Court has granted certiorari to address the issues of whether the government violates a federal contract employee's constitutional right to informational privacy when: (1) it asks in the course of a background investigation whether the employee has received counseling or treatment for illegal drug use that has occurred within the past year, and the employee's response is used only for employment purposes and is protected under the Privacy Act; or (2) it asks the employee's designated references for any adverse information that may have a bearing on the

only to specific entities and specific types of information collection.[66] The vast majority of federal privacy statutes apply to records, not necessarily to searches, surveillance, or intrusions;[67] as such, they do not apply to employment relationships.

At the state and federal levels, there are statutes that nominally address workplace communications privacy. For example, two states, Connecticut and Delaware, have statutes regulating employer monitoring of employee communications and actions, requiring the employers to first provide notice to employees of such monitoring.[68] In addition, nine states have statutes that prohibit recording communications without the consent of all parties to the conversation.[69] In practice, two-way consent requirements cannot have much impact on workplace-internal communications and activities. For instance, California's statute specifically exempts communications "in which the parties to the communication may reasonably expect that the communication

employee's suitability for employment at a federal facility, the reference's response is used only for employment purposes, and the information obtained is protected under the Privacy Act. NASA v. Nelson, 130 S. Ct. 1755, 1755 (2010).

66. For example: the Gramm-Leach-Bliley Act, 15 U.S.C. § 6802 (2006), limits information sharing by financial institutions with third parties without prior consent by customers; the Privacy Protection Act, 42 U.S.C. § 2000aa (2006), restricts the search or seizures of work product materials in the possession of third parties by government officers; the Cable Communications Policy Act, 47 U.S.C. § 551 (2006), requires notice to cable customers of any disclosure of personal information; the Video Privacy Protection Act, 18 U.S.C. § 2710 (2006), prohibits video rental stores from disclosing customer video rental and purchase information; and the Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered sections of titles 18, 26, 29, and 42 of the U.S. Code), regulates the disclosure of health information.

67. As for protecting records, the protections of the federal statutes are generally limited to when the records are in the "hands of third parties." Daniel J. Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, 75 S. CAL. L. REV. 1083, 1148 (2002). In other words, "the statutory regime does not protect records based on the type of information contained in the records, but protects them based on the particular types of third parties that possess them." *Id.*

68. CONN. GEN. STAT. § 31-48d (2011); DEL. CODE ANN. tit. 19, § 705 (2010). Three states have recently introduced similar legislation including: Massachusetts (H.R. 1862, 186th Sess. (Mass. 2009)), New York (A3871-A S4755 (N.Y. 2009), and Pennsylvania (S.B. 363 (Pa. 2009)).

69. *See* California: CAL. PENAL CODE § 632(a) (Deering 2010); Connecticut: CONN. GEN. STAT. § 52-570d(a) (2011); Florida: FLA. STAT. § 934.03(2)(d), (3)(d) (2010); Illinois: 720 ILL. COMP. STAT. 5/14-2(a)(1) (2006); Maryland: MD. CODE ANN., CTS. & JUD. PROC. § 10-402(c)(3) (2011); Massachusetts: MASS. ANN. LAWS ch. 272, § 99(b)(4), (c)(1) (LexisNexis 2010); New Hampshire: N.H. REV. STAT. ANN. § 570-A:2(1-a) (2010); Pennsylvania: 18 PA. CONS. STAT. § 5704(4) (2010); Washington: WASH. REV. CODE § 9.73.030(1)(a) (2011).

may be overheard or recorded."[70] The statute would therefore not apply if employees had been provided notice that their work-related communications are subject to recording. Furthermore, the employer can make an employee's consent to recording a condition of continued employment. The situation is different with respect to monitoring of communications between employees and external parties (such as customers, distributors, suppliers, and personal contacts of employees). Employers may find it more challenging to rule out limited, reasonable expectations of privacy for such external parties, or to obtain consent from the same.[71]

The Federal Electronic Communications Privacy Act (ECPA)[72] can also apply to work-related monitoring. However, as discussed in Section II.B.3.a, *infra*, most of the ECPA's requirements are satisfied with one party's consent (so employers can marginalize its impact by providing notice to their employees) and its application has been somewhat challenging for the courts.

a) The Electronic Communications Privacy Act

In 1968, Congress enacted Title III of the Omnibus Crime Control and Safe Streets Act of 1968,[73] which became generally known as the "Wiretap

70. CAL. PENAL CODE § 632(c).

71. Increasingly, enterprises include notices about e-mail filtering and communications monitoring in outbound e-mail footers that are automatically included in all communications through company networks; however, such notices may not reach outsiders in advance of initial contact or at all with respect to communications through channels outside the control of the employer, such as webmail, instant messenger, and text messaging. *See* Lothar Determann & Lars Brauer, *Employee Monitoring Technologies and Data Privacy—No One-Size-Fits-All Globally*, 9 IAPP PRIVACY ADVISOR 1, 4 (2009) ("But it is more difficult to inform third-party Web sites or e-mail and text message recipients of monitoring practices, let alone ask for upfront consent (as the first message presumably is subject to the monitoring)."); Lothar Determann, *When No Really Means No: Consent Requirements for Workplace Monitoring in the U.S.*, 3 WORLD DATA PROTECTION REP. 22, at 2 (2003) ("Some employers implement recordings informing callers that 'all calls can be monitored [] for quality assurance' and some ask employees to include monitoring notices in their e-mail signatures, but such notices cannot reach all third parties, especially not in the arena of first-time electronic communications.").

72. Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, tit. I, 100 Stat. 1848, 1848–59 (codified as amended at 18 U.S.C. §§ 2510–2522 (2006)); tit. II, 100 Stat. at 1860–68 (codified at 18 U.S.C. §§ 2701–2711 (2006)); tit. III, 100 Stat. at 1868–73 (codified at 18 U.S.C. §§ 3121–3127 (2006)).

73. Omnibus Crime Control and Safe Streets Act of 1968, Pub. L. No. 90-351, tit. III, § 802, 82 Stat. 197, 212 (codified as amended at 18 U.S.C. §§ 2510–2520 (2006)). Title III not only prohibited general wiretapping and electronic eavesdropping but also established requirements for state and federal officials to obtain wiretapping and eavesdropping warrants.

Act."[74] While the Wiretap Act was "the primary law protecting the security and privacy of business and personal communications in the United States," it soon became "hopelessly out of date."[75] The Wiretap Act only proscribed unauthorized aural interception of wire or oral communications—it only applied where the contents of a communication could be overheard and understood by the human ear.[76] In addition, it applied only to interceptions of communications sent via common carriers.[77] By the mid-1980s, e-mail, computer-to-computer data transmissions, cellular and cordless phones, and video conferencing were becoming commonplace; telephone calls were being transmitted by wire, microwave, and fiber optics, often in the form of digitized voice, data, and video. Additionally, many different companies, not just common carriers, were offering telephone and communications services.[78] Not only were the technological means of communication advancing, but so too were the surveillance devices and techniques to monitor such communications.[79] A 1985 Office of Technology Assessment report concluded that existing protections against telephone[80] and e-mail surveillance were "weak, ambiguous, or nonexistent."[81]

In 1986, recognizing that technology was surpassing the protections afforded by the Wiretap Act,[82] Congress recast the Wiretap Act as the Electronic Communications Privacy Act.[83] Title I of the ECPA addresses the interception of wire, oral, and electronic communications; Title II addresses

74. *See, e.g.*, Fraser v. Nationwide Mut. Ins. Co., 352 F.3d 107, 113 n.7 (3d Cir. 2003) ("The Wiretap Act was formally known as the 1968 Omnibus Crime Control and Safe Streets Act . . . . [I]t was superseded by the ECPA.").

75. S. REP. NO. 99-541, at 2 (1986) (internal quotations omitted).

76. *Id.* (citing United States v. N.Y. Tel. Co., 434 U.S. 159, 167 (1977)).

77. *Id.* (citing 18 U.S.C. § 2510(10) (1968)).

78. *Id.* at 2–3.

79. *Id.* at 3.

80. OFFICE OF TECH. ASSESSMENT, U.S. CONGRESS, FEDERAL GOVERNMENT INFORMATION TECHNOLOGY: ELECTRONIC SURVEILLANCE AND CIVIL LIBERTIES 29, 30 (1985), *available at* http://www.fas.org/ota/reports/8509.pdf.

81. *Id.* at 45.

82. *Electronic Communications Privacy Act: Hearings on H.R. 3378 Before the Subcomm. on Courts, Civil Liberties, and the Admin. of Justice of the H. Comm. on the Judiciary*, 99th Cong. 1–2 (1986) (statement of Rep. Robert Kastenmeier).

83. *See* statutes cited *supra* note 72; *see also* GINA STEVENS & CHARLES DOYLE, CONGRESSIONAL RESEARCH SERV., PRIVACY: AN OVERVIEW OF FEDERAL STATUTES GOVERNING WIRETAPPING AND ELECTRONIC EAVESDROPPING 6–7 (2008), *available at* http://digital.library.unt.edu/ark:/67531/metacrs10538/m1/1/high_res_d/98-326_2008Sep02.pdf.

access to stored wire and electronic communications and transactional records; and Title III addresses pen registers and trap and trace devices.[84]

Title I of the ECPA, still generally referred to as the Wiretap Act, makes punishable the intentional: (1) interception, or attempted interception, of "any wire, oral, or electronic communication;"[85] (2) use, or attempted use, of "any electronic, mechanical, or other device to intercept any oral communication;"[86] (3) disclosure, or attempted disclosure, "to any other person the contents of any wire, oral, or electronic communication, knowing or having reason to know that the information was obtained through the interception of a wire, oral, or electronic communication in violation of [the ECPA];"[87] or (4) the use, or attempted use, of "the contents of any wire, oral, or electronic communication, knowing or having reason to know that the information was obtained through the interception of a wire, oral, or electronic communication in violation of [the ECPA]."[88] Violations of the

84. S. REP. NO. 99-541, at 2–3. Titles I and II of the ECPA are discussed in more detail *infra*. A pen register device records outgoing address or routing information regarding a communication, *see generally* 18 U.S.C. § 3127(3) (2006 & Supp. 2009), while a trap and trace device records incoming address or routing source-identifying information, *see generally* 18 U.S.C. § 3127(4) (2006).

85. *See* 18 U.S.C. § 2511(1)(a) (2006 & Supp. 2008).

86. *See id.* § 2511(1)(b).

87. *See id.* § 2511(1)(c).

88. *See id.* § 2511(1)(d). The ECPA also prohibits the intentional disclosure, or attempted disclosure, to any other person the contents of any wire, oral, or electronic communication, intercepted by authorized means, where (1) there is knowledge that the communication was intercepted in connection with a criminal investigation; (2) the information was obtained or received in connection with a criminal investigation; and (3) there is intent to improperly obstruct, impede, or interfere with a duly authorized criminal investigation. *See id.* § 2511(1)(e). The ECPA defines "wire communication" as:

> [A]ny aural transfer made in whole or in part through the use of facilities for the transmission of communications by the aid of wire, cable, or other like connection between the point of origin and the point of reception (including the use of such connection in a switching station) furnished or operated by any person engaged in providing or operating such facilities for the transmission of interstate or foreign communications or communications affecting interstate or foreign commerce[.]

Id. § 2510(1). "Oral communication" is defined as "any oral communication uttered by a person exhibiting an expectation that such communication is not subject to interception under circumstances justifying such expectation, but such term does not include any electronic communication[.]" *Id.* § 2510(2). "Electronic communication" is defined as "any transfer of signs, signals, writing, images, sounds, data, or intelligence of any nature transmitted in whole or in part by a wire, radio, electromagnetic, photoelectronic or photooptical system that affects interstate or foreign commerce, but does not include . . . any wire or oral communication." *Id.* § 2510(12). As such, "a communication is an electronic communication protected by the federal wiretap law if it is not carried by sound waves and

ECPA carry criminal penalties.[89] In addition, "any person whose wire, oral, or electronic communication is intercepted, disclosed, or intentionally used in violation of [the ECPA]" may bring a civil action for relief, including equitable relief, money damages, and attorney's fees.[90]

Title II of the ECPA, the Stored Communications Act (SCA), makes it unlawful to access stored communications. The ECPA defines electronic storage as "any temporary, intermediate storage of a wire or electronic communication incidental to the electronic transmission thereof; and . . . any storage of such communication by an electronic communication service for purposes of backup protection of such communication."[91] The SCA prohibits, with the threat of fines and imprisonment: "(1) intentionally access[ing] without authorization a facility through which an electronic communication service is provided; or (2) intentionally exceed[ing] an authorization to access that facility; and thereby obtain[ing], alter[ing], or prevent[ing] authorized access to a wire or electronic communication while it is in electronic storage in such system . . . ."[92] Similar to the Wiretap Act, the SCA provides civil remedies for anyone aggrieved by a violation of the Act.[93]

The purpose of the SCA is to address "the growing problem of unauthorized persons deliberately gaining access to, and sometimes tampering with, electronic or wire communications that are not intended to be available to the public."[94] For example, while there is no violation of the SCA when a subscriber accesses her own stored e-mail messages, "[a]ccessing the storage of other subscribers without specific authorization to do so would be a violation . . . ."[95] "Similarly, a member of the general public authorized to access the public portion of a computer facility would violate . . . [the SCA] by intentionally exceeding that authorization and accessing the private portions of the facility."[96]

While the goals of the ECPA appear to be quite straightforward, applying the ECPA has been wrought with difficulty, particularly for alleged violations arising from the workplace. Almost from its inception, the language used within the ECPA has been subject to highly technical parsing to determine

cannot fairly be characterized as containing the human voice." S. REP. NO. 99-541, at 14. "This term also includes electronic mail . . . ." *Id.*

89. *See* 18 U.S.C. § 2511(4)(a).
90. *See id.* § 2520(a)–(b).
91. *Id.* § 2510(17).
92. *Id.* § 2701(a).
93. *Id.* § 2707.
94. S. REP. NO. 99-541, at 35 (1986).
95. *Id.* at 36.
96. *Id.*

the Act's application in the workplace. It has not been regarded as a model of statutory clarity.[97] Part of the difficulty with the ECPA has been the interplay between Title I (Wiretap Act), which prohibits the interception of wire, oral, and electronic communications, and Title II (SCA), which protects stored wire and electronic communications and transaction records.[98]

Ironically, although the principal motivation to update the 1968 Omnibus Crime Control and Safe Streets Act with the 1986 ECPA was because statutory protections against electronic eavesdropping had, as discussed *supra*, become out of date, the ECPA itself was quickly found to be behind the technological curve.[99] For example, in *Konop v. Hawaiian Airlines, Inc.*, when the Ninth Circuit Court of Appeals addressed whether an employer's executive had accessed an employee's (Konop's) private website without authorization in violation of the SCA, the court noted: "[T]he ECPA was written prior to the advent of the Internet and the World Wide Web. As a result, the existing statutory framework is ill-suited to address modern forms of communication like Konop's secure website."[100]

Konop had restricted access to his website to only pre-approved individuals, mostly pilots and other employees of Hawaiian Airlines.[101] A Hawaiian Airlines vice president asked for and received access information for Konop's website from two pilots Konop had pre-approved.[102] The vice president then used that information to access and read Konop's website.[103] "Section 2701(c)(2) of the SCA allows a person to authorize a third party's access to an electronic communication if the person is (1) a user of the

97. *See, e.g.*, Steve Jackson Games, Inc. v. United States, 36 F.3d 457, 462 (5th Cir. 1994) (referring to the Wiretap Act as "famous (if not infamous) for its lack of clarity"); Forsyth v. Barr, 19 F.3d 1527, 1542–43 (5th Cir. 1994) ("[C]onstruction of the Wiretap Act is fraught with trip wires.").

98. *See, e.g.*, United States v. Smith, 155 F.3d 1051, 1055 (9th Cir. 1998) ("[T]he intersection of the Wiretap Act and the Stored Communications Act is a complex, often convoluted, area of the law.") (citations omitted).

99. *See, e.g.*, Jeremy U. Blackowicz, Note, *E-mail Disclosure to Third Parties in the Private Sector Workplace*, 7 B.U. J. SCI. & TECH. L. 80, 104 (2001) ("Commentators are practically unanimous in calling for statutory solutions in the form of both amendments and revisions to the ECPA or a new statutory scheme to give employees some form of protection.") (citation omitted); Lee Nolan Jacobs, *Is What's Yours Really Mine?:* Shmueli v. Corcoran Group *and Penumbral Property Rights*, 14 J.L. & POL'Y 837, 876 (2006) ("With the continued evolution of technology, any protections afforded by the ECPA have become practically irrelevant.") (citation omitted).

100. 302 F.3d 868, 874 (9th Cir. 2002).

101. *See id.* at 872.

102. *See id.* at 873.

103. *See id.*

service and (2) the communication is of or intended for that user."[104] The *Konop* court noted "some indication in the legislative history that Congress believed addressees or intended recipients of electronic communications would have the authority under the SCA to allow third parties access to those communications."[105] Therefore, the Hawaiian Airlines executive would not have violated the SCA if he gained access to Konop's website by using information obtained from authorized users. However, the individuals from whom the executive obtained the access information had never actually *used* Konop's website; therefore they were never "users" under the language of the SCA. The *Konop* court relied on this technicality and perhaps overly-literal interpretation of the statute to reverse the district court's grant of summary judgment dismissing Konop's SCA claim.[106]

Exceptions within the ECPA also render much of the Act inapplicable to ordinary uses of computer and communications systems within the workplace. Section 2511(2)(a)(i) exempts officers, employees, and agents of a wire or communications service provider from liability for intercepting, disclosing, or using communications transmitted over the service in the ordinary course of business.[107] Section 2511(2)(d) exempts from liability anyone who intercepts a communication who is a party to the communication, or where one of the parties has consented to interception.[108] Based on the language of these two sections, "employers who own and provide their own e-mail [and communications] systems are exempt from the ECPA's requirements."[109] Employers who outsource their e-mail and

104. *Id.* at 880 (quotations and citation omitted); *see also* 18 U.S.C. § 2701(c)(2) (2006).

105. *Konop*, 302 F.3d at 880 (citing H.R. REP. NO. 99-647, at 66–67 (1986)).

106. *See id.* The *Konop* court upheld the lower court's dismissal of Konop's Title I Wiretap Act claims, holding that in order to "intercept" the content of Konop's website, it would have to be acquired during transmission, not while in electronic storage, *id.* at 878, presumably in transmission from Konop's computer to the storage location of the website content versus in transmission from the storage location to someone else's computer. *See* Pietrylo v. Hillstone Rest. Group, No. 06-5754 (FSH), 2009 WL 3128420 (D.N.J. Sept. 25, 2009) (upholding a jury's verdict that an employer had violated the SCA by accessing without authorization an invitation-only private MySpace chat group maintained by an employee).

107. 18 U.S.C. § 2511(2)(a)(i) (2006).

108. *Id.* § 2511(2)(d). *See also* Sporer v. UAL Corp., No. C 08-02835 JSW, 2009 WL 2761329, at *5–6 (N.D. Cal. Aug. 27, 2009) (holding the employer's monitoring of employees' e-mail messages did not violate § 2511 because employees impliedly consented to monitoring by consenting to the employer's monitoring policy). *But see* Watkins v. L.M. Berry & Co., 704 F.2d 577, 581 (11th Cir. 1983) ("[K]nowledge of the *capability* of monitoring alone cannot be considered implied consent . . . .") (alteration in original) (citation omitted) (applying pre-ECPA § 2511).

109. Lisa Smith-Butler, *Workplace Privacy: We'll Be Watching You*, 35 OHIO N.U. L. REV. 53, 67 (2009).

communications systems to service providers can also rely on the exception when they work with their service provider to intercept employee communications.[110]

Although the SCA does not specifically reference e-mail,[111] as noted, *supra*, Congress clearly intended the SCA to protect against unauthorized access of e-mail messages. The SCA prohibits unauthorized access of communications while in electronic storage;[112] however, similar to § 2511(2)(a)(i) of the Wiretap Act, § 2701(c) of the SCA exempts from liability providers of the wire or electronic storage.[113] Courts have applied § 2701(c) to employers, holding they are exempt from liability under the SCA for accessing employee e-mail messages stored on their computer systems.[114] Because of the exemptions contained in both the Wiretap Act and the SCA, commentators are in general agreement that the ECPA is ineffective in providing employees with any privacy protections relative to work-related e-mail messages and other forms of wire and electronic communications.[115]

C. INTRUSIVE WORKPLACE MONITORING AND EMPLOYEE PRIVACY

Privacy rights afforded electronic communications have a complex history that provides little guidance as technologies evolve. The surreptitious "listening" to other people's conversations has evolved from literally standing outside a home to overhear conversations,[116] to tapping phone lines

110. Leonard Court & Courtney Warmington, *The Workplace Privacy Myth: Why Electronic Monitoring Is Here To Stay*, 29 OKLA. CITY U. L. REV. 15, 28–30 (2004).

111. *See id.* at 26.

112. 18 U.S.C. § 2701(a). *See also id.* § 2510(17) (defining "electronic storage" under the ECPA).

113. *Id.* § 2701(c).

114. *See* Fraser v. Nationwide Mut. Ins. Co., 352 F.3d 107, 114–15 (3rd Cir. 2003).

115. *See, e.g.*, Jay P. Kesan, *Cyber-Working or Cyber-Shirking?: A First Principles Examination of Electronic Privacy in the Workplace*, 54 FLA. L. REV. 289, 299 (2002) ("[T]he ECPA is ineffective in regulating the employer/employee relationship.") (citation omitted); Porter II & Griffaton, *supra* note 9, at 66 (concluding the ECPA "provides employees little protection from the monitoring of their workplace electronic communications"); Lawrence E. Rothstein, *Privacy or Dignity?: Electronic Monitoring in the Workplace*, 19 N.Y.L. SCH. J. INT'L & COMP. L. 379, 401 (1999) (concluding the ECPA "has generally proven ineffective in protecting employees in the workplace from their employers' monitoring") (citation omitted). *See also* Ariana R. Levinson, *Carpe Diem: Privacy Protection in Employment Act*, 43 AKRON L. REV. 331, 340 n.37 (2010) (summarizing commentators who have criticized application of the ECPA in the employment context).

116. "Eaves-droppers, or such as listen under walls or windows or the eaves of a house, to hearken after discourse, and thereupon to frame slanderous and mischievous tales, are a common nuisance and presentable at . . . [court]." 4 SIR W.M. BLACKSTONE, COMMENTARIES ON THE LAWS OF ENGLAND 168–69 (Thomas B. Wait & Co. 1807).

to record conversations,[117] to today's incarnation of reviewing the keystrokes one types on a computer keyboard, revealing, along with passwords and the addresses of websites visited, the content of messages composed and sent to others.[118]

New instant photography and audio recording technologies prompted Warren and Brandeis in 1890 to call for a right "to be let alone."[119] In his later dissent in *Olmstead v. United States*, in which the Supreme Court ruled a warrantless wiretap of a telephone conversation did not violate the Fourth Amendment, Justice Brandeis warned that "in the application of a constitution, our contemplation cannot be only of what has been but of what may be."[120] The majority in *Katz v. United States* chose to re-evaluate the notion of eavesdropping in light of, at that time, "the vital role that the public telephone has come to play in private communication."[121]

In his dissent in *Katz*, Justice Black argued that changes in technology should not expand the reach of the Fourth Amendment.[122] This concern

117. *See, e.g.*, Olmstead v. United States, 277 U.S. 438, 456–57 (1928) (describing wire tapping as intercepting messages on telephones by inserting small wires along ordinary telephone wires); *see also* ROBERT ELLIS SMITH, BEN FRANKLIN'S WEB SITE: PRIVACY AND CURIOSITY FROM PLYMOUTH ROCK TO THE INTERNET 156 (2000) (describing one of the earliest known surreptitious recordings of a conversation, in 1895, in which a postal inspector hid a recording device in his top hat to successfully record the words of a lawyer suspected of the illegal use of the mail).

118. *See, e.g.*, Bailey v. Bailey, No. 07-11672, 2008 WL 324156, at *1 (E.D. Mich. Feb. 6, 2008) (describing a key logger as a program designed to record every keystroke made on the computer and store it in a text file on the computer's hard drive); United States v. Ropp, 347 F. Supp. 2d 831, 831 (C.D. Cal. 2004) (involving a key logger program that "recorded and stored the electronic impulses traveling down the cable between [the user's] keyboard and the computer to which it was attached").

119. Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195 (1890) (citing THOMAS M. COOLEY, A TREATISE ON THE LAW OF TORTS OR THE WRONGS WHICH ARISE INDEPENDENT OF CONTRACT 29 (1880) ("The right to one's person may be said to be a right of complete immunity: to be let alone.")). "Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that 'what is whispered in the closet shall be proclaimed from the house-tops.' " *Id.*

120. 277 U.S. at 474 (Brandeis, J., dissenting) (internal quotation marks omitted). "Ways may some day be developed by which the Government, without removing papers from secret drawers, can reproduce them in court, and by which it will be enabled to expose to a jury the most intimate occurrences of the home." *Id.*

121. 389 U.S. 347, 352 (1967).

122. Justice Black stated:

> Tapping telephone wires, of course, was an unknown possibility at the time the Fourth Amendment was adopted. But eavesdropping (and wiretapping is nothing more than eavesdropping by telephone) was . . . an ancient practice which at common law was condemned as a nuisance . . . . There can be no doubt that the Framers were aware of this

forms the crux of the uncertainty over the extent to which evolving technologies can invade one's privacy. In *Kyllo v. United States*, the Supreme Court addressed the extent to which new technologies should "shrink the realm of guaranteed privacy."[123] In *Kyllo*, the Supreme Court ruled that thermal imaging technology, used without a warrant to measure the heat emanating from a home and hence indicating whether marijuana was being grown inside, constituted a Fourth Amendment search requiring a warrant.[124] The Court reasoned that the warrant was required because information was gleaned through the use of technology that otherwise could not have been obtained without a physical "intrusion into a constitutionally protected area . . . ."[125]

But the Supreme Court has been quick to back away from an expansive application of *Kyllo*. In *Illinois v. Caballes*, the Supreme Court held contraband detected by a drug sniffing dog during a routine traffic stop did not fall within the realm of *Kyllo*.[126] The Court reasoned that the thermal imaging device at issue in *Kyllo* was able to detect lawful activity, particularly intimate details in a home, whereas "[a] dog sniff conducted during a concededly lawful traffic stop that reveals no information other than the location of a substance that no individual has any right to possess does not violate the Fourth Amendment."[127] And most recently in *Quon*, the Supreme Court expressed a cautious approach vis-à-vis technology: "The judiciary risks error by elaborating too fully on the Fourth Amendment implications of emerging technology before its role in society has become clear."[128] In fact, the latter part of that statement reflects an important qualification in *Kyllo*'s holding: the Supreme Court held the use of thermal imaging technology constituted a Fourth Amendment search—"at least where . . . the technology in question is not in general public use."[129] As a result, as Justice Stevens noted in his *Kyllo*

> practice, and if they had desired to outlaw or restrict the use of evidence obtained by eavesdropping, I believe that they would have used the appropriate language to do so in the Fourth Amendment.

Id. at 366 (Black, J., dissenting) (citations and internal quotation marks omitted).

123. 533 U.S. 27, 34 (2001).

124. *See id.*

125. *Id.* (citation and internal quotation marks omitted).

126. 543 U.S. 405 (2005).

127. *Id.* at 409–10.

128. 130 S. Ct. 2619, 2629 (2010). "A broad holding concerning employees' privacy expectations vis-à-vis employer-provided technological equipment might have implications for future cases that cannot be predicted." *Id.* at 2630.

129. *Kyllo*, 533 U.S. at 34.

dissent, "the threat to privacy will grow, rather than recede, as the use of intrusive equipment becomes more readily available."[130]

The role of emerging technology is important because, ultimately, the right to privacy is dependent upon an individual's expectation of privacy. An individual's expectation of privacy must be both actual (subjective) and one that society is willing to accept as reasonable.[131] If and to the extent an individual employee can substantiate an actual expectation of privacy, the following question as to the reasonableness of such expectation—and hence the existence and scope of the employee's rights—depends upon societal norms.[132] The "reasonableness" assessment is contextual: it depends upon the circumstances of any particular event. It must therefore be determined on a case-by-case basis.[133] There is "no talisman that determines in all cases those privacy expectations that society is prepared to accept as reasonable."[134] As such, guidance on when actionable workplace privacy exists, or does not exist, can only arise from an examination of decisions exploring the various circumstances in which an employee claims an invasion of privacy by an employer. Before turning to the parameters of such examination, it is important to remember that such examination becomes relevant only where the employee can substantiate an *actual* expectation of privacy. In practice, it is largely up to the employer whether employees are allowed to nurture such an actual expectation of privacy. Employers can—and often do—destroy any actual expectation of privacy by notifying employees in painstaking detail about the existence and intrusiveness of monitoring and surveillance

130. *Id.* at 47 (Stevens, J., dissenting). In his majority opinion, Justice Scalia differentiated *Kyllo* from *California v. Ciraolo*, 476 U.S. 207 (1986), on the basis that discovering marijuana plants from an airplane flyover was not a Fourth Amendment search because such flights were routine, whereas use of thermal imaging technology was not routine. *Kyllo*, 533 U.S. at 39 n.6.

131. Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., concurring); *see also* TBG Ins. Servs. Corp. v. Superior Court, 117 Cal. Rptr. 2d 155, 160 (Ct. App. 2002) ("When affirmative relief is sought to prevent a constitutionally prohibited invasion of privacy, the plaintiff must establish (1) a legally protected privacy interest; (2) a reasonable expectation of privacy in the circumstances; and (3) conduct by defendant constituting a serious invasion of privacy.") (citation and internal quotation marks omitted) (applying CAL. CONST. art. I, § 1).

132. *See* Oliver v. United States, 466 U.S. 170, 178 n.8 (1984), *cited with approval in* O'Connor v. Ortega, 480 U.S. 709, 715 (1987).

133. This is a clearly enunciated approach for public-sector employees. *See, e.g.*, *O'Connor*, 480 U.S. at 718 ("[T]he question whether an employee has a reasonable expectation of privacy must be addressed on a case-by-case basis."). While not stated so directly, courts do perform a case-by-case analysis to determine whether a private-sector employee has a reasonable expectation of privacy. *See, e.g.*, Thygeson v. U.S. Bancorp, No. CV-03-467-ST, 2004 WL 2066746, at *17–22 (D. Or. Sept. 15, 2004).

134. *O'Connor*, 480 U.S. at 715.

technologies deployed. Yet, occasionally employers find it difficult to keep up with technological progress, and notices become outdated. This allows for the growth of limited expectations of privacy in communication methods not covered by outdated employer notices. Also, during periods of economic growth or in industries with limited access to talent, employees gain market power, forcing employers to try harder to remain attractive to employees. In such circumstances, employers tend to keep notices and policies friendlier to employees. Thus, when notices become outdated or employees gain market power, actual expectations of privacy may develop and raise the question of whether they are "reasonable" in the face of deployments of intrusive monitoring technologies.

As a general matter, courts believe that in most circumstances, employees do not have a reasonable expectation of privacy in e-mail messages sent or received over their employer's computer systems.[135] Courts have also been reluctant to find a reasonable expectation of privacy for personal use of an employer-provided computer.[136] In particular, courts have taken the

135. *See, e.g.*, Sporer v. UAL Corp., No. C 08-02835 JSW, 2009 WL 2761329, at *6–7 (N.D. Cal. Aug. 27, 2009) (holding an employee had no expectation of privacy in e-mail messages transmitted on work computer because he was aware the employer would monitor such messages; noting the employee was fired after receiving an e-mail message containing a pornographic video from a non-employee friend and then forwarding the message to his personal e-mail account); Smyth v. Pillsbury Co., 914 F. Supp. 97, 101 (E.D. Pa. 1996) (holding an employee could not have "a reasonable expectation of privacy in e-mail communications voluntarily made by an employee to his supervisor over the company e-mail system notwithstanding any assurances that such communications would not be intercepted by management") (applying Pennsylvania law); McLaren v. Microsoft Corp., No. 05-97-00824-CV, 1999 WL 339015, at *4–5 (Tex. App. May 28, 1999) (holding an employee had no reasonable expectation of privacy in e-mail messages stored in a password-protected personal folder located on the employee's work computer). Indeed, some courts believe there is no appropriate expectation of privacy in e-mail messages once they have been sent to and received by a third party. *See, e.g.*, Rehberg v. Paulk, 598 F.3d 1268, 1281–82 (11th Cir. 2010); United States v. Lifshitz, 369 F.3d 173, 190 (2d Cir. 2004); Guest v. Leis, 255 F.3d 325, 333 (6th Cir. 2001). *But see* United States v. Forrester, 512 F.3d 500, 511 (9th Cir. 2008) (stating the contents of an e-mail message, like that of a letter, may deserve Fourth Amendment protection because it is expected to be read only by the intended recipient).

136. *See, e.g.*, United States v. Barrows, 481 F.3d 1246, 1249 (10th Cir. 2007) (holding an employee had no expectation of privacy in any files observed by co-workers when that employee connected his personal computer, located in a public work area, to his employer's computer network which allowed file sharing, left the computer running, and did not password-protect any files); *TBG Ins. Servs. Corp.*, 117 Cal. Rptr. 2d at 163 (holding the employee had no reasonable expectation of privacy in personal files stored on an employer-provided computer because he was aware of the employer's computer use policy which stated the computer was not to be used for personal purposes and its content could be monitored at any time). *But see* Curto v. Med. World Commc'ns, Inc., No. 03CV6327 (DRH)(MLO), 2006 WL 1318387, at *5–6 (E.D.N.Y. May 15, 2006) (holding an employee had a reasonable expectation of privacy in personal e-mail messages and files stored on and

approach that since employers routinely monitor employee work-related communications and computer use,[137] "the use of computers in the employment context carries with it social norms that effectively diminish the employee's reasonable expectation of privacy with regard to his use of his employer's computers."[138]

Similar to the determination of a reasonable expectation of privacy, there is no bright-line test for what constitutes a highly offensive intrusion upon seclusion.[139] One conclusion the Supreme Court clearly reached in *Quon* is that public employers do not have to use the least intrusive means possible in order to conduct a permissible warrantless search under the Fourth Amendment.[140] While courts will generally not tolerate surveillance for "repugnant" or "socially unprotected" reasons,[141] they have used a case-by-case approach to determine what is a highly intrusive invasion of privacy.[142]

---

then deleted by the employee from an employer-provided laptop computer used by the employee solely at her home and which was never connected or used through the employer's computer system).

137. *See, e.g.*, discussion *supra* note 7.

138. *TBG Ins. Servs. Corp.*, 117 Cal. Rptr. 2d at 162.

139. *See, e.g.*, Turnbull v. Am. Broad. Cos., No. CV 03-3554 SJO (FMOX), 2004 WL 2924590, at *13 (C.D. Cal. Aug. 19, 2004) ("[A] court determining the existence of 'offensiveness' would consider the degree of intrusion, the context, conduct and circumstances surrounding the intrusion, the intruder's motives and objectives, the setting into which he intrudes, and the expectations of those whose privacy is invaded.") (citation omitted); Hernandez v. Hillsides, Inc., 211 P.3d 1063, 1073 (Cal. 2009) (noting California tort law contains no bright line on determining the offensiveness of an intrusion).

140. City of Ontario, Cal. v. Quon, 130 S. Ct. 2619, 2632 (2010) ("This Court has repeatedly refused to declare that only the least intrusive search practicable can be reasonable under the Fourth Amendment.") (citation and internal quotation marks omitted).

141. *See Hernandez*, 211 P.3d at 1080 (identifying blackmail, harassment, and prurient curiosity as repugnant and socially unprotected reasons for surveillance) (citing Shulman v. Grp. W Prods., Inc., 955 P.2d 469, 493 (Cal. 1998)).

142. *Compare id.* at 1073 *with* Nelson v. Salem State Coll., 845 N.E.2d 338 (Mass. 2006). In *Hernandez*, the employer secretly installed a hidden camera in a shared office space in an attempt to ascertain who was entering the office after hours to use a computer to access pornography. 211 P.3d at 1066. Although the California Supreme Court believed the employees had a reasonable expectation of privacy in their shared office space, *id.* at 1076, it ruled the employer's video monitoring was not overly intrusive because it was limited in scope and directly related to protecting the goals of the workplace (protecting abused children). *Id.* at 1082. In *Nelson*, the employer secretly installed a hidden camera in a shared office space based on concerns of after-hours unauthorized access to the work area. 845 N.E.2d at 343. The plaintiff sued her employer for invasion of privacy after she discovered she had been recorded by the video surveillance changing clothes in the office space. *Id.* at 341. Despite the fact that the camera was set to record twenty-four hours per day for the purpose of monitoring after-hours access, the plaintiff had locked the door to the office, and her activities were recorded just before and after regular business hours, the court ruled she

There is a close relationship between a reasonable expectation of privacy and the degree of permissible intrusiveness. As courts have shown, particularly in relation to workplace monitoring, where employees are aware the employer may intrude upon their privacy for legitimate business purposes, there can be no expectation of privacy.[143] Where the employer's monitoring goes beyond legitimate business purposes, however, and intrudes on what society may consider highly personal areas beyond the scope of work, then an actionable invasion of privacy may be found.[144] There are currently three areas where monitoring by employers may, according to social norms, be considered so personal as to constitute an inappropriate intrusion: access to personal web-based applications; use of webcams; and use of location-tracking technologies.

1. Employer Access to Personal Web-Based Applications

According to a recent survey on information technology policies in the workplace, over fifty percent of responding employees accessed personal web-based e-mail accounts from work using employer-provided computers, although only seventeen percent of the respondents said their companies permitted such conduct.[145] Courts have found that employees can have an expectation of privacy in e-mail messages stored on personal web-based e-mail services, even when they have accessed those services at work through

did not have a reasonable expectation of privacy within the office because other people had keys to the office and could have walked in on her at any time. *Id.* at 349.

143. *See, e.g.*, cases cited *supra* note 142.

144. *Compare* Phillips v. Smalley Maint. Servs., 711 F.2d 1524 (11th Cir. 1983) (finding an invasion of an employee's privacy based on the employer's repeated inquiries into the employee's sex life), Johnson v. K Mart Corp., 723 N.E.2d 1192, 1196–97 (Ill. App. Ct. 2000) (reversing summary judgment granted in favor of the defendant-employer as to plaintiffs' intrusion upon seclusion claims; holding the employer's investigation concerning workplace thefts, vandalism, and drug use went too deeply into personal lives of employees, beyond any business purpose), *and* Soroka v. Dayton Hudson Corp., 1 Cal. Rptr. 2d 77, 79, 86 (Ct. App. 1991) (holding the prospective employer violated applicants' privacy with a 704-question psychological test that asked questions pertaining to religious beliefs and sexual orientation, concluding these issues had no bearing on the requirements of the applied-for job), *with* Morenz v. Progressive Cas. Ins. Co., No. 79979, 2002 WL 1041760, at *2, *4 (Ohio Ct. App. May 23, 2002) (finding no invasion of privacy where the employer asked the employee if he was gay; concluding the purpose of the question, asked in private, was merely to ascertain the employee's job satisfaction and comfort living in the south). *See generally* Marisa Anne Pagnattaro, *What Do You Do When You Are Not at Work?: Limiting the Use of Off-Duty Conduct as the Basis for Adverse Employment Decisions*, 6 U. PA. J. LAB. & EMP. L. 625, 632–34 (2004) (discussing when employers' monitoring stays within or goes beyond the scope of legitimate business purpose).

145. PONEMON INSTITUTE, TRENDS IN INSIDER COMPLIANCE WITH DATA SECURITY POLICIES 7 (2009), http://www.ponemon.org/local/upload/fckjail/generalcontent/18/file/Trends%20in%20Insider%20Compliance%20with%20Policies%20Final%203.pdf.

employer-provided computer systems. For example, in *Pure Power Boot Camp, Inc. v. Warrior Fitness Boot Camp, LLC*, the court made three specific conclusions in finding an employee had a reasonable expectation of privacy in personal e-mail messages stored on a third party's service, although the employee had accessed that outside service while at work, using employer-provided equipment.[146] The *Pure Power Boot Camp* court found that: (1) an employer's access of personal e-mail messages from an employee's web-based e-mail service without authorization violates the SCA;[147] (2) an employer's computer use and e-mail policy which explicitly prohibited personal use of the Internet at work and provided notice that all e-mail messages could be monitored did not create an implied consent on the part of the employee that his personal e-mail messages stored with an outside service provider could be monitored, even though the employee had accessed the outside service provider at work using employer-provided equipment;[148] and (3) the fact that the employee's username and password to a personal web-based e-mail account were later automatically filled in on the employee's work computer because of the employee's earlier access to the account did not imply authorization for others to access the employee's account.[149]

Stengart v. Loving Care Agency, Inc. also involved an employee's correspondence with her attorney through personal e-mail messages stored on a third-party web-based system.[150] Rather than access the messages directly through the service, Stengart's employer accessed copies of her messages that had been automatically stored on her company-provided laptop computer.[151] Although the employer's computer use policy warned that e-mail messages "are not to be considered private or personal," the court noted the policy did not provide any express notice that messages sent or received on a personal, web-based e-mail account would be subject to monitoring if company equipment was used to access the account.[152] The court concluded that Stengart had a reasonable expectation of privacy in her personal e-mail correspondence with her attorney because of the steps she

146. 587 F. Supp. 2d 548 (S.D.N.Y. 2008).

147. *Id.* at 556.

148. *Id.* at 559. The e-mail messages were "personal" in that they were in no way related to the employer's business. *See id.* at 560.

149. *Id.* at 561 (providing the analogy that had the employee left a key to his house on his desk, that would not imply authorization for anyone else to use the key to "rummage" through his house).

150. 990 A.2d 650 (N.J. 2010).

151. *Id.* at 655–56.

152. *Id.* at 659.

took to protect the privacy of those messages. In particular, she used "a personal, password-protected e-mail account instead of her company e-mail address and did not save the account's password on her computer."[153]

Pietrylo v. Hillstone Restaurant Group, though not involving e-mail messages, is similar in facts to *Konop*.[154] In *Pietrylo*, a supervisor "coerced" an employee to provide access to information in a restricted communications area within a MySpace account in which employees were making comments critical of their employer and management.[155] The court ruled such access was unauthorized and in violation of the SCA.[156] These cases indicate a clear willingness on the part of the courts to consider e-mail and other types of electronic messages stored on personal web-based accounts to be within a zone in which employees have a reasonable expectation of privacy. Furthermore, an employer's invasion of this zone constitutes an actionable invasion of privacy.

2. Webcams

Webcams and cameras built into laptop computers are becoming ubiquitous in the workplace. Worldwide sales of webcams are predicted to increase from $1.2 billion in 2006 to $6.2 billion by 2013.[157] Employers are no longer only using webcams for video conferencing or virtual training; they are also beginning to use webcams for employee monitoring.[158] Webcams built into laptop computers raise the potential of intrusive employer monitoring beyond the physical bounds of "traditional" workplace video monitoring.

A recent survey of workplace monitoring reveals that nearly fifty percent of employers use video monitoring to counter theft, violence, and sabotage.[159] As a general matter, courts find no objection to video monitoring

153. *Id.* at 663. *See also In re* Asia Global Crossing, Ltd., 322 B.R. 247, 259–61 (Bankr. S.D.N.Y. 2005) (holding certain executives had an expectation of privacy in personal e-mail messages sent through the company's e-mail system because the company's computer use policy was equivocal regarding certain uses and monitoring); Nat'l Econ. Research Assocs., Inc. v. Evans, No. 04-2618 BLS2, 2006 WL 2440008, at *3–5 (Mass. Super. Ct. Aug. 3, 2006) (holding same on similar facts).

154. Pietrylo v. Hillstone Restaurant Grp., No. 06-5754 (FSH), 2009 WL 3128420 (D.N.J. Sept. 25, 2009); *see* discussion *supra* Section II.B.3.a. regarding Konop v. Hawaiian Airlines, Inc., 302 F.3d 868 (9th Cir. 2002).

155. *Pietrylo*, 2009 WL 3128420, at *3.

156. *Id.*

157. Michelle V. Rafter, *Smile, You're on the Company Webcam*, INC. TECH. (Mar. 1, 2008), http://technology.inc.com/hardware/Articles/200803/webcams.html.

158. *Id.*

159. AMA & EPOLICY INST., *supra* note 7, at 3 (reporting also that only seven percent of employers use video surveillance to track employees' on-the-job performance).

in open workplaces.[160] Where courts require more detailed and specific notices to negate a reasonable expectation of privacy on the employee's side is surveillance of areas in which employees tend to have a greater actual expectation of privacy: restrooms and dressing rooms.[161] But, even with respect to highly sensitive circumstances, courts have not acknowledged a core expectation of privacy that is protected against waivers, consents, and notices as a matter of public policy.[162] In some cases, state legislatures have stepped in and prohibited certain forms of surveillance outright. For example, section 435 of the California Labor Code prohibits "audio or video

160. For example, in *Vega-Rodriguez v. Puerto Rico Telephone Co.*, the court explained:

> [N]o legitimate expectation of privacy exists in objects exposed to plain view as long as the viewer's presence at the vantage point is lawful. And the mere fact that the observation is accomplished by a video camera rather than the naked eye, and recorded on film rather than in a supervisor's memory, does not transmogrify a constitutionally innocent act into a constitutionally forbidden one.

110 F.3d 174, 181 (1st Cir. 1997) (citation and footnote omitted); *see also* Acosta v. Scott Labor LLC, 377 F. Supp. 2d 647, 651 (N.D. Ill. 2005) (holding the use of hidden cameras in an open office setting does not automatically transform a non-private area into a private one).

161. *See, e.g.*, Williams v. City of Tulsa, Okla., 393 F. Supp. 2d 1124, 1137–38 (N.D. Okla. 2005), *aff'd*, 204 F. App'x 762 (10th Cir. 2006) (holding city could potentially be liable for violations of the Fourth Amendment and the ECPA as well as for intentional infliction of emotional distress for alleged surreptitious video monitoring of the restroom, but the plaintiffs did not present sufficient evidence it had occurred). The court had no issue with video cameras hidden in clocks which recorded activities in general work areas, *id.* at 1134, nor with audio recording equipment discovered in the air conditioner vent above one supervisor's office, *id.* at 1134–36. *See also* Rosario v. United States, 538 F. Supp. 2d 480 (D.P.R. 2008) (denying employer Department of Veterans Affairs' motion to dismiss employee's Fourth Amendment violation of privacy claims based on video surveillance of locker room); Trujillo v. City of Ontario, 428 F. Supp. 2d 1094 (C.D. Cal. 2006), *aff'd sub nom.* Bernhard v. City of Ontario, 270 F. App'x 518 (9th Cir. 2008) (holding male police officers had a reasonable expectation of privacy against video surveillance in their locker room where employer did not provide notice of surveillance). *But see* Thompson v. Johnson Cnty. Cmty. Coll., No. 96-3223, 1997 WL 139760, at *2 (10th Cir. Mar. 25, 1997) (unpublished opinion) (holding employees had no expectation of privacy in locker area which was located in a room that housed heating and air conditioning equipment and a storage area and for which access was not restricted). An interesting variation on dressing room surveillance, but which highlights the contextual nature of the expectation of privacy, is found in *Bevan v. Smartt*, 316 F. Supp. 2d 1153, 1160–61 (D. Utah 2004) (holding night club dancers had no expectation of privacy vis-à-vis video surveillance of their dressing room by club security personnel, but the dancers did have an expectation of privacy when government agents viewed the same surveillance without a warrant). *See also* Colorado v. Galvadon, 103 P.3d 923 (Colo. 2005) (holding same as to store manager and video surveillance of back room in store with no public access).

162. Feminist Women's Health Ctr. v. Superior Court, 61 Cal. Rptr. 2d 187, 196 (Ct. App. 1997).

recording to be made of an employee in a restroom, locker room, or room designated by an employer for changing clothes, unless authorized by court order."[163] But, the existence of such narrowly-drafted statutes only confirms the general rule that—in the absence of such statutes—employers can destroy the expectation of privacy with detailed notices.

One key concern with webcams is their portability when installed in a laptop computer. They can record an employee's conduct in front of the computer while the employee is at work, at home, or even in a hotel room while traveling on business. The employee may not know if the webcam is activated, or even if it is installed on the laptop being used. While there have not been any reported claims of violations of privacy by employers activating laptop webcams, there has been at least one highly-publicized case involving laptop webcams.

In November 2009, a Pennsylvania high school student and his family learned the school district had obtained video images of the student allegedly engaging in improper behavior in his home from the student's district-issued laptop computer webcam.[164] The webcam on this particular student's laptop computer, like the ones on all the district-issued laptops, could be activated and monitored remotely without the student's or his family's knowledge.[165] The plaintiff alleged the school district had thousands of photos of students in their homes, including some showing students or the family sleeping or in various states of undress.[166] Although the FBI opened an investigation into the incident,[167] authorities decided not to prosecute because they concluded

163. CAL. LAB. CODE § 435(a) (Deering 2010). In addition, this prohibition cannot be waived or derogated from by notice: "No recording made in violation of this section may be used by an employer for any purpose. This section applies to a private or public employer, except the federal government." *Id.* § 435(b). This provision "represent[s] society's understanding that a locker room is a private place requiring special protection." *Trujillo*, 428 F. Supp. 2d at 1106.

164. Complaint at 6, Robbins v. Lower Merion Sch. Dist., No. 10-CV-00665-JD (E.D. Pa. Feb. 16, 2010). The student's allegedly improper behavior was ingesting drugs while using the laptop; in fact, the student was eating candy at the time. John P. Martin, *1,000s of Web Cam Images, Suit Says*, PHILA. ENQUIRER, Apr. 16, 2010, at A1.

165. Complaint, *supra* note 164, at 7.

166. Motion for Sanctions ¶¶ 2, 4, *Robbins*, No. 10-CV-00665-JD (Apr. 15, 2010), ECF No. 44.

167. *See* Order, *Robbins*, No. 10-CV-00665-JD (May 10, 2010), ECF No. 61 (allowing the Government access to the school district's computers and servers); Press Release, Dep't. of Justice, Inquiry into Lower Merion School District Activating Web Cams on Student Issued Computers (Feb. 22, 2010), http://philadelphia.fbi.gov/dojpressrel/pressrel10/ph022210a.htm.

there was no criminal intent on the part of the school district's employees.[168] The school district subsequently settled all cases brought against it by students.[169]

An employer's use of webcams to monitor employee behavior at home can be particularly troublesome for employers. The Supreme Court clearly identifies the home as a bastion of intimacy and privacy.[170] While most courts have ruled against an invasion of privacy based on video recording in the workplace, courts often draw the line in areas society perceives to be intimate.[171] It is within these areas of intimacy that individuals have a reasonable expectation of privacy, an invasion of which could easily be perceived as highly offensive.

3. GPS

Location-tracking technologies allow employers to monitor the exact location of employees, both at the workplace and off-site. One of the principal means of location-tracking, Global Positioning System (GPS) devices, and its employee privacy implications, are discussed in this Section.

GPS uses a satellite positioning system to record both the precise location of a GPS device—as well as the person carrying or using such a device—and the time of positioning.[172] GPS devices are typically installed in vehicles as well as cell phones and can provide tracking information such as the route travelled, the address of all stops, the duration of stops, the amount of time spent traveling between stops, the maximum speed between stops, and whether the device (or person) has entered or exited a pre-determined boundary.[173] Employers primarily use GPS devices to track employee use of vehicles, often to ensure employees are going where they are supposed to be

168. *See* Press Release, Dep't. of Justice, No Criminal Charges Filed Following Lower Marion School District Student Computer Monitoring Investigation (Aug. 17, 2010), http://philadelphia.fbi.gov/dojpressrel/pressrel10/ph081710.htm.

169. Chloe Albanesius, *Pa. School District Settles Webcam Spying Case for $610K*, PCMAG.COM (Oct. 12, 2010), http://www.pcmag.com/Article2/0,2817,2370622,00.asp.

170. *See, e.g.*, Kyllo v. United States, 533 U.S. 27, 37 (2001) ("In the home, our cases show, *all* details are intimate details, because the entire area is held safe from prying government eyes.").

171. *See supra* notes 160–61.

172. *See* Mason Weisz, *Monitoring Employee Location with GPS and RFID in 2005: Workplace Privacy Issues*, *in* WORKPLACE PRIVACY: PROCEEDINGS OF THE NEW YORK UNIVERSITY 58TH ANNUAL CONFERENCE ON LABOR 69, 78–79 (Jonathan Remy Nash & Samuel Estreicher eds., 2010); *see also* Jill Yung, *Big Brother IS Watching: How Employee Monitoring in 2004 Brought Orwell's* 1984 *to Life and What the Law Should Do About It*, 36 SETON HALL L. REV. 163, 170–72 (2005) (providing a description of GPS technology).

173. *See* Weisz, *supra* note 172, at 80.

going and not wasting time during or in between trips.[174] Though recent surveys indicate employers have been slow to adopt GPS technology,[175] the National Workrights Institute predicts work-related GPS tracking will increase because the technology is quickly becoming an affordable option for small businesses.[176]

As with other forms of work-related monitoring, the privacy implications of GPS tracking are unsettled. California is the only state with a statute directly addressing GPS tracking, prohibiting any person or entity within the state from using "an electronic tracking device to determine the location or movement of a person."[177] However, the statute still permits employers to use GPS devices to track the location of their vehicles.[178] Some commentators have concluded that GPS tracking does not fit within any of the types of communications covered under the ECPA.[179]

---

174. *See, e.g.*, *id.* at 81 (recounting a trash hauling business that reduced weekly overtime claims from 300 to 70 hours after installing GPS devices in the company's trucks; also noting transit systems combine GPS with weather and traffic monitoring systems to predict arrival times); Johnathon Williams, *Get a Handle on Your Overhead: Technology Is Making It Easier for You To Keep Tabs on Your Business's Resources*, ENTREPRENEUR, Apr. 21, 2009, http://www.entrepreneur.com/article/201342 (describing one employer who gave employees a per diem for hotels on business trips of over 150 miles only to learn through GPS tracking that some employees were instead driving to and from the work site each day, pocketing the per diem and increasing the mileage and gasoline costs for the company vehicles).

175. *See, e.g.*, AMA & EPOLICY INST., *supra* note 7, at 3 (reporting that only eight percent of employers used GPS to track company vehicles and only three percent used GPS to track company-provided cell phones).

176. NAT'L WORKRIGHTS INST., ON YOUR TRACKS: GPS TRACKING IN THE WORKPLACE 5–6, http://workrights.us/wp-content/uploads/2011/02/NWI_GPS_Report.pdf (last visited Mar. 29, 2011); *see also* Williams, *supra* note 174 (noting that the CEO of a company providing GPS tracking services claims sales have recently grown in "astronomical proportions").

177. CAL. PENAL CODE § 637.7(a) (Deering 2010).

178. *Id.* § 637.7(b) ("This section shall not apply when the registered owner, lessor, or lessee of a vehicle has consented to the use of the electronic tracking device with respect to that vehicle."). Though Connecticut has a statute requiring employers to give employees notice of any electronic monitoring, it is limited to the collection of information at an employer's premises. CONN. GEN. STAT. § 31-48d(3) (2011) (" 'Electronic monitoring' means the collection of information *on an employer's premises* concerning employees' activities or communications by any means other than direct observation . . . .") (emphasis added). Delaware's statute requiring employers to provide employees notice of electronic monitoring only applies to the monitoring or interception of any "telephone conversation or transmission, electronic mail or transmission, or Internet access or usage." DEL. CODE ANN. tit. 19, § 705(b) (2010). *See generally supra* Section II.B.3 (discussing Connecticut's and Delaware's statutory requirements that employers provide notice to employees of workplace electronic monitoring).

179. *See, e.g.*, Weisz, *supra* note 172, at 86.

As discussed *infra*, the contours of permissible GPS tracking are currently being established in federal criminal cases, addressing the constitutionality of warrantless GPS tracking. Though they are not directly applicable to employment scenarios, particularly in the private employment sector, as with *Katz v. United States*, these cases may lay the doctrinal foundation for determining the degree of privacy that may be afforded employees vis-à-vis employer use of GPS tracking.[180]

Most courts have held that the use of GPS devices to track the movements of criminal suspects does not require a warrant based on the U.S. Supreme Court's holding in *United States v. Knotts*.[181] In *Knotts*, a beeper device was attached to a drum of chemicals and then used by law enforcement agents to track the transport of the drum from its point of purchase to the suspect's secluded cabin.[182] The Supreme Court ruled the use of the device did not require a warrant because a "person traveling in an automobile on public thoroughfares has no reasonable expectation of privacy in his movements from one place to another."[183] The rationale used by the Court is that:

> One has a lesser expectation of privacy in a motor vehicle because its function is transportation and it seldom serves as one's residence or as the repository of personal effects. A car has little capacity for escaping public scrutiny. It travels public thoroughfares where both its occupants and its contents are in plain view.[184]

However, in *United States v. Maynard*, the D.C. Circuit Court of Appeals held a suspect's reasonable expectation of privacy was violated by the FBI's warrantless continuous surveillance of the defendant for approximately one month through the installation of a GPS tracking device on the defendant's automobile.[185] The *Maynard* court considered *Knotts* inapplicable because *Knotts* concerned a discreet journey of approximately 100 miles, whereas the

180. 389 U.S. 347 (1967); *see* discussion *supra* Section II.B.2.

181. 460 U.S. 276 (1983); *see, e.g.*, United States v. Marquez, 605 F.3d 604 (8th Cir. 2010); United States v. Garcia, 474 F.3d 994 (7th Cir. 2007); United States v. Jesus-Nunez, No. 1:10-CR-00017-01, 2010 WL 2991229 (M.D. Pa. July 27, 2010). *But see* People v. Weaver, 909 N.E.2d 1195, 1201 (N.Y. 2009) ("The massive invasion of privacy entailed by the prolonged use of the GPS device was inconsistent with even the slightest reasonable expectation of privacy.").

182. *Knotts*, 460 U.S. at 277.

183. *Id.* at 281.

184. *Id.* (citation and internal quotation marks omitted).

185. 615 F.3d 544 (D.C. Cir.), *reh'g en banc denied*, 625 F.3d 766 (D.C. Cir. 2010), *cert. granted sub nom.* United States v. Jones, No. 10-1259, 2011 U.S. LEXIS 4956 (U.S. June 27, 2011).

surveillance in *Maynard* was continuous, twenty-four hours per day, lasting twenty-eight days.[186] As such, the *Maynard* court concluded the prolonged surveillance of the defendant's movements revealed an "intimate picture" of his life that he would expect no one else to have—i.e., he had a reasonable expectation of privacy in his movements over the course of the twenty-eight days.[187]

In contrast, the Ninth Circuit Court of Appeals has held that repeatedly monitoring a suspect's vehicle over a four-month period using various types of mobile tracking devices did not require a warrant.[188] The Ninth Circuit concluded the police obtained no more information than they could have by physically following the suspect and that the use of the tracking devices merely made their work more efficient, which is not unconstitutional.[189] When the Ninth Circuit denied a rehearing en banc, Judge Kozinski wrote in his dissent, "1984 may have come a bit later than predicted, but it's here at last."[190] Judge Kozinski differentiated the beeper technology at use in *Knotts* (which "could help police keep vehicles in view when following them, or find them when they lost sight of them," but "still required at least one officer[,] and usually many more[,] to follow the suspect[]") from current GPS technology (which "can record the car's movements without human intervention[,] quietly, invisibly, with uncanny precision").[191] Judge Kozinski's primary concern is that "[b]y tracking and recording the movements of millions of individuals the government can use computers to detect patterns and develop suspicions. It can also learn a great deal about us because where we go says much about who we are."[192]

Courts are beginning to re-examine the fundamental basis for *Knotts*—that location tracking outside the home is analogous to physical surveillance and therefore does not require a warrant—in light of evolving technology.[193] As expressed by the District Court for the Eastern District of New York:

---

186. *Id.* at 556, 558.

187. *Id.* at 563 (referencing Katz v. United States, 389 U.S. 347 (1967)).

188. United States v. Pineda-Moreno, 591 F.3d 1212, 1213 (9th Cir. 2010), *reh'g denied*, 617 F.3d 1120.

189. *Id.* at 1216.

190. United States v. Pineda-Moreno, 617 F.3d 1120, 1121 (9th Cir. 2010) (Kozinski, J., dissenting).

191. *Id.* at 1124.

192. *Id.*

193. *See, e.g.*, *In re* An Application of the U.S. for an Order Authorizing the Release of Historical Cell-Site Info., 736 F. Supp. 2d 578, 596 (E.D.N.Y. 2010) (concluding the Fourth Amendment prohibits as an unreasonable search and seizure an order for cell phone-based locational data in the absence of a showing of probable cause); *see also In re* An Application of

> [T]echnology has progressed to the point where a person who wishes to partake in the social, cultural, and political affairs of our society has no realistic choice but to expose to others, if not to the public as a whole, a broad range of conduct and communications that would previously have been deemed unquestionably private.[194]

Arguably, employers and employees may find themselves in an ever-changing landscape of privacy protection vis-à-vis the use of GPS tracking devices. Initially, most courts do not consider the use of such devices as an invasion of privacy. However, as their use becomes more sophisticated and continuous, revealing a portrait of personal activities versus merely a recording of location after location, courts begin to recognize an invasion of a reasonable expectation of privacy. This privacy protection may be lost, however, as reflected in the reasoning of *Kyllo v. United States*, once continuous GPS tracking becomes common.[195] U.S. laws condition privacy protections on "actual" expectations of privacy and recognize the validity of implied consent by employees who continue to show up for their "at will" employment. As long as these legal conditions remain, it is up to employers to specifically notify employees of all monitoring and surveillance practices, however intrusive the practices may be, and thus destroy any "actual" expectation of privacy.

D. WORKPLACE PRIVACY TRENDS IN THE UNITED STATES

Absent specific state laws limiting intrusive employee monitoring—which tend to be few and narrowly drafted—employers are free to destroy U.S. employees' expectations of privacy via detailed notices, and without an actual expectation of privacy, employee privacy is not protected against monitoring under federal law and general state privacy laws. The U.S. Supreme Court could have changed this situation in *Quon* by developing core privacy rights that cannot be destroyed or limited through notices, but the Court chose not to. Instead, the Court found it prudent to not "establish far-reaching premises that define the existence, and extent, of privacy expectations enjoyed by employees when using employer-provided communication devices."[196] The Court was not concerned with whether Quon had a reasonable expectation of privacy in his text communications; as long as it had a legitimate business purpose, the City of Ontario could review his text

---

the U.S. for an Order Authorizing the Release of Historical Cell-Site Info., 10-MC-897 (NGG), 2011 U.S. Dist. LEXIS 93494 (E.D.N.Y. Aug. 22, 2011) (holding same).

194. *In re An Application*, 736 F. Supp. 2d at 582 (citations omitted).

195. *Kyllo*'s reasoning suggests that increased "general public use" might diminish an expectation of privacy. *See supra* Section II.C.

196. City of Ontario, Cal. v. Quon, 130 S. Ct. 2619, 2629 (2010).

messages.[197] And when subsequently confronted with the question of whether employees had a reasonable expectation of work-related privacy, the Supreme Court again refused to discuss its contours. Instead, it merely assumed the existence of a right to informational privacy before deciding that legitimate employer needs justified investigatory background checks regarding employees' personal lives.[198]

Even without acknowledging a fundamental or core privacy right that is beyond destruction via employer notices, U.S. courts can protect employee privacy by holding the level of detail provided in employer notices to high and increasing standards. For example, if an employer notifies employees about e-mail monitoring in general terms, courts may find that such notice is not sufficient to destroy an expectation of privacy in private e-mail (even if accessed at work), text messaging, or instant messaging. But, the jurisprudence on this point is mixed and U.S. courts tend to interpret monitoring notices broadly in favor of employer monitoring, so long as the monitoring serves legitimate business purposes. For example, in *Stengart v. Loving Care Agency, Inc.*, the Superior Court of New Jersey ruled that an employee had a reasonable expectation of privacy in e-mail messages she sent to her personal attorney using an employer-provided computer.[199] However, the *Stengart* court declined to "attempt to define the extent to which an employer may reach into an employee's private life or confidential records through an employment rule or regulation."[200] And a California Court of Appeal, in a situation very similar to *Stengart*, held that an employee did not have a reasonable expectation of privacy in e-mail messages she sent to her personal attorney using an employer-provided computer because the employer had "unequivocally" informed its employees that those who used the company's computers to send personal e-mail would have "no right of privacy" in the information sent.[201]

In summary, we see two trends emerging from cases addressing employees' reasonable expectations of privacy in work-related

197. *Id.* at 2631.

198. NASA v. Nelson, 131 S. Ct. 746 (2011).

199. 973 A.2d 390, 401 (N.J. Super. Ct. App. Div. 2009) ("A policy imposed by an employer, purporting to transform all private communications into company property—merely because the company owned the computer used to make private communications or used to access such private information during work hours—furthers no legitimate business interest.") (citation omitted).

200. *Id.*

201. Holmes v. Petrovich Dev. Co., 119 Cal. Rptr. 3d 878, 896–97 (Ct. App. 2011) (internal quotation marks omitted). The *Holmes* court described the plaintiff's sending personal e-mails through the company's computer system as "akin to consulting her lawyer in her employer's conference room, in a loud voice, with the door open." *Id.* at 883.

communications. First, courts, particularly the U.S. Supreme Court, have shied away from acknowledging a core privacy right that employers cannot destroy by way of notice. Thus, employers in the United States are free to completely or partially destroy employee privacy expectations—and with the expectations, also destroy most forms of legal protections for data privacy under U.S. law, because privacy protections are conditioned on privacy expectations. Second, any limited expectation of privacy that may exist due to too narrowly or poorly drafted employer notices can be negated based on a broad interpretation of the applicable notice if the actual monitoring at hand is supported by an employer's legitimate business interest in monitoring employee communications. While the courts appear willing to address the legitimacy of business interests, the issue will most likely be decided on a case-by-case basis. And in evaluating the legitimacy of a business interest, the guidance is not necessarily clear. In *Stengart*, the court concluded the employer's policies were ambiguous as to employees' personal use of computer systems,[202] while the *Holmes* court concluded the employer's policies were unambiguous.[203] Finally, in *Quon*, the Supreme Court acknowledged that the employer's policies did not explicitly address text messages but concluded that verbal notice that text messages were considered the same as e-mail messages was sufficient to incorporate text messages into the formal city policies.[204]

Thus, employees should anticipate very minimal expectations of privacy in workplaces within the United States.

III. EMPLOYER MONITORING AND EMPLOYEE PRIVACY—EUROPEAN PERSPECTIVE

Employers in Europe have access to the same monitoring technologies that are available to employers in the United States. Furthermore, multinational groups operating in Europe and the United States tend to face technical pressures to implement technologies uniformly across the global enterprise. Thus it is not surprising that in a few cases, employers have become entangled in some of the same monitoring-related disputes in European courts as discussed in Part II, *supra*, with respect to the United States.[205] But, the following review of European cases will show that: first, the monitoring activities challenged in European courts tend to be far less

202. *Stengart*, 973 A.2d at 396.
203. *Holmes*, 119 Cal. Rptr. 3d at 896–97.
204. City of Ontario, Cal. v. Quon, 130 S. Ct. 2619, 2625 (2010).
205. For discussion of U.S. law, see *supra* Part II.

intrusive as in some U.S. cases; second, European employees tend to win privacy-based lawsuits; third, European employers are not required or expected to engage in intrusive employee monitoring; and fourth, European employers are typically unable to defend their practices based on notices.

A. LAWS IN EUROPE—OVERVIEW

First, in Europe, there is technically no uniform body of "European law" that directly applies between employers and employees. In most if not all European countries, however, some laws agreed to or enacted on a supranational level apply in one form or another, as implemented into national law or with immediate legal effect at the national law level. To understand current employee privacy law in Europe, one must note the different legal regimes, legislatures, and courts in Europe that have been making and interpreting law in this area, including: national legislatures, international treaties, and the European Union (and its predecessor organizations). Second, the concept of data protection in Europe does not completely mirror the concept of privacy in the United States.

B. CIVIL RIGHTS PROTECTIONS FOR PRIVACY AT THE EUROPEAN LEVEL

The European Convention on Human Rights was signed in 1950 by ten founding member states.[206] Today, a larger group of European countries has signed and implemented the European Convention on Human Rights, adjudicated by the European Court of Human Rights in Strasbourg. The Convention expressly protects individual privacy against government interference:

> Article 8—Right to respect for private and family life
>
> 1. Everyone has the right to respect for his private and family life, his home and his correspondence.
>
> 2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.[207]

206. Belgium, Denmark, France, Ireland, Italy, Luxembourg, the Netherlands, Norway, Sweden, and the United Kingdom. Convention for the Protection of Human Rights and Fundamental Freedoms, June 1, 2010, C.E.T.S. No. 194, *available at* http://conventions.coe.int/treaty/en/Treaties/Html/005.htm [hereinafter Human Rights Convention].

207. Human Rights Convention, *supra* note 206, § 1.

Article 8 has been refined in a number of cases by national courts and the European Court of Human Rights, which have applied the right to various forms of intrusion into data privacy. However, the courts have usually only applied the Article 8 right to privacy in situations where a state actor (i.e., a government entity), and not a private sector employer, has interfered. For example, in *Halford v. United Kingdom*[208] and *Kopp v. Switzerland*,[209] the European Court of Human Rights acknowledged that state employees have a right to privacy with respect to phone calls made from their government-operated work locations.

In *Copland v. United Kingdom*,[210] the European Court of Human Rights expanded Article 8 protection to e-mails and ruled that phone connection data (e.g., time of connection and numbers called) as well as e-mail and internet usage information (e.g., websites visited and numbers of e-mails sent and received) were also protected. In *Copland*, a state-owned college had monitored the e-mail and internet usage of its employees for purposes of determining abuse. Details of the monitoring program were disputed by the parties to the underlying litigation, but the court decided that even the undisputed minimal amounts of the college's monitoring (e.g., recording and analysis of connection information) were incompatible with the protections in Article 8 of the Convention without a specific statutory basis.[211] The general statutory authorization of the college to do " 'anything necessary or expedient' for the purposes of providing higher and further education" was insufficient as a basis for the monitoring; hence, the monitoring violated the plaintiff's rights.[212]

Given this finding, the court in *Copland* was neither required nor given the opportunity to decide what the college—or another government institution—could monitor, given more specific statutory authorizations; but based on other cases, the court would likely have applied a balancing test of the individual's right to privacy and any legitimate purposes of government that are necessary in a democratic society.[213] The European Court of Human Rights weighed heavily the fact that the plaintiff had a "reasonable expectation" of privacy which was not met due to the lack of notices and

208. Halford v. United Kingdom, 24 Eur. Ct. H.R. 523 (1997).

209. Kopp v. Switzerland, App. No. 23224/94, 27 Eur. H.R. Rep. 91 (1998).

210. 45 Eur. Ct. H.R. 253 (2007).

211. *Id.*

212. *Id.* ¶ 47.

213. *See* Gerrit Hornung, *EGMR: Überwachung Privater E-Mail und Internetnutzung am Arbeitsplatz* [*Monitoring of Private E-mail and Internet Use at the Workplace*], 12 MULTIMEDIA UND RECHT [MMR] 431, 432 (2007).

specific legislation or other publicized rules.[214] Therefore, it appears that the court might have accepted a statutory basis for monitoring that both requires the employer to give employees reasonable prior notice and applies some restraint on the methods used with respect to their intrusiveness.[215] But how strict a degree of scrutiny the court would apply to such legislation remains unclear.

The European Convention and Court of Human Rights exist independently of the European Union (EU), which is not a member of the convention and which has grown out of efforts to achieve economic integration rather than civil rights protections.[216] In 1957, France, Germany, Italy, and the Benelux countries—a different group of founders than for the European Convention of Human Rights—signed the Treaty of Rome to establish the European Economic Community (EEC), whose primary goal was to achieve integration via trade with a view to economic expansion.[217] In 1992, a larger group of countries signed the Maastricht Treaty and eliminated "Economic" from the name of the new European Community (EC), reflecting a collective determination to expand the Community's powers to non-economic domains.[218]

The Maastricht Treaty also created the European Union to document ambitions for non-economic integration.[219] But while the European Union was to serve largely symbolic functions, the European Community remained the law-making body, and economic integration remained the main driver of EC activities for the remainder of the last century.[220] Therefore, much of the supranational European legislation remained focused on removing barriers to trade and protecting the economic freedoms of businesses in Europe.[221]

---

214. *Copland*, 45 Eur. Ct. H.R. 253, ¶ 42.

215. *See* Hornung, *supra* note 213, at 432.

216. According to Article 6 of the Treaty of Lisbon Amending the Treaty on European Union and the Treaty Establishing the European Community, the Union shall accede to the European Convention for the Protection of Human Rights and Fundamental Freedoms. 2007 O.J. (C 306) 1, 135 (Dec. 17, 2007), *available at* http://www.ecb.int/ecb/legal/pdf/en_lisbon_treaty.pdf [hereinafter Treaty of Lisbon].

217. *See* Treaty of Rome, Mar. 25, 1957, *available at* http://ec.europa.eu/economy_finance/emu_history/documents/treaties/rometreaty2.pdf.

218. *See* The Maastricht Treaty: Provisions Amending the Treaty Establishing the European Economic Community with a View to Establishing the European Community, Feb. 7, 1992, 31 I.L.M. 247 (1992), *available at* http://www.eurotreaties.com/maastrichtec.pdf [hereinafter Maastricht Treaty].

219. *See id.* tits. IX–XI.

220. *See id.*

221. *See, e.g.*, Peter Tettinger, *Die Charta der Grundrechte der Europäischen Union* [*The Charter of Fundamental Rights of the European Union*], 54 NEUE JURISTISCHE WOCHENSCHRIFT [NJW] 1010 (1014), 2001 (Ger.).

Consequently at the European level, there was relatively little mandate or perceived need to protect privacy. Whenever national authorities and legislatures restricted individual freedoms or civil rights, EC law offered protective rights to individuals to challenge restrictions in the economic sector, and national constitutional laws offered additional protection for individual rights in the economic and private spheres.[222] But, in its effort to harmonize economic conditions, the EEC (and later the EC and EU) not only struck down trade-restricting national legislation and regulations but also increasingly imposed harmonized legislation, primarily through Directives that the member states were required to implement into national law.[223] EEC legislation covered any topic considered economically relevant (e.g., environmental and product safety standards, consumer contracts, advertising) and sought to create a level playing field for businesses in Europe.[224] As such, the legislation had the potential to restrict individual freedoms as much as previously-national legislation did.

Given the supremacy of EC law (and later EU law) over national law, national constitutions could no longer fully protect individual rights without endangering European harmonization and integration. To reduce the risk of challenges to European laws under national constitutional laws as well as the risk of diverging national standards on this topic, the European Court of Justice (the European Community's Court and now the European Union's Court) tried to fill the "civil rights vacuum" by inventing a suite of European constitutional principles that could be used to challenge EC/EU legislation, and of which the European Court of Justice remained the ultimate guardian.[225] The European Court of Justice developed these European civil rights in reference to principles in the European Convention of Human Rights and national constitutional laws, as assessed and defined by the European Court of Justice from time to time.[226] EU member states perceived this "ad hoc" development of human rights protections as unsatisfactory. After long negotiations, the EU member states agreed on a Charter of Fundamental Rights of the European Union in 2000.[227] This Charter expressly protects privacy and personal data:

222. THOMAS DIETERICH ET AL., ERFURTER KOMMENTAR ZUM ARBEITSRECHT [COMMENT ON LABOR] ¶¶ 114–117, at 23–24 (10th ed. 2010).

223. *See* Tettinger, *supra* note 221.

224. *See id.*

225. *See, e.g.*, *id.* at 1014.

226. *See id.*

227. Charter of Fundamental Rights of the European Union, 2000 O.J. (C 364) 1 (Dec. 18, 2000), *available at* http://www.europarl.europa.eu/charter/pdf/text_en.pdf.

> Article 7—Respect for private and family life
>
> Everyone has the right to respect for his or her private and family life, home and communications.
>
> Article 8—Protection of personal data
>
> 1. Everyone has the right to the protection of personal data concerning him or her.
>
> 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
>
> 3. Compliance with these rules shall be subject to control by an independent authority.[228]

Under the Treaty of Lisbon, the Charter is legally binding.[229] However, due to the jurisdictional limitations of EU law, the EU Charter of Fundamental Rights applies only if and to the extent that EU member states implement or enforce EU law over their respective national laws. Courts and scholars increasingly reference EU law, usually without clarifying whether the existence of a particular civil right protection in the EU Charter actually changed the legal situation as a matter of law, rather than as a matter of public policy.[230]

C. THE EC'S DATA PROTECTION DIRECTIVE

In 1995, because diverging national standards and cross-border data transfer restrictions had become an obstacle to trade in the Common Market,[231] the European Community attempted to harmonize data protection laws across the EC member states through the EC Data Protection Directive.[232] In order to secure approval from EC member states with

228. *Id.* arts. 7–8.

229. Article 6 of the Treaty on the European Union, as amended by the Lisbon Treaty, is binding on all EU member states, except for member states with an opt-out for this provision. *See* Treaty of Lisbon, *supra* note 216, at 156.

230. *See, e.g.*, Tettinger, *supra* note 221, at 1014.

231. The German State of Hessen passed the world's first data protection law in 1970. *See Privacy in Hessen*, LANDESPORTAL HESSEN, *available at* http://www.hessen.de/irj/hessen_Internet?cid=098693b3bbacadc19b81045a1c2300f2 (last visited Jan. 26, 2011). Other German states and European countries quickly followed suit. *See Law Texts and Comments*, VIRTUAL PRIVACY OFFICE, *available at* http://www.datenschutz.de/recht/gesetze/ (last visited Mar. 6, 2011).

232. Directive 95/46/EC, 1995 O.J. (L 281) 31 (Nov. 23, 1995), *available at* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML [hereinafter EC Data Protection Directive].

historically high data protection standards, the EC adopted a general prohibition on the processing of any personal data and a particular ban on transfers outside the European Economic Area (EEA), subject to a number of narrow, enumerated exceptions.[233] All EEA member states[234] had to implement these substantive requirements into national legislation,[235] but they retained jurisdiction to legislate administrative details such as notification and approval requirements, penalties, and enforcement procedures. Given the jurisdictional limitations of the European Community to regulate economic activity via Directives, the EC Data Protection Directive covers data processing activities by private sector employers and possibly government-owned businesses, but not by government entities in their capacity as state actors.[236]

A primary objective of the national legislation that prompted the EU harmonization initiative was to regulate and limit automated processing of personal data because of perceived danger from government—Big Brother

233. The European Economic Area (EEA) is comprised of the twenty-seven EU member states, plus three more—Iceland, Liechtenstein, and Norway—which agreed under a separate treaty to adopt certain EU laws. *See* Agreement on the European Economic Area (EEA), 1994 O.J. (L 1) 3 (May 2, 1992), *available at* http://ec.europa.eu/world/agreements/prepareCreateTreatiesWorkspace/treatiesGeneralData.do?step=0&redirect=true&treatyId=1.

234. The EEA was established in 1994, following an agreement between the member states of the European Free Trade Association (EFTA) and the EC, later the EU. The treaty allows Iceland, Liechtenstein, and Norway to participate in the EU's single market without a conventional EU membership. In exchange, they are obliged to adopt all EU legislation related to the single market, except laws regarding agriculture and fisheries. One EFTA member, Switzerland, has not joined the EEA.

235. The EC Commission collects unofficial English translations of national legislation. *See Policy Papers from National Data Protection Authorities*, EUR. COMM'N, *available at* http://ec.europa.eu/justice/policies/privacy/policy_papers/policy_papers_en.htm (last updated Aug. 6, 2010).

236. Article 3.2 of the EC Data Protection Directive provides:

> This Directive shall not apply to the processing of personal data . . . in the course of an activity which falls outside the scope of Community law, such as those provided for by Titles V and VI of the Treaty on European Union and in any case to processing operations concerning public security, defence, State security (including the economic well-being of the State when the processing operation relates to State security matters) and the activities of the State in areas of criminal law.

EC Data Protection Directive, *supra* note 232. This limitation is a function of the principles of limited competences and subsidiary, which are codified in Articles 4 and 5 of the EU Treaty, as amended by the Lisbon Treaty, whereby the EU has limited competences and the EU shall only exercise its competences to the extent the Member States cannot effectively legislate a particular topic. *See* Consolidated Version of the Treaty on European Union, 2010 O.J. (C 83) 13 (Mar. 3, 2010), *available at* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:083:0013:0046:EN:PDF.

watching the "transparent citizen"[237]—and the private sector—businesses creating large databases that could be accessed and abused by government.[238] Consequently, EC data protection laws, on a national level as well in the Directive, prohibit the processing of personal data unless a specific statutory exemption applies.[239]

In contrast to U.S. data *privacy* laws, European data *protection* laws do not condition protection on an expectation of privacy. The data protection laws protect the right to privacy and regulate the processing of personal data within the European Union. These laws define "personal data" and "processing" very broadly and cover even publicly available data. Any information relating to an identifiable individual is "personal data"[240] and any

---

237. The term "transparent citizen" originates from the German term "gläserner Bürger" (literally translated: glass citizen) used by scholars and politicians to illustrate the dangers of government and private surveillance. *See, e.g.*, Hans U. Buhl & Günter Müller, *The "Transparent Citizen" in Web 2.0: Challenges of the "Virtual Striptease,"* 4 BUS. & INFO. SYS. ENGINEERING 203 (2010), *available at* http://www.bise-journal.org/pdf/1_60896.pdf.

238. *See* Stefan Krempel, *Vom gläsernen Bürger zum gläsernen Staat* [*From the Glass Citizen to the Glass State*], TELEPOLIS, June 18, 2000, *available at* http://www.heise.de/tp/artikel/8/8262/1.html; *see also* ENTSCHEIDUNGEN DES BUNDESVERFASSUNGSGERICHTS [BVerfGE] [Federal Constitutional Court] 37 NJW 419 (422), 1984 (Ger.) (holding in the census decision that there are no insignificant dates given the technical development, and deriving the right to informational self-determination from the general personality right interpreting Article 2, ¶ 1 in conjunction with Article 1, ¶ 1 of the German Constitution). The decision had a profound impact both in Germany and Europe—the principles laid down in it appear in the state data protection acts the following years, as well as in the General Amendment to the German Federal Data Protection Act of 1990.

239. Article 7 of the EC Data Protection Directive:

> [P]ersonal data may be processed only if: (a) the data subject has unambiguously given his consent; or (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or (c) processing is necessary for compliance with a legal obligation to which the controller is subject; or (d) processing is necessary in order to protect the vital interests of the data subject; or (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed; or (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection under Article 1 (1).

EC Data Protection Directive, *supra* note 232, art. 7.

240. Article 2(a) of the EC Data Protection Directive:

> "[P]ersonal data" shall mean any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to

collection, use, and transfer—even the redaction and deletion thereof—constitutes "processing."[241]

Employers in the European Economic Area routinely rely on three exemptions for their processing of personal data: (1) a necessity to perform contractual obligations with the data subject, (2) individual consent from the data subject, and (3) a legal requirement to process personal data based on statutory obligations or orders from the government of the country whose data protection laws apply.[242] In extraordinary situations they may be able to rely on a balancing-of-interests test,[243] but in practice, employers typically

---

> an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

Id. art. 2(a). Switzerland and some EEA Member States, including Austria, expand the definition of personal data to information relating to a specific legal entity. *See* BUNDESGESETZ ÜBER DEN DATENSCHUTZ [DPA] [FEDERAL ACT ON DATA PROTECTION], SR 235.1 (1992), art. 3(a), (b) (Switz.), *available at* http://www.admin.ch/ch/e/rs/2/235.1.en.pdf ("data subjects: natural or legal persons whose data is processed"); BUNDESGESETZ ÜBER DEN SCHUTZ PERSONENBEZOGENER DATEN [DSG] [FEDERAL ACT CONCERNING THE PROTECTION OF PERSONAL DATA] BUNDESGESETZBLATT I [BGBL I], No. 165/1999, art. 2, pt. 1, § 4, ¶ 3 (Austria), *available at* http://www.dsk.gv.at/DocView.axd?CobId=41935 (last visited Apr. 26, 2011) (" 'Data Subject' . . . : any natural or legal person or group of natural persons not identical with the controller, whose data are processed . . . .").

241. Art. 2(b) of the EC Data Protection Directive:

> "[P]rocessing of personal data" ("processing") shall mean any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.

EC Data Protection Directive, *supra* note 232, art. 2(b).

242. There are a variety of situations exemplifying this point. Employers are typically required to report certain information to local tax authorities, and local law enforcement agencies can demand the disclosure of certain personal data so long as procedural and formal safeguards are observed. However, multinational companies cannot always rely on these exceptions where data collection or disclosure obligations follow from statutes or government agencies in another country, in particular from countries outside the EEA. For instance, airlines were for a while caught in a crossfire of conflicting data protection/disclosure obligations. *See* Lothar Determann, *Conflicting Data Laws: Airlines Are Damned If They Do, Don't*, S.F. DAILY J., Sept. 23, 2003, at 5. More recently, the Society for Worldwide Interbank Financial Telecommunication (SWIFT) was caught between European data protection requirements and subpoenas from U.S. tax authorities. *See* Press Release, European Commission, The SWIFT Case and the American Terrorist Finance Tracking Program (June 28, 2007), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/07/266&format=HTML&aged=0&language=EN&guiLanguage=en.

243. Article 7 of the EC Data Protection Directive also allows processing if necessary for (c) "compliance with a legal obligation to which the controller is subject" (but legal obligation means typically "legal obligation under local law" or "under laws that conform to EC law"); (d) "in order to protect the vital interests of the data subject" (interests are

find it difficult to meet the high standards applied by courts and data protection authorities.[244] Principal exceptions to processing personal information are discussed *infra*.

1. Necessity Under Contract

Contractual duties serve as justification only if the processing is truly necessary for the performance of a contract between the data subject and the data controller.[245] Necessity can be assumed where the employer processes personal data to enable employees to do their job (e.g., by offering e-mail, internet connectivity, data storage, etc.). However, with respect to monitoring, employers will typically not be able to show a necessity under employment contracts, because European employers do not include express duties to monitor in their agreements and they are not obligated to monitor employees under applicable statutes.[246]

2. Consent

Unlike in the United States, it is not possible within the European Union to unilaterally destroy an expectation of privacy—the employer must affirmatively seek employee consent in order to rely on it as an exception from the general prohibition of monitoring activities involving data processing.[247] Further, consent is valid only if the data subject grants it in an informed, voluntary, express, specific, and written manner.[248] The

considered vital in cases of medical emergencies, but probably not in most cases of commercial convenience or in most other situations in which companies would like to refer to this exception); (e) "for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed;" or (f) "for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection under Article 1 (1)" (again, national data protection authorities apply high standards). EC Data Protection Directive, *supra* note 232, art. 7.

244. *See Frequently Asked Questions Relating to Transfers of Personal Data from the EU/EEA to Third Countries*, EUROPEAN COMM'N, 49, http://ec.europa.eu/justice/policies/privacy/docs/international_transfers_faq/international_transfers_faq.pdf (last updated Aug. 6, 2010) [hereinafter FAQ].

245. *See* EC Data Protection Directive, *supra* note 232, arts. 2(h), 7, 26.

246. Achim Lindemann, *Betriebsvereinbarungen zur E-Mail-, Internet- und Intranet-Nutzung* [*Operating Agreements for E-Mail, Internet and Intranet Use*], DER BETRIEBSBERATER 1950, 1951 (2001).

247. *See* Lindemann, *supra* note 246.

248. *See* EC Data Protection Directive, *supra* note 232, arts. 2(h), 7(a). Recent Mexican data privacy legislation follows the EU model in many respects, but it accepts implied consent upon receipt of a sufficiently detailed notice, which is similar to the U.S. approach in this respect. *See* Lothar Determann & Sergio Legorreta, *New Data Privacy Law in Mexico*, 10

"voluntariness" requirement raises significant difficulties in the employment context. The national data protection authorities in most EEA member states[249] presume that employee consent is coerced, and hence involuntary, given the typical balance of power in the employment relationship.[250] In order to overcome this presumption, the employer must give conspicuous notice that each employee is entitled to withhold consent without any unduly adverse consequences, so that employees are truly in a position where they can voluntarily grant or deny consent. As a practical matter, however, employers can then expect that some employees will deny or later revoke their consent. This alone tends to render any systematic deployment of monitoring technologies based on employee consent impractical.

3. Statutory Obligations

Employers are not legally required to monitor employees in most EU member states, nor do they face the same kinds of liabilities as U.S. employers that provide indirect motivation for monitoring programs.[251] But automated monitoring programs such as e-mail filtering and blocking of potentially harmful websites may best address the obligations imposed by data protection laws and laws requiring enterprises to maintain controls and transparency.[252] For example, Germany enacted a statute in 1998 regarding controls and transparency in enterprises, which requires companies to establish risk management, protection, and control systems, as well as monitoring programs to enforce such controls.[253] Also, data protection laws

IAPP PRIVACY ADVISOR 1 (Nov. 2010), *available at* https://www.privacyassociation.org/publications/2010_10_26_new_data_privacy_law_in_mexico/.

249. Article 29 Working Party Working Document on Surveillance and Monitoring of Electronic Communications in the Workplace (Article 29 Data Prot. Working Party, Working Paper No. 55, Reference 5401/01, 2002), *available at* http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2002/wp55_en.pdf. The "national data protection authorities" are in reference to independent government agencies and not a body of law. For a list, see *Privacy and Data Protection Authorities*, COUNCIL OF EUR., http://www.coe.int/t/dghl/standardsetting/dataprotection/Supervisory%20_Authorities_en.asp (last visited May 22, 2011).

250. FAQ, *supra* note 244, at 50; *see also* Determann & Brauer, *supra* note 71; Determann, *supra* note 71.

251. *See supra* Section II.A.

252. *See* Michael Schmidl, *E-Mail-Filterung am Arbeitsplatz* [*E-mail Filtering in the Workplace*], 13 MMR 343, 345–46 (2005); Michael Schmidl, *Aspekte des Rechts der IT-Sicherheit* [*Aspects of the Law of IT-Security*] [Feb. 18, 2010], 63 NJW 476 (478), 2010 (Ger.).

253. *See* GESETZ ZUR KONTROLLE UND TRANSPARENZ IM UNTERNEHMENSBEREICH [KONTRAG] [CORPORATE SECTOR SUPERVISION AND TRANSPARENCY ACT], Mar. 5, 1998, DEUTSCHER BUNDESTAG: DRUCKSACHE [BT] 13/10038 art. 1, § 9(c) (Ger.), *available at* http://www.sicherheitsforum-bw.de/x_loads/KonTraG.pdf (last visited May 22, 2011) (adding a section 91, paragraph 2 of the German Share Corporation Act (Aktiengesetz)

require technical and administrative security measures to protect the integrity and confidentiality of personal data.[254]

4. Balancing Test

Pursuant to Article 7(f) of the EC Data Protection Directive, EU member states have enacted exceptions from the general prohibition to process personal information if and to the extent that "data processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection . . . ."[255] In one example, based on the German version of this "balancing test exception," the German Federal State-owned railway company, Deutsche Bahn AG, engaged in data mining, comparing names in its human resources database with names in its accounts-payable database to identify matches that might warrant further investigations into self-dealing, bribery, nepotism, and favoritism regarding suppliers with family connections to employees.[256] The company apologized publicly and acknowledged that it had conducted automated comparisons of the addresses and bank account information of 175,000 employees with those of Deutsche Bahn suppliers.[257] Prosecutors announced investigations of the management of Deutsche Bahn.[258] Public outcry with respect to this monitoring program, as well as allegations regarding individual follow-up investigations, caused the CEO of Deutsche Bahn to resign and the German government to amend the Federal Data Protection Act. A new Section 32 clarified that employers may generally

according to which share companies have to establish risk management controls and monitoring programs, arguably also including information technology control mechanisms).

254. For example, section 9 of the German Federal Data Protection Act requires technical protection measures. For additional examples and references, see Lothar Determann & Jesse D. Hwang, *Data Security Requirements Evolve: From Reasonableness to Specifics*, 26 COMPUTER & INTERNET LAW., Sept. 2009, at 6, 10.

255. EC Data Protection Directive, *supra* note 232, art. 7(f).

256. Nicolas Mähner, *Neuregelung des § 32 BDSG zur Nutzung Personenbezogener Mitarbeiterdaten am Beispiel der Deutschen Bahn AG* [*Revision of § 32 BDSG [German Federal Data Protection Act] on the Use of Personal Employee Data Using the Example of the Deutsche Bahn AG*], 13 MMR 379 (2010).

257. Von M. Bauchmüller & Klaus Ott, *Mehdorn Verschweigt Weiteren Daten-Skandal* [*Mehdorn Conceals New Data Scandal*], SUEDDEUTSCHE.DE, Feb. 3, 2009, *available at* http://www.sueddeutsche.de/wirtschaft/386/457048/text/; Brett Neely, *Deutsche Bahn Chief Mehdorn Apologizes to Workers on Data Probe*, BLOOMBERG (Feb. 6, 2009), http://www.bloomberg.com/apps/news?pid=20601100&sid=aMDxC5iRb7nM.

258. Sabine Siebold, *Staatsanwaltschaft prüft Ermittlungen gegen Bahn-Chef* [*Public Prosecutors Consider Investigations of Deutsche Bahn Chief*], REUTERS (Feb. 12, 2009), http://de.reuters.com/article/deEuroRpt/idDELC60224520090212.

process personal data of employees only for purposes of concluding, maintaining, and terminating employment relationships. Personal data of employees may be collected and otherwise processed for purposes of uncovering criminal actions only if and to the extent that (1) actual documented facts create a suspicion of criminal activities, (2) the processing is necessary, and (3) the interests of the individual employee do not outweigh the interests of the employer.[259] These specific rules do not contemplate routine monitoring or processing of personal data for purposes of investigating infractions that do not amount to criminal acts (such as potential violations of a company-wide code of conduct or similar rules).[260] Employee consent is not mentioned as a possible means of legitimizing monitoring programs.

D. NATIONAL WIRETAP LAWS IN EUROPE (CASE STUDY: GERMANY)

In addition to data privacy laws, employers must observe restrictions under anti-wiretap laws, which have not (yet) been harmonized throughout Europe. Under German federal telecommunications law, for example, employers who expressly allow or tolerate some private use of the Internet, e-mail, or other electronic communications systems are treated like telecommunications service providers and are fully subject to telecommunications secrecy provisions.[261] As such, employers cannot even implement anti-spam filtering or anti-virus filtering technologies without

259. Mähner, *supra* note 256.

260. On August 25, 2010, the German government presented a bill amending the Federal Data Protection Act. The bill, *inter alia*, deals with the automated matching of employee data for internal compliance investigations. The bill provides for a two stage escalation model. In the first stage, only anonymous or aliased data may be matched for the purpose of disclosing severe breaches of duty, especially crimes committed during the employment (e.g., corruption). Marie-Theres Tinnefeld, Thomas Petri & Stefan BrinkTinnefeld, *Aktuelle Fragen um ein Beschäftigtendatenschutzgesetz—Eine erste Analyse und Bewertung* [*Current Topics Around Employees' Data Protection*], 13 MMR 727, 732 (2010); Norton Rose LLP, *Neuer Gesetzesentwurf zum Beschäftigtendatenschutz* [*New German Draft Bill Regarding Employee Data Protection*], *available at* http://www.nortonrose.com/knowledge/publications/2010/pub30760.aspx?lang=de-de&page=all (last visited Apr. 29, 2011). According to the draft bill, employers may initiate the first stage without cause for suspicion of breach. Routine spot tests are permissible. *See* Michael Schmidl & Benjamin Baeuerle, *German Employee Data Protection Law Proposed by Government*, 10 WORLD DATA PROTECTION REP., no. 9, 2010, at 28, 28–29.

261. *See, e.g.*, Thorsten B. Behlinger, *Compliance Versus Fernmeldegeheimnis* [*Compliance Versus Privacy of Telecommunications*], 19 BETRIEBS BERATER 892, 892 (2010); René Hoppe & Frank Braun, *Arbeitnehmer-E-Mails: Vertrauen ist Gut—Kontrolle ist Schlecht—Auswirkung der neusten Rechtsprechung des Bundesverfassungsgerichts auf das Arbeitsverhältnis* [*Employees' E-mails: Faith Is Good, Checks Are Bad—Consequences for the Employer-Employee Relationship Arising Out of the Latest Decisions by the Federal Constitutional Court*], 13 MMR 80, 81 (2010).

valid, individual consent, which is extremely difficult, if not impractical, to obtain from individuals. Without such consent, filtering technologies can only be deployed as necessary to protect the network, without an option for the employer to access individual filter reports or quarantined e-mails for productivity or compliance monitoring.[262] Theoretically, German employers can avoid this consequence by strictly prohibiting personal use of communications systems, because the German telecommunications laws only apply to public systems, not to closed systems. But, in practice, employees expect, and employers allow, limited personal use of company communications systems.

E. WORK-RELATED ELECTRONIC MONITORING

French courts have been even stricter by protecting employees from e-mail searches whether or not the employer allows private use. In one case, an employer was sanctioned for terminating the employment contract of an employee who had been running a competing consulting business from his workplace, using the employer's e-mail system to accept orders and process engagements for services that were similar to those that the employer offered to customers.[263] The court reprimanded the employer for the fact that it had not notified the employee of the possibility of searches into e-mail folders that were labeled "personal," as well as for the fact that the employer had not submitted required notifications to the French data protection authorities.[264] Consequently, the court invalidated the termination, ordering reinstatement and damages for the employee.

A number of EC member states, including Germany, Italy, the Netherlands, Spain, and the United Kingdom, strictly prohibit ongoing monitoring of employee communications and permit electronic monitoring

---

262. Michael Schmidl, *Decision 2 BvR 902/06 of the German Constitutional Court: The End of E-Mail Screening in the Workplace*, 9 WORLD DATA PROTECTION REP., no. 8, 2009, at 15, 15; Schmidl, *E-Mail-Filterung*, *supra* note 252, at 345.

263. Cour d'Appel [CA] [regional court of appeal] Versailles, Apr. 2, 2003, Aff. No. 02/00293 (Fr.); *see also* Kunz Kömpf, *Kontrolle der Nutzung von Internet und E-Mail am Arbeitsplatz in Frankreich und Deutschland* [*Controlling the Use of Internet and E-mail in the Workplace in France and Germany*], 26 NEUE ZEITSCHRIFT FÜR ARBEITSRECHT [NZA] 1341, 1343 (2007) (Ger.); Christiane Féral-Schuhl, *Cyber Surveillance at Work*, UNI GLOBAL UNION, 22, *available at* http://www.uniglobalunion.org/Apps/UNIPub.nsf/vwLkpById/F6403CF3DFEEBF01C125757C00367650/$FILE/feral-schuhl_cybersurveillance-en.pdf (last visited Apr. 29, 2011).

264. *See* Cour d'Appel [CA] [regional court of appeal] Versailles, Apr. 2, 2003, Aff. No. 02/00293 (Fr.). *Contra* McLaren v. Microsoft Corp., No. 05-97-00824-CV, 1999 WL 339015, at *4 (Tex. App. May 28, 1999) (holding no right to privacy in e-mail messages stored in a password-protected "personal" folder).

only in very limited circumstances (e.g., where an employer already has concrete suspicions of wrong-doing against particular employees),[265] subject to significant restrictions with respect to the duration, mode, and subjects of the monitoring activities.[266] Several jurisdictions worldwide, including France, the Netherlands, and Israel, require filings with data protection or labor authorities, while others, including France, Germany, Italy, the Netherlands, and China, require employers to consult or at least notify trade unions or other employee representative bodies before subjecting their employees to surveillance measures.[267]

Complaints by an employee in a country with a high level of data protection can trigger investigations and lawsuits by data protection authorities, trade unions, consumer watchdogs, and similar organizations, and can also lead to criminal complaints.[268] Employers found in non-compliance may face steep penalties, damages awards, and possibly even prison time, along with plenty of bad press, as some recent examples demonstrate.

In September 2008, German authorities ordered discount retailer Lidl to pay fines totaling around 1.5 million euros for a variety of alleged data protection violations, including monitoring employees and customers through the use of in-store hidden cameras to counter a wave of theft.[269] In

265. Astrid Wellhörner & Phillip Byers, *Datenschutz im Betrieb—Alltägliche Herausforderung für den Arbeitgeber* [*Data Protection at Work—Everyday Challenges for Employers*], 18 BETRIEBS BERATER 2310, 2311 (2009).

266. For instance, in the context of internal audits in Germany it is often necessary to inform employees in detail about the reasons for the internal audit, the controller's identity, the categories of data collected, etc. *See* Michael Schmidl, *Germany: Internal Audits and Protection of Employee Data*, 7 WORLD DATA PROTECTION REP., no. 6, 2007, at 10, 10–11 (2007).

267. *See, e.g.*, German Works Constitution Act § 87(1)(6), Sept. 25, 2001, BGBL. I at 2518, repromulgated Dec. 23, 2003, BGBL. I at 2848, art. 81 (Ger.), *available at* http://www.bmwi.de/English/Redaktion/Pdf/__Archiv/labour-law/works-constitution-act1,property=pdf,bereich=bmwi,sprache=de,rwb=true.pdf.

268. *See, e.g.*, STRAFGESETZBUCH [STGB] [PENAL CODE] § 202a (Ger.) (illegally spying on data—with a maximum penalty of three years' imprisonment); *id.* § 206 (telecommunication secrecy—with a maximum penalty of five years' imprisonment); *id.* § 303a (deleting or changing data—with a maximum penalty of two years' imprisonment); BUNDESDATENSCHUTZGETZ [BDSG] [FEDERAL DATA PROTECTION ACT], Jan. 14, 2003, § 44 (Ger.) (with a maximum penalty of two years' imprisonment).

269. *Millionen-Strafe für die Schnüffler* [*A Penalty of Millions for Snoops*], SUEDDEUTSCHE, Sept. 11, 2008, *available at* http://www.sueddeutsche.de/wirtschaft/860/309795/text/. As a result of this and similar incidents, the degree of possible fines and penalties based on the German Federal Data Protection Act increased on November 1, 2009. *See* DEUTSCHER BUNDESTAG: BESCHLUSSEMPFEHLUNG UND BERICHT [DECISION AND RECOMMENDATION

2009, Deutsche Telekom came under scrutiny when the company admitted to having collected and reviewed telephone call data of its directors and executives in order to investigate management irregularities.[270] Deutsche Telekom reacted by creating a management board position dedicated to data privacy and security matters.[271] Despite a historic emphasis on data protection, even companies based in Europe struggle with privacy compliance. This suggests that it is imperative for U.S. companies with operations abroad to obtain legal advice on the implications of their contemplated monitoring activities under the laws of all jurisdictions in which affected employees are located.

In Europe, public companies are not required or encouraged to establish whistleblower hotlines, monitor employees, or conduct investigations. In fact, employers must obtain various authorizations from national authorities, which tend to require that electronic monitoring programs protect employee data privacy.[272] Employment contracts in Europe are also not "at will" agreements, and employees are protected against termination more generally. Consequently, employers are less exposed to vicarious liability claims based on employee wrong-doings, perhaps because European laws seem to acknowledge employers' lesser degree of control over their employees' communications and other activities.[273]

TO MODIFY THE FEDERAL DATA PROTECTION ACT] [BT] 16/13657 (Ger.), *available at* http://dip21.bundestag.de/dip21/btd/16/136/1613657.pdf.

270. *Telekom Bespitzelte Aufsichtsräte, Manager und Journalisten* [*Telecom Spied on Board of Directors, Managers, and Journalists*], SPIEGEL ONLINE (May 24, 2008) (Ger.), http://www.spiegel.de/wirtschaft/0,1518,555148,00.html.

271. *See Manfred Balz*, DEUTSCHE TELEKOM [T-MOBILE], http://www.telekom.com/dtag/cms/content/dt/en/579544 (last visited July 12, 2011).

272. Melissa Klein Aguilar, *Finally: German Whistleblower Guidelines Released*, COMPLIANCE WK., May 1, 2007, *available at* http://www.eapdlaw.com/files/News/684ef9c3-942d-4a1a-a43e-5e91edd73573/Presentation/NewsAttachment/3b385aac-c8cf-4165-9f64-67ee78bc64a4/Finally_German%20Whistleblowers%20GuidlinesReleased_pdf.pdf; Cynthia Jackson, *A Global Whistle-Stop Tour*, DAILY J., Feb. 19, 2009, *available at* http://www.bakermckenzie.com/files/Publication/b3442009-d314-4585-a396-f1ec419acc6e/Presentation/PublicationAttachment/4840fa55-490c-448a-983f-fbdb28b9f7f5/ar_sfpa_DJ8GlobalWhistleStopTour_feb09.pdf.

273. It is difficult to prove a negative, but the authors note a dearth of reported cases on employer liability for harassment or unlawful contact of employees from European jurisdictions.

IV. DIFFERENCES IN POLICY, LAW, AND PRACTICE—AND THE IMPACT ON GLOBAL EMPLOYERS AND EMPLOYEES

True to their respective, fundamentally different approaches to data privacy and employment relations in principle, the United States and the European Union offer entirely different parameters for workplace privacy and employer monitoring—in law and practice. In the United States, some state statutes increasingly seek to protect employees' privacy rights from overly intrusive monitoring; however, for the most part key differences between the U.S. and European privacy regimes still exist. As such, global employers must be cognizant of the two contrasting privacy regimes.

In the United States, privacy is legally protected only where an actual and reasonable expectation of privacy exists. Employers are free to eliminate actual employee privacy expectations through detailed, specific notices and deploy even highly intrusive monitoring technologies, except where prohibited by a few, narrowly worded statutory prohibitions of extremely intrusive employer monitoring in some states (such as video surveillance in locker rooms and restrooms).[274] Courts could apply increasingly higher requirements for the level of detail required to allow employer notices and find limited expectations of privacy where employer notices are outdated or incomplete. But, many U.S. courts have interpreted notices broadly in the employer's favor and found either no actual or no reasonable privacy expectations where employers pursued legitimate interests with their monitoring efforts.

In Europe, companies are generally prohibited from collecting and processing personal data under data protection laws that are intended to

274. Some state laws and state courts have begun to consider privacy claims in working environments and employee privacy has gained ground in the United States. N.Y. GEN. BUS. LAW § 395-b (2010) bans the use of surveillance devices—whether they're video or conventional "peep-hole" types—by business owners, and covers areas such as changing rooms or areas, bathrooms, and any other place where a reasonable expectation of personal privacy exists. Employers who violate this law are subject to fines up to $300, fifteen days in jail, or a combination of fines and time served. New York State has passed an "eavesdropping" statute similar to the Federal Wiretapping Statutes as well. *See* N.Y. PENAL LAW §§ 250.00, 250.05 (2010). A person is guilty of the felony of eavesdropping when he or she unlawfully engages in, *inter alia*, "wiretapping" or "intercepting or accessing of an electronic communication." In California it is a crime to intercept or eavesdrop upon any confidential communication, including a telephone call or wire communication, without the consent of all parties. CAL. PENAL CODE §§ 631–632 (2010). The appellate court has ruled that using a hidden video camera in a private place violates the statute. California v. Gibbons, 263 Cal. Rptr. 905 (Ct. App. 1989).

minimize the existence of personal data.[275] Employers are not required or encouraged to deploy intrusive monitoring technologies. Employees can freely deny or revoke consent to monitoring programs and their consent is presumed to be invalid as coerced, unless employers can prove that employees consented voluntarily (i.e., are given the option to say "no" without adverse consequences), which in practice limits or precludes the implementation of monitoring technologies altogether.

Global employers therefore have to navigate the contrasting legal environments carefully. There are ample opportunities for pitfalls and difficulties, for example, in connection with the global deployment of e-mail and web servers as well as anti-spam and virus protection filters, investigations into potential wrong-doings involving employees in multiple jurisdictions, management of multi-country teams and reporting lines, or short- and long-term secondments of employees.

Practical options for global employers include the following three approaches:

(1) Country-specific monitoring protocols tailored to local requirements and permissions. A multinational enterprise could determine, on a country-by-country basis, how much monitoring is necessary and permissible, and then develop information security and monitoring policies that are optimal for the particular jurisdiction. But this approach severely limits the ability to maintain global systems and policies, involves significant costs (for legal research, system design, and compliance maintenance), and does not even guarantee full compliance or optimal compromises for situations that involve several jurisdictions (e.g., investigations into alleged illegal practices that involve employees of several different subsidiaries, or concerns regarding harassment across borders where employees in one country e-mail offensive materials to employees in another).

(2) Reducing global surveillance to the standards permissible in the most restrictive jurisdiction. For example, a company with presences in the United States, the United Kingdom, and Germany could deploy only monitoring technologies and processes that comply with German data protection laws.

275. The collection, processing, and use of data is governed by the principles of data avoidance and data economy. In the interest of collecting as little data as possible, personal data shall only be collected to the extent required for the purposes of processing the relationship between the parties. For instance, the Federal Data Protection Act states the "data omission and data parsimony" principle, ensuring that no or as little as possible person-related data is collected, processed, and used. BUNDESDATENSCHUTZGETZ [BDSG] [FEDERAL DATA PROTECTION ACT], Jan. 14, 2003, as amended, § 3a (Ger.). Consequently the technical infrastructure shall already minimize the amount of collected, processed, and used data. This data shall be kept in anonymous or pseudonymous form if possible.

This approach enables globally uniform systems and should help minimize potential exposure to employee privacy claims. However, this approach may leave the multinational enterprise unreasonably exposed to liability arising from employee misconduct in jurisdictions where monitoring is permissible and expected by governments and in courts applying due diligence standards, such as the United States.

(3) Regional combination approaches whereby intrusive technologies are deployed in the United States and jurisdictions with similarly lenient privacy laws, whereby restrictive jurisdictions are excluded. This approach tries to mitigate the disadvantages of the first option (high costs, fragmented systems and processes) and second option (undue exposure in jurisdictions where monitoring is permissible and expected) by differentiating on a regional basis or "country category basis." The global enterprise could develop policies that implement two or more levels of employee monitoring for certain jurisdictions or regions. This approach offers the comfort of remaining relatively close to local practices and requirements without investing extensively in legal research and country-specific systems and processes. But, as with any compromise, this approach tends to involve some trade-offs; for instance, without closely analyzing local requirements, companies cannot be sure that their practices fully comply with applicable law.

Global employees also have choices. In the United States, they can look for employers with less intrusive monitoring policies and quit when they receive notice that the policies have changed. If enough employees are sufficiently concerned, employer policies can be expected to change according to the dynamics of the labor marketplace. In the meantime, the employees may depart in favor of a different working situation. If the employees move to Europe, they will find a different legal environment. European employers cannot rely on notices that destroy the employees' privacy expectations and employees can freely deny or revoke consent to monitoring and surveillance at any time.

Thus, in Europe, employees have (but do not need, as a condition for legal protection) reasonable privacy expectations, whereas in the United States, employees currently do not have (but need, as a condition for legal protection) reasonable privacy expectations.

THE CASE FOR LIBERAL SPECTRUM LICENSES: A TECHNICAL AND ECONOMIC PERSPECTIVE

Thomas W. Hazlett[†] *& Evan T. Leo*[††]

TABLE OF CONTENTS

---

† Professor of Law & Economics and Director, Information Economy Project, George Mason University. Professor Hazlett received his Ph.D. (Economics) from the University of California, Los Angeles; formerly held faculty positions at the University of California, Davis, Columbia University, and the Wharton School; and served as Chief Economist of the Federal Communications Commission.

†† Partner, Kellogg, Huber, Hansen, Todd, Evans & Figel, P.L.L.C. Mr. Leo received his J.D. from The George Washington University. The authors express their appreciation to Peter Huber and Charles Jackson for substantive technical comments and thank Mary Ann Endo for invaluable research assistance. All liability remains with the authors.

I. INTRODUCTION

The traditional system of radio spectrum allocation has inefficiently restricted wireless services. Alternatively, liberal licenses that cede de facto spectrum ownership rights yield incentives for operators to maximize airwave value. These authorizations have been widely used for mobile services in the United States and internationally, leading to the development of highly productive services and waves of innovation in technology, applications, and business models.

Serious challenges to the efficacy of such a spectrum regime have arisen, however. While the world marvels at the emergence of vast wireless networks, now serving over five billion global subscribers,[1] many leading policy advocates in the United States have concluded that ceding de facto ownership of the airwaves through cellular licenses is a barrier to innovation and social progress. Advancing the examples of cordless phones, Wi-Fi, and Bluetooth, they credit unlicensed bandwidth—spectrum without exclusive ownership rights—as a Petri dish for disruptive technologies. These advocates extrapolate from examples to produce a template for restructuring the airwaves. Exclusive spectrum rights are obsolete, they claim; expanding "spectrum commons" would be more economically productive. Federal regulators have begun accepting this argument, shifting policies to favor allocations of unlicensed spectrum.

The marketplace, however, demonstrates that spectrum scarcity is alive and well. Costly conflicts over airwave use not only continue but have intensified with scientific advances that dramatically improve the functionality of wireless devices and so increase demand for spectrum access. These developments have increased the importance of social coordination in the use of wireless technologies by intensifying the competition between rival, mutually exclusive employments of frequencies. Alternative property

1. *Mobile Broadband Subscriptions To Hit One Billion in 2011*, INT'L TELECOMM. UNION (Feb. 7, 2011, 11:56 AM), http://www.itu.int/ITUD/ict/newslog/Mobile+Broadband+Subscriptions+To+Hit+One+Billion+In+2011.aspx.

rules are available to guide this coordination. Selecting rules that most reliably allow spectrum to generate the highest economic gains is socially efficient.

This Article evaluates the economic and technical arguments underlying this choice of regulatory regime. We first trace the path from traditional licenses, which systematically squandered valuable wireless opportunities, to reforms creating liberal licenses. Next, we examine the claim that advanced wireless technologies can effectively eliminate spectrum scarcity and, with it, the social utility of exclusionary rules for access to airwaves. We show that interference between radio signals is real and that conflicts between rival users are expensive. To productively use spectrum inputs for one set of applications or technologies constrains what such inputs can supply for alternatives. New and improved spectrum-sharing technologies do not eliminate these trade-offs but instead increase the value of communications, *further exacerbating* the competition for airspace. Overwhelming marketplace evidence demonstrates that liberal licenses promote beneficial social coordination, uniquely shifting spectrum to innovative uses, organizing investment in large-scale network infrastructure, and creating complex contracts permitting intensive spectrum sharing. Indeed, exclusive frequency rights are so broadly accommodating that they efficiently supply "spectrum commons," just as public parks are most productively provided within the context of private ownership of real estate.

II. RIVAL SPECTRUM MODELS

The U.S. mobile phone industry has achieved remarkable success. More than 302.9 million Americans—roughly ninety-six percent of the population—purchase wireless service.[2] The nation's wireless carriers spend over $24 billion a year[3] building network infrastructure; about $22 billion[4] more is spent on handsets and other wireless devices. U.S. companies like Qualcomm and Motorola have developed cutting-edge wireless technologies sold throughout the world. Firms like Apple, Palm, and Research in Motion (Blackberry) have assumed leading positions as device and application suppliers without owning wireless infrastructure by contracting with carriers

2. CTIA—The Wireless Ass'n ("CTIA"), *Wireless Quick Facts* (2010), http://www.ctia.org/media/industry_info/index.cfm/AID/10323. The term "wireless service" is used interchangeably with "mobile service," "cellular service," and "CMRS" (Commercial Mobile Radio Service, the FCC's official service designation) in this Article.

3. *Id.*

4. Consumer Elecs. Ass'n ("CEA"), CEA Historical Sales Data (2009) (on file with authors) [hereinafter CEA Database] (totaling 2009 retail cell phone sales ($8.6 billion) and smartphone sales ($13.6 billion)).

who do. Application providers such as Yahoo!, Google, Twitter, and ESPN, while also lacking wireless assets, have likewise been able to reach mass market audiences through partnerships with wireless firms. The U.S. wireless industry as a whole generates $160 billion in revenues per year[5]—more than broadcast and cable television combined.[6] According to conservative estimates, the industry creates more than $150 billion per year in additional consumer benefits.[7]

Radio spectrum is a key input to the wireless industry. Licenses issued by the Federal Communications Commission ("FCC") enable firms to supply services via designated airwaves; the nature of the spectrum rights the FCC grants affect the volume, quality, cost, and scope of services that can be provided to customers. Through 2008, mobile networks could access only about 194 MHz,[8] just seven percent of the prime bandwidth below 3 GHz (the range most economically viable for broadcasting and mobile services, commonly considered "beachfront property").[9] In September 2006, Advanced Wireless Service ("AWS") licenses representing an additional 90 MHz of frequency space in the 1.7 GHz and 2.1 GHz bands were auctioned, with the U.S. Treasury collecting $13.7 billion.[10] These frequencies were

5. CTIA, *supra* note 2.

6. *Compare* CTIA, *supra* note 2, *with* Agata Kaczanowska, *Television Broadcasting in the US*, IBISWORLD, 3 (Feb. 2011), http://clients.ibisworld.com/industryus/default.aspx?indid=1261 (estimating that television broadcasting revenue totaled $36.1 billion in 2010), *and* Nat'l Cable & Telecomms. Ass'n, *Industry Data*, http://www.ncta.com/Statistics.aspx (last visited Sept. 2, 2011) (estimating cable revenue totaled $89.9 billion in 2009).

7. Jerry A. Hausman, *Cellular 3G Broadband and WiFi*, *in* FRONTIERS OF BROADBAND, ELECTRONIC AND MOBILE COMMERCE 9, 11 (R. Cooper & G. Madden eds., 2004); *see also* Roger Entner, Ovum, *The Increasingly Important Impact of Wireless Broadband Technology and Services on the U.S. Economy*, CTIA, 2 (2008), http://files.ctia.org/pdf/Final_OvumEconomicImpact_Report_5_21_08.pdf (estimating that by 2016 the wireless industry will help bring about $427 billion per year in productivity gains, more than the productivity gains currently provided by the motor vehicle and pharmaceutical industries combined).

8. RYSAVY RESEARCH, MOBILE BROADBAND AND SPECTRUM DEMAND 23 (2008), http://www.rysavy.com/Articles/2008_12_Rysavy_Spectrum_Demand_.pdf. For a detailed description of mobile (and other) allocations under 3 GHz, see Evan Kwerel & John Williams, *A Proposal for a Rapid Transition to Market Allocation of Spectrum* (Fed. Commc'ns Comm'n, Working Paper No. 38, 2002), *available at* http://wireless.fcc.gov/auctions/conferences/combin2003/papers/masterevanjohn.pdf.

9. Om Malik, *700 MHz Explained in 10 Steps*, GIGA OM (Mar. 14, 2007, 6:30 AM), http://gigaom.com/2007/03/14/700mhz-explained/ ("Due to its broadcast-attractive physics (like its ability to penetrate walls), this spectrum is desirable for both broadband communications in general and public-safety uses in particular.").

10. Public Notice, Fed. Commc'ns Comm'n ("FCC"), Auction of Advanced Wireless Services Licenses Closes (Sept. 20, 2006), http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-06-1882A1.pdf. Due to regulatory lags, AWS bandwidth was not generally available to licenses until well into 2007 or even 2008. There is also a lag between the time

encumbered by a wide range of government users, were not generally available to licensees through 2008,[11] and are gradually being deployed.[12] In March 2008, the FCC conducted further auctions for rights to use 52 MHz in the 700 MHz band, collecting another $18.958 billion in winning bids.[13] The 700 MHz frequencies were occupied, in part, by analog TV broadcast stations that were switched off June 12, 2009 as part of the transition to digital television.[14] These frequencies are being deployed in emerging Fourth Generation ("4G") wireless services, including Long Term Evolution ("LTE"), offering far higher data speeds and capacities than existing wireless broadband networks.[15]

licenses are assigned and networks are built. T-Mobile, the largest AWS bidder, first began serving customers using these frequencies in May 2008. *See* Katherine Noyes, *T-Mobile's 3G Network Touches Down in NYC*, TECHNEWSWORLD (May 5, 2008), http://www.technewsworld.com/story/62876.html?wlc=1235087208.

11. *See* RYSAVY RESEARCH, *supra* note 8, at 23 & n.48.

12. Comments of T-Mobile USA, Inc. to the Nat'l Telecomms. and Info. Admin. at 1–2, Relocation of Federal Systems in the 1710–1755 Frequency Band, Review of the Initial Implementation of the Commercial Spectrum Enhancement Act, No. 0906231085-91085-01 (U.S. Dep't of Commerce Aug. 21, 2009), *available at* http://www.ntia.doc.gov/comments/2009/CSEA/T-Mobile_CSEA_NOI_Comments_8-21-09.pdf.

13. *Auction 73: 700 MHz Band*, FED. COMMC'NS COMM'N (Feb. 10, 2009), http://wireless.fcc.gov/auctions/default.htm?job=auction_summary&id=73; *Revised 700 MHz Band Plan for Commercial Services*, FED. COMMC'NS COMM'N (Sept. 5, 2007), http://wireless.fcc.gov/auctions/data/bandplans/700MHzBandPlan.pdf.

14. In 1997 the FCC assigned each TV station a new digital TV broadcasting license, placing the digital stations on channels 2–51 to allow the remaining channels assigned to TV broadcasting (52–69) to be later reallocated. By regulatory mandate, stations were broadcasting in digital formats on their new digital channel assignments by 2002. The end of analog broadcasting on TV Channels 52–69 had been mandated by Congress to occur December 31, 2006, but the mandate contained conditions unlikely to be met in the vast majority of markets. Long delays were anticipated. Hence, in the Digital Transition and Public Safety Act of 2005, Congress set February 17, 2009 as the new analog switch-off date. Just days before the deadline, however, Congress, responding to a request from the new Obama administration, voted to delay the switch-off until June 12, 2009. This deadline held. Ending analog TV broadcasting on channels 52–69 made 108 MHz (6 MHz per channel) available for reallocation. Some 70 MHz of this "digital dividend" was allocated to liberal licenses auctioned by the FCC in 2002, 2003, and 2008. For an analysis of how the analog switchover to digital television worked, see Matthew Lasar, *A Year of Digital TV: Who Won the Transition?*, ARS TECHNICA (June 15, 2010, 8:45 AM), http://arstechnica.com/telecom/news/2010/06/a-year-of-digital-tv-who-won-the-transition.ars.

15. *See* Marguerite Reardon, *The 411 on AT&T's 4G Strategy (FAQ)*, CNET NEWS (Dec. 21, 2010, 4:00 AM), http://news.cnet.com/8301-30686_3-20026253-266.html ("Currently, AT&T is using 700 MHz spectrum holdings as well as spectrum it acquired in the FCC's Advanced Wireless Services spectrum auction."); Maravedis, *Verizon vs. Clearwire: A 4G Comparison*, LTE WORLD (Dec. 8, 2010), http://lteworld.org/blog/verizon-vs-clearwire-4g-comparison (stating that Verizon's "LTE network will be [deployed using] . . . its spectrum resources of 34 MHz in the 700 MHz band").

A. LIBERAL LICENSES

Much of the case for unlicensing spectrum begins with the argument that broadcast television wastes valuable radio spectrum and that an unlicensed regime can make more productive use of frequencies.[16] For empirical support, commons advocates rely heavily on the success of Wi-Fi, which is now widely used in local area networks, public hot spots, and other applications. Wi-Fi, as one commons advocate explains, is "the most prominent unlicensed wireless technology available today" and "a great case study for the impact of dynamic wireless technologies."[17] Introduced in the late 1990s,[18] Wi-Fi radios now provide high-speed, digital connections to millions of users using unlicensed bandwidth.

TV broadcasters, by contrast, provide video service via exclusive licenses. But TV band airwaves are dramatically under-utilized, littered with "white spaces" where little to no communications travel. Indeed, *all* over-the-air TV reception could be transferred to cable and/or satellite TV systems at a small fraction of the cost of the TV airwaves that would be released for more valuable services.[19] Thus, exclusive spectrum rights impede innovation and promote inefficient use of the airwaves.

Both ends of the comparison are confused. Broadcast TV licenses are locked into inefficient market structures precisely because of "command and control" regulation that economists have long condemned as "Gosplan."[20]

---

16. Paul Baran, Symposium Paper, Is the UHF Frequency Shortage a Self Made Problem? (Marconi Centennial Symposium, Bologna, Italy, June 23, 1995), *available at* http://www.interesting-people.org/archives/interesting-people/199507/msg00023.html.

17. KEVIN WERBACH, RADIO REVOLUTION: THE COMING AGE OF UNLICENSED WIRELESS 22 (2003), *available at* http://werbach.com/docs/RadioRevolution.pdf; *see also* Fed. Commc'ns Comm'n, *Connected & On the Go: Broadband Goes Wireless* 5 (Feb. 2005), *available at* http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-256693A1.pdf [hereinafter FCC Task Force Paper] (noting the tremendous success of Wi-Fi devices); Yochai Benkler, *Some Economics of Wireless Communications*, 16 HARV. J.L. & TECH. 25, 30 (2002) (noting the "breathtaking growth" of Wi-Fi devices that "rely on utilizing frequencies that no one controls").

18. *See generally* Kevin Negus & Al Petrick, *History of Wireless Local Area Networks (WLANs) in the Unlicensed Bands*, 11 INFO, no. 5, 2009, at 35.

19. Richard H. Thaler, *The Buried Treasure in Your TV Dial*, N.Y. TIMES, Feb. 28, 2010, at BU7; Thomas W. Hazlett, *The U.S. Digital TV Transition: Time To Toss the Negroponte Switch* (AEI-Brookings Joint Ctr. for Regulatory Studies, Working Paper No. 01-15, 2001) [hereinafter Hazlett, *U.S. Digital TV Transition*]; Thomas W. Hazlett, *Transition to Yesterday: Subsidizing the Killer App of 1952*, ARS TECHNICA (Nov. 3, 2008, 10:40 PM), http://arstechnica.com/old/content/2008/11/dtv-transition-to-yesterday.ars [hereinafter Hazlett, *Transition to Yesterday*].

20. Gerald R. Faulhaber & David J. Farber, *Spectrum Management: Property Rights, Markets, and the Commons* 6 (AEI-Brookings Joint Ctr. for Regulatory Studies, Working Paper No. 02-

Ronald H. Coase's classic critique of the FCC focused on these licenses, leading him to recommend adoption of private property rights in spectrum to replace "public interest" assignments by regulators.[21] Coase's proposal was considered radical. When Coase explained his proposal at a 1959 FCC hearing, the first question a Commissioner posed was, "Is this all a big joke?"[22]

It was not. Were licensees able to freely control the spectrum allocated to their licenses, rather than being granted narrow rights to transmit in ways regulators specified, these de facto property owners would have naturally sought to expand value creation—to make productive use of "white spaces." It took a quarter century for regulators to start (implicitly) embracing Coase's proposal. Beginning with licensing of cellular phone service in the 1980s, carriers were given far more control over the use of their spectrum as compared to radio and TV broadcasting licensees.[23]

Why the shift? For one thing, the political interest in regulating cellular was lower because the content transmitted over cellular networks is private, whereas broadcasting is inherently public, transmitting content influencing social, cultural, and political developments. For another thing, the cost of regulating cellular was much higher. Instead of a single, one-way transmitter, cellular systems involve the integration of thousands of base stations and millions of handsets, each a receiver and a transmitter moving in space. The increasingly liberal licenses granted to mobile phone operators have enabled firms to design their own services, adapt new technologies, determine network architectures, and experiment with new business models as profit criteria dictate—a radical departure (for allocated spectrum)[24] from the traditional broadcast license.

The historical broadcast license is quite distinct from a liberal license. Broadcast licensees are endowed with extremely delimited property rights in spectrum; their specific use permits authorize only a broadcasting service operated according to fixed technical standards and pre-specified service definitions. Licensees cannot allocate the spectrum allocated to their licenses

12, 2002). The term references bureaucratic management techniques for industrial markets in the Soviet Union. *Id.* at 6.

21. Ronald H. Coase, *The Federal Communications Commission*, 2 J.L. & ECON. 1 (1959).

22. Ronald H. Coase, *Comment on Thomas W. Hazlett*, 41 J.L. & ECON. 577, 579 (1998).

23. Thomas W. Hazlett & Matthew L. Spitzer, *Advanced Wireless Technologies and Public Policy*, 79 S. CAL. L. REV. 595, 623–31 (2006).

24. The reforms begun with cellular licensing did not constitute full liberalization in that only modest amounts of bandwidth had been allotted to liberal licenses. More widespread reforms have occurred in other countries. *See* Thomas W. Hazlett, *Property Rights and Wireless License Values*, 51 J.L. & ECON. 563 (2008).

to higher valued uses, even if alternatives are more profitable. The gross inefficiencies that result are not due to licensing, but to truncating private ownership of spectrum rights. As Coase recognized, granting licensees ownership of spectrum, as in broad, "flexible use" permits,[25] allows market forces to divert spectrum resources to their most socially valuable employments.[26]

These market forces can be seen at work in recent FCC sales of new wireless licenses. These sales will bring the total amount of licensed spectrum available to mobile carriers in the United States up to levels comparable to those in the European Union. By 2001, EU regulators had issued mobile licenses allocating an average of about 266 MHz per country, about fifty percent higher than the amount then allocated in the United States.[27] The recent AWS and 700 MHz auctions bring the U.S. total up to about 360 MHz available for mobile service,[28] but EU countries are now preparing to

25. Kwerel & Williams, *supra* note 8, at 3–4 (using the term "flexible use" as a proxy for exclusive spectrum rights). Hazlett and Spitzer expand the term to EAFUS: "exclusively-assigned, flexible-use spectrum." Hazlett & Spitzer, *supra* note 23, at 623.

26. With colleagues, Coase outlined a detailed policy proposal for private property rights in spectrum in 1962, a considerable feat given the relative dearth of market data to rely on. *See* Ronald Coase, William H. Meckling & Jora Minasian, Problems of Radio Frequency Allocation (Sept. 1995) (unpublished manuscript originally written in 1963), *available at* http://www.rand.org/pubs/drafts/2008/DRU1219.pdf. The efficiency of the approach suggested has, indeed, proven successful as per many economic analyses. Perhaps the most influential source is the 2002 report on spectrum policy commissioned by the government of the United Kingdom. MARTIN CAVE, REVIEW OF RADIO SPECTRUM MANAGEMENT (2002), *available at* http://www.ofcom.org.uk/static/archive/ra/spectrum-review/2002review/1_whole_job.pdf.

27. Thomas W. Hazlett & Roberto E. Muñoz, *Spectrum Allocation in Latin America: An Economic Analysis*, 21 INFO. ECON. & POL'Y 261, 262 (2009). Note that the smaller U.S. allocations cannot be explained by cross-country size or cost disparities. A reduction in available bandwidth means that, in any given market, wireless operators have less capacity, and slower average speeds, to accommodate a given number of customers, all else (including network infrastructure investments) equal.

28. *See* Blair Levin et al., What 700 MHz Winners Can Do with Their Spectrum 4 (Apr. 15, 2008) (unpublished manuscript) (on file with authors). These totals do not include spectrum to "fixed" broadband services, such as the 3.5 GHz band in Europe, and the 2.5 GHz band in the United States. While these wireless services are being adapted for mobile use, the migration—undoubtedly promising as a future source of competition—is yet nascent. Clearwire is developing a nationwide wireless broadband network using WiMax technology delivered over 2.5 GHz frequencies. There is potentially some 190 MHz available for use there, divided between EBS (Educational Broadband Services) licenses held by non-profit (mostly educational) institutions and BRS (Broadband Radio Service) licenses held by commercial operators. Transaction costs associated with re-aggregating the dispersed, truncated, and conflicting transmission rights have been formidable. *See* Thomas W. Hazlett, *Spectrum Tragedies*, 22 YALE J. ON REG. 242, 258 (2005). As of year-end 2010, Clearwire reported 4.4 million U.S. subscribers. Dan Meyer, *Clearwire Posts Strong Q4 Wholesale Growth,*

make major new allocations that will extend the advantage.[29] To meet the perceived U.S. deficit, the FCC has announced that it "recommends making 500 MHz of spectrum newly available for broadband by 2020, with a benchmark of making 300 MHz available by 2015."[30] The goal is ambitious; previous experience suggests that even much smaller allocations take much longer to implement.[31]

The FCC issued cellular licenses in the United States between 1982 and 1989, with Personal Communications Services ("PCS") license awards beginning in 1995.[32] These authorizations represented a paradigm shift in the regulation of spectrum use, rejecting the traditional regime crafted for and typified by broadcasting licenses. For radio and television stations, licensees receive narrowly crafted operating permits that define the services they can provide, the technology they may employ, the physical locations where they must place transmitters, transmitter height, and the business model (advertising-based, non-subscription) they must use. A TV broadcaster cannot, for example, forgo video transmissions and instead use its licensed spectrum to provide high-speed internet service.

Cost Conservation Remains Intact, RCR WIRELESS NEWS (Feb. 17, 2011, 7:08 PM), http://www.rcrwireless.com/article/20110217/CARRIERS/110219933/1097.

29. Comments of CTIA—The Wireless Association, In the Matter of Implementation of Section 6002(b) of the Omnibus Budget Reconciliation Act of 1993, WT Docket No. 09-66, at 82–83 (Fed. Commc'ns Comm'n Sept. 30, 2009) ("Ofcom, the regulator in United Kingdom, is in the process of reallocating 355 MHz of spectrum for commercial wireless services, which would bring the U.K.'s total up to 710 MHz . . . . Similarly, in Germany, 340 MHz of spectrum has been identified for reallocation, which will bring the total up to 645 MHz . . . .").

30. FED. COMMC'NS COMM'N, CONNECTING AMERICA: THE NATIONAL BROADBAND PLAN 26 (2010), *available at* http://download.broadband.gov/plan/national-broadband-plan-chapter-2-goals-for-high-performance-america.pdf.

31. *See id.* at 79. The six- to thirteen-year lags estimated by the FCC are conservative. For instance, the estimated length of the original cellular allocation was given to be eleven years. In fact, AT&T first petitioned the FCC for cellular airwaves in 1958. Tom Farley, *The Cell-Phone Revolution*, 22 AM. HERITAGE INVENTION & TECH. 23 (2007). Yet, cellular licenses were not distributed until lotteries held between 1984 and 1989 were completed. Thomas W. Hazlett & Robert J. Michaels, *The Cost of Rent Seeking: Evidence from the Cellular Telephone License Lotteries*, 59 S. ECON. J. 425 (1993). Hence, the actual regulatory lag may well have been closer to thirty years than to the National Broadband Plan's estimate of eleven.

32. Thomas W. Hazlett, *Assigning Property Rights to Radio Spectrum Users: Why Did FCC License Auctions Take 67 Years?*, 41 J.L. & ECON. 529, 532, 535 (1998). Note that while the United States was the first country to widely assign cellular licenses (for analog voice services), EU countries issued 2G (Second Generation) licenses (for digital voice services) between 1989 and 1992. *See* LAURENT BENZONI & EVA KALMAN, THE ECONOMICS OF RADIO FREQUENCY ALLOCATION (1993). This was years ahead of the comparable U.S. allocation for PCS, which began assigning licenses in 1995. *See* Peter Cramton, Evan Kwerel & John Williams, *Efficient Relocation of Spectrum Incumbents*, 41 J.L. & ECON. 647, 661 (1998).

Wireless phone licenses—and particularly PCS licenses—were the first major implementation of two fundamental policy innovations: (1) awarding licenses by competitive bidding, thus abandoning assignments by regulatory fiat or lottery;[33] and (2) permitting licensees wide discretion in using allotted frequency space.[34] Auctions were efficiency enhancing,[35] but the latter policy was of much greater significance for consumer welfare.[36] With licensees free to choose services, technologies, architectures, and business models, market forces could, for the first time, optimize radio spectrum use. The emergence of vigorous economic activity, including high levels of network investment, led regulators to adopt liberal licenses as the new standard for wireless services.

B. UNLICENSED

In parallel to the evolution of the liberal license model, a second distinct FCC policy regime was developing. Traditionally, the FCC set aside bands for "unlicensed" use by low power, short-range radios—remote controls, short-range security systems, and baby monitors, for example. The FCC has also allocated such unlicensed bands for the use of non-communications devices, like microwave ovens and medical or scientific equipment that emit radiation, which potentially conflicts with the use of communications systems. The first unlicensed bands were established in 1938.[37] In 1985, however, the FCC took a decisive step in authorizing a whole class of new unlicensed devices, thus eliminating the process of regulatory pre-approval under vague "public interest" criteria. This deregulatory initiative, which

33. License auctions were first authorized in Section 6002 of the Omnibus Budget Reconciliation Act of 1993, Pub. L. No. 103-66, § 6002, 107 Stat. 312, 387–97. The first spectrum auctions were for PCS awards. Lotteries had been used to assign cellular licenses. *See generally* JAMES B. MURRAY, JR., WIRELESS NATION (2001).

34. For example, cellular licenses were originally issued with a mandate that operators use the specific analog transmission format AMPS; there was no mandated format for digital PCS licenses.

35. Peter Cramton, *Spectrum Auctions*, *in* 2 HANDBOOK OF TELECOMMUNICATIONS ECONOMICS 605, 607 (Sumit Majumdar et al. eds., 2002).

36. *See* Thomas W. Hazlett & Roberto E. Muñoz, *A Welfare Analysis of Spectrum Allocation Policies*, 40 RAND J. ECON. 424, 437–38 (2009).

37. *See* Kenneth R. Carter, Ahmed Lahjouji & Neal McNeil, *Unlicensed and Unshackled: A Joint OSP-OET White Paper on Unlicensed Devices and Their Regulatory Issues* 6 (Fed. Commc'ns Comm'n, OSP Working Paper No. 39, 2003), *available at* http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-234741A1.pdf. Amateur bands, including CB (citizens' band) frequencies, are organized on a similar basis to unlicensed bands, although radio operators are technically required to be licensed by the FCC. The requirement is loosely enforced. Moreover, licenses help to assure that users comply with FCC rules and do not cede control of spectrum space. In this sense, unlicensed (amateur) bands have been in use since before the Radio Act of 1927.

began under President Jimmy Carter's FCC and was implemented under President Ronald Reagan's, aimed to reduce barriers to entry for new technologies.[38] In place of a case-by-case regulatory process, the FCC set forth technical criteria, including power limits, to which new "Part 15" unlicensed devices would need to adhere.[39] This policy paved the way for widespread use of unlicensed devices in the so-called Industrial, Scientific, and Medical ("ISM") frequencies in the 900 MHz band (26 MHz), the 2.4 GHz band (83.5 MHz), and at 5.8 GHz (125 MHz). In subsequent years, the market introduced thousands of unlicensed devices under the "Part 15" framework, including cordless phones and Wi-Fi radios connecting computers in local area networks.[40] One of the lead FCC engineers who worked on the regulatory initiative recounts that such devices were neither planned nor anticipated.[41]

In recent years, the FCC has moved aggressively to allocate more bandwidth to unlicensed (or "license-exempt") spectrum.[42] In 1985, there

38. *See* Kenneth R. Carter, *Unlicensed to Kill: A Brief History of the Part 15 Rules*, 11 INFO, no. 5, 2009, at 8; Michael J. Marcus, *Wi-Fi and Bluetooth: The Path from Carter and Reagan-Era Faith in Deregulation to Widespread Products Impacting Our World*, 11 INFO, no. 5, 2009, at 19; Negus & Petrick, *supra* note 18.

39. *See* 47 C.F.R. §§ 15.1–.717 (2010). In addition to the limiting technical constraints, Part 15 requires that an operator accept whatever interference is received and correct whatever interference is caused. Should harmful interference occur, the operator is required immediately to correct the interference problem, even if correction of the problem requires ceasing operation of the Part 15 system causing the interference. *See* Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission Systems (*Revision of Part 15*), 17 FCC Rcd. 7435, ¶ 6 n.2 (2002).

40. Devices are regulated under three methods at the FCC: verification, declaration of conformity, and certification. The latter category relies largely on tests performed by private firms. *See Equipment Authorization (EA)*, FED. COMMC'NS COMM'N, http://www.fcc.gov/oet/ea/procedures.html#sec1 (last visited Jan. 21, 2011).

41. Marcus, *supra* note 38, at 29–30 ("In the 1981–85 period when these rules were drafted, wireless LANs were not a common topic of discussion. Indeed, Ethernet and other LAN installations were rare outside technical organizations and unheard of in homes. The deliberations had raised the possibility of 'wireless data terminals' as an example, but did not specifically 'tilt' in favor of this application in the resulting rules. The Carter and Reagan era faith in deregulation laid the foundation for the future development of a variety of products without the need for government action.").

42. The rationale for this policy shift was laid out in the FCC's Spectrum Policy Task Force Report issued in November 2002 and was strongly endorsed by then FCC Chair Michael Powell, appointed by President George W. Bush. *See* Spectrum Policy Task Force, FCC Report, ET Docket No. 02-135 (Nov. 2002); *see also* Lawrence Lessig, *Technology over Ideology*, WIRED (Dec. 2004), *available at* http://www.wired.com/wired/archive/12.12/view.html?pg=5 ("When Powell took charge, most thought the FCC would quickly launch massive spectrum auctions. The reigning ideology was that spectrum is land, and that markets allocate land most efficiently. But Powell's FCC quickly sabotaged this idea. . . . Auctions were slowed; spectrum commons were encouraged.").

was 234.5 MHz of spectrum in the ISM bands available to unlicensed devices.[43] By 2004, the FCC had allocated approximately 665 MHz of spectrum in the same frequency range to unlicensed use. In comparison, as of that same date, about 385 MHz in this range had been allocated to liberal licenses—an unlicensed-to-licensed ratio of 1.7.[44] This tends to be substantially higher than in other countries, where the ratio is generally less than one.[45]

The U.S. government push favoring unlicensed bandwidth was further seen in an important 2005 decision. Regulators allocated a band of 50 MHz (3.65 GHz to 3.70 GHz) for terrestrial services including WiMax, an emergent wireless broadband technology often referred to as "wi-fi on steroids."[46] Despite the general use of neighboring frequencies (known generically as "the 3.5 GHz band") in international markets as licensed spectrum, and the development of WiMax radios using these airwaves by equipment makers, the FCC chose to allocate the entire band for unlicensed (non-exclusive) access.[47] This generated controversy even among the major vendors of radios for use in unlicensed spectrum, Intel and Alvarion, which had opposed the FCC's approach.[48] Then, in December 2008, in a much larger and more valuable band where the Commission sought to choose between licensed and unlicensed models, the FCC ruled that the frequencies previously set aside for TV broadcasts would be opened for the use of unlicensed devices.[49] This decision made available up to an additional 240

43. In 2002, the requirement that devices using ISM bands conform to spread spectrum formats was abolished, while power limits and other technical constraints were retained. Amendment of Part 15 of the Commission's Rules Regarding Spread Spectrum Devices, 17 FCC Rcd. 10,755 (2002).

44. *See also* Hazlett, *supra* note 28, at tbl.2 (comparing 648.5 MHz of unlicensed spectrum to 189 MHz of "flexible use" licensed spectrum).

45. *See infra* Figure 1.

46. *See, e.g.*, George Ou, *White Space Backhauls—A Penny Wise and a Pound Foolish*, DIGITAL SOC'Y (Mar. 16, 2010), http://www.digitalsociety.org/2010/03/white-space-backhauls-a-penny-wise-and-a-pound-foolish.

47. Wireless Operations in the 3650–3700 MHz Band; Rules for Wireless Broadband Services in the 3650–3700 MHz Band; Additional Spectrum for Unlicensed Devices Below 900 MHz and in the 3 GHz Band; Amendment of the Commission's Rules with Regard to the 3650–3700 MHz Government Transfer Band (*Wireless Operations in the 3650–3700 MHz Band*), 20 FCC Rcd. 6502, ¶ 25 (2005).

48. *See* Jerry Brito, *The Spectrum Commons in Theory and Practice*, 2007 STAN. TECH. L. REV. 1, ¶¶ 52, 87.

49. Unlicensed Operation in the TV Broadcast Bands; Additional Spectrum for Unlicensed Devices Below 900 MHz and in the 3 GHz Band (*Unlicensed Operation in the TV Broadcast Bands*), 23 FCC Rcd. 16,807 (2008); *see also Unlicensed Operation in the TV Broadcast Bands*, 25 FCC Rcd. 18,661 (2010).

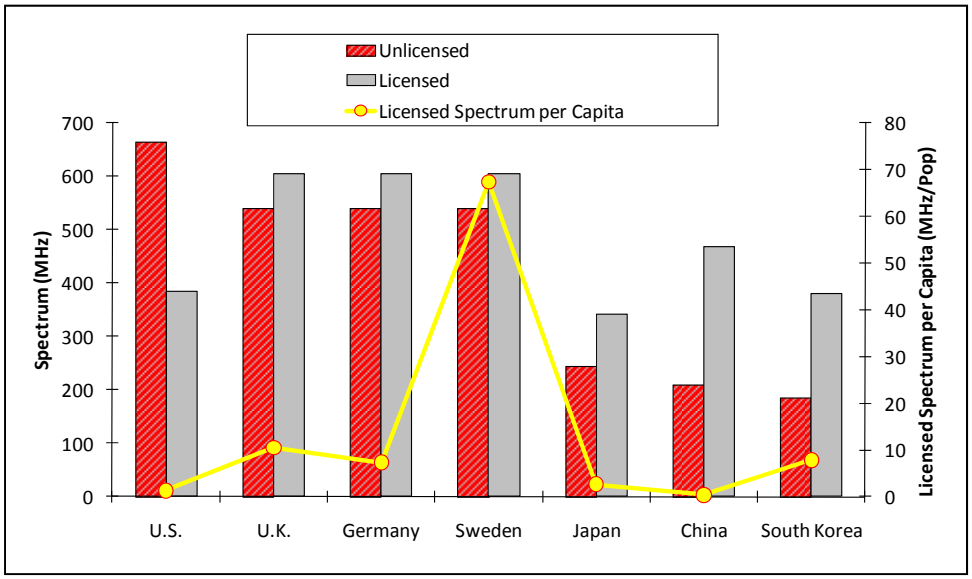
MHz of bandwidth in the median U.S. market[50] and brought the total unlicensed allocation to 955 MHz.[51] By comparison, as of year-end 2008, approximately 422 MHz had been allocated to liberal licenses, bringing the ratio of unlicensed to liberal-license spectrum to about 2.3:1.[52]

50. Under FCC rules, unlicensed devices will be permitted to access TV band spectrum not occupied by TV broadcasters. The average TV market features 8.6 stations (there are 1800 full-power stations and 210 TV markets), and each market is allocated forty-nine TV channels, allocated 6 MHz each (294 MHz total). This leaves about 240 MHz for unlicensed use. We note that the bandwidth that would actually be available for unlicensed devices is much below the 240 MHz set aside. *See* Michael Calabrese, *Broadcast to Broadband: Unlicensed Access to Unused TV Channels*, IEEE INTERNET COMPUTING, 71 (Apr. 2008), http://www.newamerica.net/files/Broadcast_to_Broadband.pdf. That is because the FCC usage restrictions, ostensibly employed to protect over-the-air TV viewers on the one side and unlicensed device users on the other, leave many unused channels in place to separate rival applications. Comments of Charles L. Jackson & Dorothy Robyn at i, *Unlicensed Operation in the TV Broadcast Bands*, ET Docket Nos. 04-186/02-380 (Fed. Commc'ns Comm'n Jan. 31, 2007).

51. This total includes 28 MHz in the 900 MHz band and 83.5 MHz in the 2.4 GHz band. Carter, Lahjouji & McNeil, *supra* note 37, at 7. It also includes 50 MHz in the 3.7 GHz band. *See Wireless Operations in the 3650–3700 MHz Band*, 20 FCC Rcd. 6502, ¶ 1. In addition, there is a total of 555 MHz of unlicensed allocation in the 5 GHz band. *See* Revision of Parts 2 and 15 of the Commission's Rules to Permit Unlicensed National Information Infrastructure (U-NII) Devices in the 5 GHz Band (*Revision of Parts 2 and 15*), 18 FCC Rcd. 24,484, ¶¶ 4–5 (2003). There is then another 240 MHz in the "white spaces" TV band discussed immediately above. *See* discussion *supra* note 50.

52. By year-end 2008, some 422 MHz of spectrum was allocated for liberal licenses, although much of it was encumbered. In particular, much of the capacity of 700 MHz licenses was blocked by TV broadcasts ongoing until the June 2009 analog switch-off. Licensed spectrum allocations are calculated as: 50 MHz (800 MHz cellular), 120 MHz (1.9 GHz PCS), 14 MHz (SMR, 1.9 GHz), 90 MHz (1.7/2.1 GHz AWS), 70 MHz (700 MHz), 78 MHz (2.5 GHz, 136-MHz BRS channels). The 30 MHz of WCS spectrum is not included because, while license rules are liberal in terms of services and technologies, emission rules are exceedingly stringent. The WCS licenses, auctioned in 1997, attracted extremely low bids as a result. *See* Cramton, *supra* note 35, at 635. Ironically, the FCC blocked a 2006 bid by satellite radio licensee XM to buy WCS licenses. With Satellite Digital Audio Radio Service licenses allocated spectrum adjacent to the WCS band, integration of ownership could have easily solved the externality problem. *See* Tony Sanders, *FCC Delay Scotches WCS Merger*, RADIO MONITOR (May 22, 2006), *available at* http://www.allbusiness.com/services/motion-pictures/4479825-1.html. On how ownership structures impact transaction costs, see Harold Demsetz, *Ownership and the Externality Problem*, *in* PROPERTY RIGHTS: COOPERATION, CONFLICT, AND LAW 282 (T.L. Anderson & F.S. McChesney eds., 2003).

Figure 1: Ratios of Unlicensed-to-Licensed Spectrum Under 6 GHz (2005)



An influential coalition composed of major technology firms such as Intel,[53] Microsoft,[54] Apple,[55] Cisco,[56] and Google,[57] together with computer manufacturers and several academics,[58] has been urging the FCC to expand the unlicensed "spectrum commons." Advanced low-power radios, they argue, can use embedded sensors and digital intelligence to sense each other's

53. *See, e.g.*, P. Pitsch, *The Future of Radio Spectrum Policy*, TECHNOLOGY@INTEL MAG., Mar. 2004.

54. *See, e.g.*, Comments of Microsoft Corp. at 2–3, *Revision of Parts 2 and 15*, ET Docket No. 03-122 (Fed. Commc'ns Comm'n Sept. 3, 2003).

55. *See, e.g.*, Comments of Wireless Info. Networks Forum, Apple Computer, Inc. Petition for Rulemaking To Allocate Spectrum in the 5 GHz Band To Establish a Wireless Component of the National Information Infrastructure, RM-8653 (Fed. Commc'ns Comm'n July 10, 1995).

56. *See, e.g.*, *Wireless and Spectrum Management*, CISCO SYS. (Jan. 2005), http://www.cisco.com/web/about/gov/networks/wireless_spectrum_management.html.

57. *See, e.g.*, Stephanie Condon, *Google Launches Free the Airwaves*, CNET NEWS (Aug. 18, 2008, 7:38 AM), http://news.cnet.com/8301-1035_3-10018917-94.html.

58. *See, e.g.*, LAWRENCE LESSIG, THE FUTURE OF IDEAS: THE FATE OF THE COMMONS IN A CONNECTED WORLD 73–84 (2001); WERBACH, *supra* note 17, at 47; Yochai Benkler, *Overcoming Agoraphobia: Building the Commons of the Digitally Networked Environment*, 11 HARV. J.L. & TECH. 287 (1998) [hereinafter Benkler, *Overcoming Agoraphobia*]; Benkler, *supra* note 17; Kevin Werbach, *Supercommons: Toward a Unified Theory of Wireless Communication*, 82 TEX. L. REV. 863 (2004); J.H. Snider, *Reclaiming the Vast Wasteland: The Economic Case for Re-allocating to Unlicensed Service the Unused Spectrum (White Space) Between TV Channels 2 and 51* (New Am. Found. Working Paper, 2006), *available at* http://www.newamerica.net/publications/policy/the_economic_case_for_re_allocating_the_unused_spectrum_white_space_between_tv_channels_2_and_51_to_unlicense.

presence and avoid interfering with each other's transmissions. In the most aggressive form of the argument, all carriers and consumers should therefore be allowed to transmit radio signals in any frequency band using one of these smart radios.

With this technology, these advocates claim, exclusive, property-like rights in spectrum are obstructive anachronisms. The FCC should therefore designate more frequency bands as unlicensed—i.e., to be used only by FCC-permitted, low-power radios. And the FCC should allow anyone to operate low-power transmitters in licensed bands allocated for broadcasting services, wireless phone companies, and others, so long as these "underlay" devices operate below some power threshold set by the FCC and incorporate smart protocols to avoid interfering with licensed transmissions.[59]

The FCC has been receptive to this vision. In 2002, an FCC-convened task force advocated further use of unlicensed bands.[60] On the predicate that unlicensed bands have been a "tremendous success,"[61] the Commission designated several new bands for unlicensed access.[62] The FCC also authorized underlay rights for ultra-wideband ("UWB") radios in bands above 10 GHz, provided they use very low power to leave other communication signals undisturbed.[63] The FCC further considered authorizing unlicensed devices to access licensed spectrum in its "Interference Temperature" proceeding, launched in November 2003; the initiative failed, however, and the proceeding was terminated in May 2007.[64] And the "3650 MHz" and "TV Band White Spaces" proceedings, mentioned *supra*, were clear victories for champions of unlicensed spectrum access who

59. *See* sources cited *supra* note 58.

60. *See* FED. COMMC'NS COMM'N, REPORT OF THE SPECTRUM EFFICIENCY WORKING GROUP (2002), *available at* http://www.fcc.gov/sptf/files/SEWGFinalReport_1.pdf.

61. Notice of Inquiry, Additional Spectrum for Unlicensed Devices Below 900 MHz and in the 3 GHz Band (*900 MHz and 3 GHz*), 17 FCC Rcd. 25,632, ¶ 6 (2002).

62. This includes 2 GHz in the 57–59 GHz band (2000); 255 MHz in the 5.470–5.725 GHz band; 2.9 GHz in the 92.0–94.0 and 94.1–95.0 GHz bands (2003); and 50 MHz in the 3.65–3.7 GHz band (2005). Amendment of Part 2 of the Commission's Rules To Allocate Additional Spectrum to the Inter-Satellite, Fixed, and Mobile Services and To Permit Unlicensed Devices To Use Certain Segments in the 50.2–50.4 GHz and 51.4–71.0 GHz Bands, 15 FCC Rcd. 25,264, ¶ 2 (2000); *Revision of Parts 2 and 15*, 18 FCC Rcd. 24,484 (2003); Allocations and Service Rules for the 71–76 GHz, 81–86 GHz and 92–95 GHz Bands, 18 FCC Rcd. 23,318, ¶ 4 (2003); *Wireless Operations in the 3650–3700 MHz Band*, 20 FCC Rcd. 6502 (2005).

63. *Revision of Part 15*, 17 FCC Rcd. 7435, ¶ 19 (2002).

64. Establishment of an Interference Temperature Metric To Quantify and Manage Interference and To Expand Available Unlicensed Operation in Certain Fixed, Mobile and Satellite Frequency Bands (*Interference Temperature Metric*), 22 FCC Rcd. 8938, ¶ 3 (2007).

prevailed in head-to-head match-ups where the Commission considered whether to apply licensed or unlicensed rules for specific frequencies.

III. WI-FI, TELEVISION, AND WIDE-AREA WIRELESS

Commons advocates misread the market evidence on Wi-Fi. Although highly effective and popular as a method for connecting PCs to broadband networks, fixed, short-range data links create wireless local area networks ("WLANs") that primarily serve to complement wide area networks ("WANs"). This is quite distinct from offering a wireless WAN ("WWAN") substitute. Wi-Fi nodes have flourished in enterprises as high-speed fiber connections became available; similarly, residential adoption has soared as companies extend cable modem, digital subscriber line ("DSL"), and fiber subscriptions to U.S. households. In this context, the driver of demand for unlicensed devices is the deployment of the WANs, which rely on private property rights to "spectrum in a tube."[65] This causality is more easily seen, perhaps, with cordless phones, an appendage of fixed networks. There, the clarity of the symbiotic economic relationship has generally prevented commentators from asserting that plain-old telephone service can more efficiently be supplied using a "spectrum commons." It would be nonsensical to claim that the use of the edge device (the cordless handset) was in any way undermining the utility of the ownership rights to the fixed network and the bandwidth created by its investors.

Unlicensed frequencies are not "open access" regimes that enable users to appropriate spectrum resources without constraint.[66] Nor are unlicensed bands organized within a "commons," where collective resource owners set usage rules to maximize joint returns.[67] Group owners do not set resource

65. The expression describes wired communications, which utilize the same electromagnetic spectrum as employed in wireless communications but create additional capacity by housing signals in constructed conduits.

66. The law and economics literature recognizes four standard property regimes: open access, state property, common property, and private property. *See* Dean Lueck & Thomas J. Miceli, *Property Law*, *in* 1 HANDBOOK OF LAW AND ECONOMICS 183, 190–200 (A. Mitchell Polinsky & S. Shavell eds., 2007). The U.S. spectrum regime has not, strictly speaking, nationalized airwaves or privatized airwaves. Rather, the law is that the public owns the airwaves, with the federal government regulating access so as to protect these public resources. *See* Thomas W. Hazlett, *The Wireless Craze, the Unlimited Bandwidth Myth, the Spectrum Auction Faux Pas, and the Punchline to Ronald Coase's "Big Joke": An Essay on Airwave Allocation Policy*, 14 HARV. J.L. & TECH. 335, 459–61 (2001). As a practical matter, this arrangement constitutes state ownership, also known as administrative allocation. Resource use is determined by state regulators, not by private or group owners.

67. This is the key characteristic of the self-organizing commons studied in such classic works as ELINOR OSTROM, GOVERNING THE COMMONS (1990).

appropriation terms; government regulators do. These regulators evaluate the trade-offs for adapting different rules on behalf of the public. Regulation of unlicensed bands creates non-exclusive use rights, and the conditions imposed exclude particular types of behavior so as to protect others. For example, a high-powered TV broadcast station is not allowed to blast emissions, and thus diminish opportunities for low-power radios to provide home networking links in the same market. This is intended to limit conflicts between rival users of a scarce resource,[68] rather than risk resource dissipation due to unproductive competition for rights.[69]

Property rules are an antidote to the "tragedy of the commons."[70] Where entirely unrestricted access prevails and scarcity exists, the marginal user will crowd into the "free" resource even as the cost borne by the group of users exceeds the newly created benefit. The standard example is an open pasture that becomes "over-grazed" by the owners of cattle, each of whom realizes some gain until the resource is destroyed for all.[71]

Perhaps because of the compelling manner in which the "over-grazing" problem is formulated, it is often missed that the regulation of unlicensed devices—prescriptions on technology standards and power limits—is just another approach to avert a tragedy of the commons. The regime seeks to permit some behaviors and to block others. The proper test of social efficiency is not whether "interference" has been reduced or eliminated, but whether the most valuable outcomes result.[72]

Three implications arise. First, anarchy does not reign. The power limits and technology restrictions imposed by regulators protect some applications and users at the expense of others. Scarcity is not eliminated; indeed, the effort to advance what Benkler labels a "well regulated commons"[73] is itself a rejection of open access. Allocating spectrum for unlicensed usage necessarily excludes certain wireless alternatives, implicating trade-offs that need not be made in the case of true resource abundance.

Second, tragedy of the commons may obtain even when there is little or no interference between users. Markets that experience over-utilization yield the most widely recognized "tragedy," but rules to mitigate resource

68. *See* Harold Demsetz, *Toward a Theory of Property Rights*, 57 AM. ECON. REV. 347 (1967).

69. *See* Henry E. Smith, *Exclusion Versus Governance: Two Strategies for Delineating Property Rights*, 31 J. LEGAL STUD. 453 (2002).

70. Garret Hardin, *The Tragedy of the Commons*, 162 SCIENCE 1243 (1968).

71. *See id.* at 1244.

72. Coase, *supra* note 21, at 27.

73. Benkler, *Overcoming Agoraphobia*, *supra* note 58, at 394.

dissipation can easily result in under-utilization, which is equally inefficient. Such an outcome includes the instance where investments in technology, infrastructure, or economic organization—say, paying incumbents to move their wireless operations—are socially efficient but fail to occur due to the lack of spectrum ownership. Market failure of this sort has occurred in unlicensed bands, such as Unlicensed PCS ("U-PCS"), where spectrum allocated for over a decade to data transmissions saw not a single device approved for use by the FCC.[74]

Third, even when abundant use is made of unlicensed bands, the allocation may be socially destructive. For example, unlicensed rules may exclude services that are more valuable than the protected activities. Equivalently, the social value created in the use of unlicensed devices may be achieved more economically via exclusive spectrum rights, provided either by market competitors or via the acquisition of bandwidth by a non-profit organization supplying a "spectrum park." The active secondary market in spectrum access vividly demonstrates how device makers can, rather than request unlicensed allocations for the use of their radios, contract with mobile licensees exercising de facto spectrum ownership. Apple arranges for its iPhone customers to access radio spectrum via bulk contracts it arranges with carriers in United States and internationally. Amazon sells its Kindle book reader with embedded wireless functionality for digital content downloads by contracting with the Sprint phone network. General Motors operates its emergency OnStar radio service by contracting with Verizon Wireless.[75] The market supplies radio spectrum bundled with network services to supply an extremely wide range of applications. Rival application providers effectively bid against each other for available wireless resources. To the degree that regulators make available spectrum inputs with exclusive ownership rights, the standard economic optimization results.

74. For a source discussing the failure of unlicensed PCS, see, e.g., Kenneth R. Carter, *Policy Lessons from Personal Communications Services: Licensed vs. Unlicensed Spectrum Access*, 15 COMMLAW CONSPECTUS 93, 97 (2006) ("[I]n 1993 the . . . FCC assigned PCS spectrum both by licenses awarded in competitive bidding auctions and through an unlicensed model. . . . It is hard to argue that licensed PCS has not been a huge success at lowering prices and spurring competition with cellular service. Conversely, unlicensed PCS has at best been a very late bloomer, and at worst, dead.") (footnotes omitted).

75. John W. Mayo & Scott Wallsten, *Enabling Efficient Wireless Communications: The Role of Secondary Spectrum Markets*, 22 INFO. ECON. & POL'Y 61, 65 (2010); Thomas W. Hazlett, *Modular Confines of Mobile Networks: Are iPhones iPhony?* (George Mason Law & Econ., Research Paper No. 10-01, 2010), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1533441.

IV. WI-FI: THE STARBUCKS FALLACY

The Wi-Fi standard for wireless, high-speed, local-area networks was ratified in 1999.[76] By 2002, global annual sales of Wi-Fi equipment had exceeded $1 billion;[77] by 2006, such sales reached an estimated $3.8 billion.[78] The secret of Wi-Fi's success, commons advocates maintain, is that the standard defines low-power, spread-spectrum devices that can operate "without the requirement of spectrum licensing to prevent interference."[79] With Wi-Fi, "there is no need for service providers, cell towers, controlled hardware markets, or expensive spectrum licenses."[80]

While Wi-Fi was taking off, comparable wireless data services offered in licensed bands were allegedly doing "exactly the opposite."[81] Although wireless phone companies soon began offering high-speed data services, these networks cost ten times more,[82] and "[n]one . . . has yet become a mass-market success."[83] Wi-Fi, it is suggested, is superior even to high-speed wireline connections: there is "great consternation in the communications industry" about the inadequacies of DSL and cable modem broadband services, and Wi-Fi service "costs half as much."[84]

Even putting aside the factual predicates and premature nature of these arguments,[85] the comparisons are fatally flawed. First, assuming that unlicensed devices are popular, it does not necessarily follow that more spectrum should be allocated to unlicensed instead of licensed uses. In 2008, U.S. Wi-Fi devices and cordless phones were out-sold by well over an order

76. *See IEEE Get Program 802*, INST. OF ELEC. & ELECS. ENG'RS, INC. (IEEE), http://standards.ieee.org/getieee802/802.11.html (last visited Apr. 7, 2011).

77. *See* Carter, Lahjouji & McNeil, *supra* note 37, at 32 (citing Goldman Sachs).

78. BERNIE MAHON & LOUIS GERHARDY, MORGAN STANLEY, Q1 2006 GLOBAL TECHNOLOGY DATABOOK 22 (2006), *available at* http://www.morganstanley.com/institutional/techresearch/pdfs/global_techdatabook0306.pdf.

79. WERBACH, *supra* note 17, at 23.

80. *Id.*

81. *Id.* at 22.

82. *Id.* at 23 ("[A] Wi-Fi network costs . . . one-tenth as much as a third-generation cellular network.").

83. *Id.* at 22.

84. *Id.* at 23.

85. Licensed wireless 3G services have become extremely popular with the advent of smartphones and other advanced wireless devices, while wireline broadband connections (particularly cable modem and new fiber-to-the-premises networks) now offer speeds up to 100 Mbps, far in excess of what a Wi-Fi connection provides. *See* Annual Report and Analysis of Competitive Market Conditions with Respect to Commercial Mobile Services, 24 FCC Rcd. 6185 (2009); Saul Hansell, *Cablevision Goes for U.S. Broadband Speed Record*, N.Y. TIMES BITS BLOG (Apr. 28, 2009, 12:01 AM), http://bits.blogs.nytimes.com/2009/04/28/cablevision-goes-for-us-broadband-speed-record/.

of magnitude by digital TV sets, which garnered $27 billion in sales.[86] Yet, the airwaves dedicated to broadcast television are severely misallocated because video signals can be delivered more efficiently on alternative platforms, freeing the TV band for more productive employments.[87] From a social welfare perspective, the relevant policy question is how incremental bandwidth can best create value. In evaluating trade-offs, it is important to recognize that licensed regimes can provide the same applications as unlicensed spectrum. Just as the private property regime for land governs the supply of public parks, public, private non-profit, or private for-profit enterprises can host diverse spectrum sharing arrangements.[88] Indeed, this is seen in the operation of mobile networks, which arrange for millions of customers to use specific licensed frequencies. To identify an efficient outcome, the analysis must determine that the costs of pre-empting alternative liberal license allocations are less than unlicensed benefits. No amount of economic activity in unlicensed bands makes the case for allocating more unlicensed spectrum.

Second, as an empirical matter, the social value of the economic activity associated with liberal licenses far exceeds that achieved in unlicensed bands. Licensed WWANs, relying on licensed spectrum, are used by more than 270 million[89] U.S. subscribers who pay over $148 billion[90] a year and generate at

86. Press Release, CEA, Consumer Electronics Industry Issues 2009 Forecast (Jan. 8, 2009), http://www.ce.org/Press/CurrentNews/press_release_detail.asp?id=11666 (stating that digital TV sets accounted for fifteen percent of the $172 billion shipment revenues in 2008).

87. Hazlett, *Transition to Yesterday*, *supra* note 19.

88. FCC spectrum policy experts have noted the efficiencies of such an institutional approach:

> [When] making decisions about the amount of spectrum allocated to unlicensed use, the government should face the opportunity cost of limiting or foreclosing other use. Just as the government decides how much land to purchase for public parks, it would decide how much spectrum to set aside for unlicensed devices. A market system would also provide the opportunity for private spectrum licensees in flexible bands to compete with the government for the provision of spectrum for low-power devices, just as private facilities that charge admission compete with public parks. Licensees might find it profitable to do so by charging manufacturers of such devices to operate on their spectrum. This would allow private licensees to compete on the technical protocols and other quality factors instead of relying on government or industry committees.

Kwerel & Williams, *supra* note 8, at 7.

89. CTIA, SEMI-ANNUAL WIRELESS INDUSTRY SURVEY (2008), *available at* http://files.ctia.org/pdf/CTIA_Survey_Year-End_2008_Graphics.pdf.

90. *Id.*

least another $150 billion[91] in consumer surplus. Equipment sales tell a similar story. In 2006, global sales for WWANs using liberal licenses were about $225 billion (including handsets), while wireless local area networks ("WLANs"), using unlicensed frequencies, totaled about $3.8 billion.[92]

FCC statistics for broadband access in the United States tell the same story. While unlicensed bandwidth is available everywhere for the use of Internet Service Providers ("ISPs"), and while hundreds of wireless ISPs ("WISPs") use airwaves for this purpose,[93] they operate in small deployments, usually in rural areas where low population densities ensure minimal interference.[94] As a result, unlicensed frequencies, while useful in providing Wi-Fi as a WLAN appendage in millions of homes or businesses that subscribe to wide area broadband services, are virtually non-existent in supplying WAN services themselves. The FCC recorded only 525,000 fixed wireless customers as of year-end 2009, a total which includes ISPs using licensed and unlicensed frequencies, as compared to more than 52 million mobile high-speed data customers (all delivered via licensed spectrum).[95] In addition, the FCC reported over 80 million fixed high-speed connections (cable modem, DSL, and fiber), services delivered via networks relying on the ownership of spectrum encased in wires. The bottom line: more than 130 million subscribers receive high-speed data service (fixed and mobile) via exclusively owned bandwidth, as compared to just a few hundred thousand

91. Hazlett & Muñoz, *supra* note 36, at 425. Professor Jerry Hausman produces consistent estimates using a different model. *See* Jerry Hausman, *Mobile Telephone*, *in* 1 HANDBOOK OF TELECOMMUNICATIONS ECONOMICS 563, 585–86 (Martin Cave et al. eds., 2002).

92. MAHON & GERHARDY, *supra* note 78, at 22, 24.

93. The Wireless Internet Service Providers Association ("WISPA") "was founded in 2004 and represents the interests of more than 300 wireless internet service providers ('WISPs'), vendors, system integrators and others interested in promoting the growth and delivery of fixed wireless broadband services to Americans. WISPA estimates that more than 2,000 WISPs operate in the United States today." Comments of the Wireless Internet Serv. Provider Ass'n at 2, Inquiry Concerning the Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion, and Possible Steps To Accelerate Such Deployment Pursuant to Section 706 of the Telecommunications Act of 1996, as Amended by the Broadband Data Improvement Act, A National Broadband Plan for Our Future, GN Docket Nos. 09-47, -51 & -137 (Fed. Commc'ns Comm'n Dec. 7, 2009), *available at* http://fjallfoss.fcc.gov/ecfs/document/view?id=7020351635; WISP Directory, WISPA, http://www.wispdirectory.com/index.php?option=com_mtree&task=listcats&cat_id=73&Itemid=53 (last visited Mar. 3, 2011) (listing 1,845 WISPs in the United States).

94. This is observed in comparing the large number of WISPs relative to the small number of "fixed wireless" subscribers (*see infra* Table 1), as well as examining the WISP map produced by WISPA. WISP Directory, *supra* note 93.

95. *See infra* Table 1.

subscribers—at most—to WISPs and those accessing the Internet via a "spectrum commons."

Table 1: U.S. Broadband Subscribers (Dec. 2005 to Dec. 2009)[96]
(*Thousands of Subscribers*)

	2005	2006	2007	2008	2009
ADSL	19,515	25,413	29,449	30,198	30,971
SDSL	369	345	293	241	225
Traditional Wireline	373	545	605	705	716
Cable Modem	26,558	31,982	36,507	40,273	43,128
Fiber	298	894	1,849	2,884	3,975
Satellite	427	572	791	938	1,116
Fixed Wireless	257	483	707	486	525
Mobile Wireless	3,128	22,288	51,016	25,040	52,486
Power Line and Other	5	5	5	5	5
Total Lines	50,930	82,525	121,222	100,770	133,148

The rise and fall of Municipal Wi-Fi networks is also revealing evidence. "Muni Wi-Fi" was widely touted as the next big thing in broadband in 2006 and 2007.[97] At one point, there were plans to deploy scores of networks covering major urban areas such as Philadelphia, Houston, Boston, Chicago, Los Angeles, Atlanta, and San Francisco.[98] By 2008, however, virtually all of these plans had collapsed.[99] Access to "free" airwaves was supposed to provide cheap service, but the structure of property rights—with regulatory limitations on power, technology, and the inability to exclude (and, therefore, contract with) potentially rivalrous spectrum users—has rendered the model of limited practical use.[100] Gradually, this reality has set

96. FED. COMMC'NS COMM'N, HIGH-SPEED SERVICES FOR INTERNET ACCESS: STATUS AS OF DECEMBER 31, 2009, tbl.7 (2011). Note that the large drop in mobile wireless subscribers, 2007 to 2008, was caused entirely by a switch in FCC accounting practices. Previously, all high-speed capable wireless devices used by subscribers were counted. Beginning in the 2008 data, only subscriptions "which [include] a data plan for transferring, on a monthly basis, either a specified or an unlimited amount of data to and from Internet sites of the subscriber's choice" were included in the "Mobile Wireless" total. *Id.* at 81.

97. *See* Robert McChesney & John Podesta, *Let There Be Wi-Fi*, WASH. MONTHLY, Jan.–Feb. 2006, *available at* http://www.washingtonmonthly.com/features/2006/0601.podesta.html.

98. Thomas W. Hazlett, *Philadelphia Freedom*, ARS TECHNICA (Dec. 22, 2008, 11:05 PM), http://arstechnica.com/telecom/news/2008/12/muni-wifi-fcc-free-wireless.ars.

99. *Id.*

100. This is why the hundreds of private WISPs did not see such services as profit opportunities, leading community activists and policymakers to seek local government support. The deals struck by municipalities included subsidies, monopoly rights to access

in. "Story after story after story highlight[s] how wide-area WiFi is a lot more complicated than many in the industry (and the press) would have you believe."[101]

At the end of the day, Wi-Fi offers only very limited substitution possibilities for high-speed DSL, cable, fiber, or wide area wireless. As commons advocates themselves acknowledge, it is "a short-range technology designed primarily for connections to a nearby hotspot."[102] Wi-Fi radios typically operate at such low power that "[e]ven if every home in a neighborhood had a Wi-Fi access point, few of those nodes would *see* one another."[103] Wi-Fi devices, in short, rarely interfere with each other because they provide such limited range. The secret of Wi-Fi's success is the secret of Wi-Fi's failure. By contrast, DSL, cable modems, fiber, and wireless broadband service (e.g., EV-DO) provide wide-area coverage across cities and markets, cost-effectively scaling to national and international networks.

Some predict, however, that more and better Wi-Fi lies ahead and that radio-frequency engineering advances will overcome all of these limitations.[104] Such range-extending technology includes phased-array antennas, mesh networking boxes that automatically create ad hoc mesh networks with each other, a follow-up standard (802.16) for wireless metropolitan-area-network ("MAN") technology, and so forth.[105]

There is no doubt that innovative advances will come, but they will not be restricted to use in unlicensed bands. Rapid technological change is also producing disruption in networks using licensed spectrum. Airwaves are

---

street lights (for placement of wireless nodes), or exclusive contracts with public agencies (anchor tenants).

101. Michael Masnick, *Can Mesh WiFi Solve the Net Neutrality Issue?*, TECHDIRT (June 20, 2006, 3:24 AM), http://techdirt.com/articles/20060620/0318228.shtml.

102. WERBACH, *supra* note 17, at 39 ("Basic WiFi or its variants . . . cannot simply be put into service for last-mile deployments.").

103. *Id.* (emphasis added).

104. *See, e.g.*, David Weinberger, *The Myth of Interference*, SALON (Mar. 12, 2003), *available at* http://www.salon.com/tech/feature/2003/03/12/spectrum/print.html.

105. WERBACH, *supra* note 17, at 39; Björn Wellenius & Isabel Neto, *The Radio Spectrum: Opportunities and Challenges for the Developing World* 7 (World Bank Policy, Working Paper No. 3742, 2005) ("Other recent innovations include smart radios and antennas, software-defined radios, cognitive radios, and mesh, ad-hoc, or viral networks. As a group, these technologies enable users not to cause insurmountable interference to each other even when transmitting at the same time, in the same place, and on the same parts of the spectrum.") (footnotes omitted); *see also* Fulvio Minervini, *Emerging Technologies and Access to Spectrum Resources: The Case of Short-Range Systems*, 67 COMM. & STRATEGIES 107, 112–15 (2007), *available at* http://mpra.ub.uni-muenchen.de/6786/1/MPRA_paper_6786.pdf (arguing that advanced antennas and mesh networks enhance access to the radio spectrum, while not disrupting the traditional framework of spectrum usage).

agnostic to regulatory regimes; users and investors are not. When network owners can optimize a given spectrum space by contracting to share it, they are often able to bring more resources and greater social coordination to bear. This is observed in liberal-licensed spectrum, which is being far more intensely developed and utilized, in economic value terms, than spectrum allocated to traditional licenses or unlicensed bands. As U.S. smartphone sales reach $14 billion per year,[106] iconic wireless innovations such as the Apple iPhone and RIM Blackberry rely on WWANs, and the licensed spectrum they use, to revolutionize communications. Hence, as the benefits of using unlicensed spectrum rise, the opportunity costs of taking allocations away from liberal licenses rise pari passu.

Which brings us to Starbucks. With all options other than low-power radios eliminated (through government regulation), exclusive rights in *real estate* may afford some of the protections of exclusive rights in *spectrum*. Users can themselves limit the number of transmitters and/or receivers competing for access to spectrum in corporate offices, on university campuses, and in Starbucks coffee shops.[107] Intel, for example, carefully restricts unauthorized use of unlicensed frequencies within its corporate office space.[108] Carnegie Mellon protects the spectrum on its campus by effectively "privatizing the commons," i.e., carefully selecting what services and users to include in its network (and thereby which to exclude), while adopting technologies and distributing wireless access points to control interference problems.[109] This is not the "end of scarcity," but the operations of a property system to manage competing, mutually exclusive activities.[110] Conflicts still exist, and "open access" would be socially destructive.

106. CEA estimates 2009 U.S. sales of 37 million smartphones, with total revenues of $14 billion. CEA Database, *supra* note 4. This is about forty times cordless phone sales. Wi-Fi sales are not charted by CEA.

107. *See* Hazlett & Muñoz, *supra* note 27, at 264–66.

108. *See* Mike Chartier, Intel, *Local Spectrum Sovereignty: An Inflection Point in Allocation*, *in* PROCEEDINGS OF THE INTERNATIONAL SYMPOSIUM ON ADVANCED RADIO TECHNOLOGIES 29, 32–33 (2004), *available at* http://www.its.bldrdoc.gov/pub/ntia-rpt/04-409/04-409.pdf ("Failure to fulfill the above terms and conditions [for non-IT WLANs] will result in I.T.'s disconnecting and or taking possession of the Experimental W-LAN Access Points.").

109. Airspace Guideline for 2.4 GHz Radio Frequency at Carnegie Mellon University, *cited in* Chartier, *supra* note 108, at 33 ("While we will not actively monitor use of the airspace for potential interfering devices, we will seek out the user of a specific device if we find that it is actually causing interference and disrupting the campus network. In these cases, Computing Services reserves the right to restrict the use of all 2.4 GHz radio devices in university-owned buildings and all outdoor spaces on the Carnegie Mellon Campus.").

110. Weinberger, *supra* note 104.

In short, regulatory exclusions police Starbucks stores' airwaves and limit unlicensed frequencies to the use of cordless phones, wireless routers, or other similar low-power devices. Starbucks may eliminate some further conflicts; instead of inviting any and all ISPs to set up local area networks in its stores, it designates and contracts with an exclusive provider.[111] The service may require paying a fee for access, as it was offered for many years; free access may be extended to paying customers, a model to which Starbucks then switched;[112] or access may be open to all who receive the signal, a model to which Starbucks more recently switched.[113] Prior to opening access to all, the "spectrum commons" had been locally privatized, allowing a service supplied solely by AT&T, one of the nation's largest mobile wireless carriers.[114] Such services provide social value. What remains unclear, given the lack of a market for the spectrum inputs consumed, is whether the opportunities consumed by the government's spectrum allocation do not exceed these benefits. Net social value may well have been higher had companies like Starbucks, AT&T, and Cisco (a large maker of Wi-Fi hotspot routers) been forced to economize on spectrum resources by purchasing them in the market—just as AT&T does when acquiring billions of dollars worth of liberal licenses, making wireless access available to its millions of mobile subscribers.

What works in a Starbucks does not necessarily produce benefits that exceed social opportunity costs. Nor does it scale to other useful wireless applications. Indeed, the rules that allow the beneficial deployment of short range devices in a home or enterprise make the deployment of WANs—particularly for services involving the economic and technical complexity of mobile access—extremely problematic. The local Wi-Fi link that connects clustered, fixed users to a WAN relies on private property rights to radio

111. In 2008, Starbucks switched its designated Wi-Fi partner from T-Mobile to AT&T. Steve Stroh, *AT&T Upgrades 3G Network and Takes Over Starbucks Wi-Fi*, TECHREPUBLIC (Feb. 11, 2008), http://www.techrepublic.com/blog/wireless/at-t-upgrades-3g-network-and-takes-over-starbucks-wi-fi/185.

112. David Sarokin, *How To Get Free Wi-Fi at Starbucks*, EHOW.COM (Aug. 30, 2009), http://www.ehow.com/how_2269126_wifi-starbucks.html.

113. Jennifer van Grove, *How Starbucks Plans To Capitalize on Free Wi-Fi*, MASHABLE (Aug. 12, 2010), http://mashable.com/2010/08/12/starbucks-digital-network/. It is interesting to note that, while customer access may be free to those in (or near) a Starbucks, company websites featured on the start-up page are limited to Starbucks' strategic partners. In the "Starbucks Digital Network," the store chain receives a cut of online revenues generated by companies like Apple, Rodale, and the New York Times, which are granted preferential access to Wi-Fi users in exchange. *Id.*

114. *See High Speed Wireless Internet Access*, STARBUCKS, http://www.starbucks.com/coffeehouse/wireless-internet (last visited Jan. 21, 2011).

spectrum to transport data to distant networks. Just as the cordless phone depends on, and does not replace, the telephone network, the cordless PC depends on, and does not displace, the wired or wireless WAN.

A. BROADCASTING LICENSES: A BLAST FROM THE PAST

According to the exclusivity critique, the central problem in spectrum is that ownership rights result in the waste of bandwidth. In any given geographic area, many channels in a licensed system are empty much of the time.[115] This was not always the case, according to spectrum commons advocates. Licensing was perhaps needed in 1927, but only because the broadcast radios of that era were not smart. Things are very different today. Technology, they claim, makes once-scarce spectrum plentiful.[116] To prove this proposition, commons advocates have measured "actual usage of the most active channels of the broadcast bands . . . during peak hours in the highly populated, Dupont Circle area of Washington, DC."[117] These

115. WERBACH, *supra* note 17, at 8; Lynnette Luna, *Start-up Looks To Jump Start Secondary Spectrum Market*, FIERCEBROADBANDWIRELESS (Mar. 10, 2008, 7:59 AM), http://www.fiercebroadbandwireless.com/story/start-up-looks-to-jump-start-secondary-spectrum-market/2008-03-10 (quoting Rick Rotondo, Spectrum Bridge's Vice President of Marketing, stating that "in any given time and place, 80 percent to 94 percent of all allocated spectrum in the U.S. goes unused"); *see* Promoting Efficient Use of Spectrum Through Elimination of Barriers to the Development of Secondary Markets (*Promoting Efficient Use*), 15 FCC Rcd. 24,203, ¶ 2 (2000) ("[R]adio spectrum may be used inefficiently by its current licensees or even lie fallow, especially in rural areas, limiting availability of valuable services to many."); Michael Calabrese, *The End of Spectrum 'Scarcity': Building on the TV Bands Database To Access Unused Public Airwaves* 1, 16 (New Am. Found., Working Paper No. 25, 2009), *available at* http://www.newamerica.net/files/Calabrese_WorkingPaper25_EndSpectrumScarcity.pdf ("[I]n every community across the country, large swaths of valuable spectrum lie fallow the majority of the time. This underutilized spectrum represents enormous, untapped, public capacity for high-speed and pervasive broadband connectivity. . . . [S]tudies show that only a fraction of even prime frequencies below 3 GHz are in use, even in the largest cities, at any particular place or time. Federal agencies sit on hundreds of MHz that are unused in most areas; and many private licensees are warehousing spectrum, particularly in rural areas."); *see also Spectrum Reports*, SHARED SPECTRUM CO., http://www.sharedspectrum.com/papers/spectrum_reports/ (last visited Mar. 23, 2011) (showing spectrum occupancy measurements in six east coast locations including New York City, New York).

116. *E.g.*, WERBACH, *supra* note 17, at 21 ("Technology is making the wireless world look more and more like the ocean."); *see also, e.g.*, U.S. GEN. ACCOUNTING OFFICE, GAO-03-277, TELECOMMUNICATIONS: COMPREHENSIVE REVIEW OF U.S. SPECTRUM MANAGEMENT WITH BROAD STAKEHOLDER INVOLVEMENT IS NEEDED 9, 55–56 (2003) ("[A]dvances in technology could also help to accommodate more services and users."); Paul Kolodzy, *Communications Policy and Spectrum Management*, *in* COGNITIVE RADIO TECHNOLOGY 64 (Bruce A. Fette ed., 2006) ("SDRs and other advanced technologies can potentially alleviate many of the conflicts by making spectrum more plentiful through more efficient access.").

117. WERBACH, *supra* note 17, at 8.

measurements are believed to establish that "[m]ost of the spectrum is empty in most places most of the time."[118]

But the measurements equate *emitted radiation* with actual usage, disregarding the economic value generated. Under this scale, a broadcast tower emitting a 1,000,000-watt test pattern signal is not a waste of spectrum and electricity, but a highly utilized frequency band. And even if spectrum use measurements were correctly done, they would reveal an even more sweeping indictment: the absence of effective airwave ownership cripples the process by which frequency spaces are bid into their highest-valued uses.[119] Administrative allocation leaves regulators in charge of resource choices, meaning that interest group competition rules the roost. The ugly result is that airwaves are systematically under-utilized, maximizing not consumer welfare but the political interests of influential coalitions.

The historical origins of this system are instructive. In 1920, Westinghouse inaugurated the nation's first successful radio station, KDKA, in Pittsburgh.[120] Hundreds of other new stations began broadcasting shortly thereafter. Each transmitter was required to obtain a license from the Department of Commerce under the 1912 Radio Act.[121] While anyone could register for a license, and the Commerce Department had no basis on which to deny them, the agency was permitted to issue licenses under terms "minimizing interference."[122] From 1921–1925, the Department established rules for allocating frequency slots on a first-come, first-served basis.[123] The

118. *Id.*; *see also* Max Vilimpoc & Mark McHenry, *Dupont Circle Spectrum Utilization During Peak Hours*, NEW AM. FOUND. (2003), http://www.newamerica.net/files/nafmigration/archive/Doc_File_183_1.pdf.

119. *See* Comments of 37 Concerned Economists at 7, *Promoting Efficient Use*, WT Docket No. 00-230 (Fed. Commc'ns Comm'n Feb. 7, 2001) (encouraging the FCC to "adopt market-oriented rules opening the radio spectrum and capturing its full potential for society"); Gerald R. Faulhaber, *The Future of Wireless Telecommunications: Spectrum as a Critical Resource*, 18 INFO. ECON. & POL'Y 256 (2006); Hazlett, *supra* note 66; Gregory L. Rosston & Jeffrey S. Steinberg, *Using Market-Based Spectrum Policy To Promote the Public Interest*, 50 FED. COMM. L.J. 87 (1997); Pablo T. Spiller & Carlo Cardilli, *Towards a Property Rights Approach to Communications Spectrum*, 16 YALE J. ON REG. 53 (1999); Lawrence J. White, *'Propertyzing' the Electromagnetic Spectrum: Why It's Important, and How To Begin*, 9 MEDIA L. & POL'Y (2000).

120. Joseph E. Baudino & John M. Kittross, *Broadcasting's Oldest Stations: An Examination of Four Claimants*, 21 J. BROAD. 61 (1977); *see also Milestones: Westinghouse Radio Station KDKA, 1920*, IEEE GLOBAL HISTORY NETWORK (June 1994), http://www.ieeeghn.org/wiki/index.php/Milestones:Westinghouse_Radio_Station_KDKA,_1920.

121. *See* Thomas W. Hazlett, *The Rationality of U.S. Regulation of the Broadcast Spectrum*, 33 J.L. & ECON. 133, 135 (1990).

122. 1912 Radio Act, ch. 287, § 4, 37 Stat. 302, 304.

123. *See* CLARENCE C. DILL, RADIO LAW: PRACTICE AND PROCEDURE 66–71 (1938) (discussing changes to regulations by the Secretary of Commerce in the 1920s).

government then delayed or encumbered new licenses (mandating time-sharing agreements, for example) to effectively protect existing stations from encroachment by entrants.[124]

The de facto property system, based on common law principles of *priority-in-use* or *right of user*, successfully launched the emerging medium. By the mid-1920s, and prior to any "public interest" licensing law, radio had become an extremely popular mass-market commodity.[125] Yet key policymakers were not happy with this result, as the first-come, first-served approach severely limited their degree of freedom. The Secretary of Commerce, Herbert Hoover, was vocal in his support for more administrative discretion, as was Senator Clarence C. Dill (D-WA), a congressional leader in the area.[126]

Just as importantly, large incumbent radio stations sought a greater level of security. While priority-in-use protected their signals, it also created risk of competitive entry. That is to say, new spectrum could potentially be claimed by new rivals. The broadcast stations sought a regulatory solution under which barriers could be legally erected to prevent this; the standard of "public interest, convenience, or necessity"—first suggested by the newly-formed National Association of Broadcasters in 1925—was such a rule.[127] Large commercial radio stations formed a coalition with key policymakers and ultimately gained passage of the Radio Act of 1927, placing radio broadcasting under the new Federal Radio Commission ("FRC").[128] The FRC was empowered to license transmitters, assign frequencies, prescribe service limits, and approve the locations and power levels of transmitters according to "public interest" criteria. These policies were carried forward by the Communications Act of 1934, which transplanted the FRC into a newly-constituted FCC.[129]

Both policymakers and powerful incumbent stations gained from the bargain in the 1927 and 1934 Acts. Regulators obtained considerable control

---

124. Hazlett, *supra* note 121, at 143–47.

125. Thomas W. Hazlett, *Physical Scarcity, Rent Seeking, and the First Amendment*, 97 COLUM. L. REV. 905, 913–19 (1997).

126. Hazlett, *supra* note 121, at 162–63. Dill was a principal author of the Radio Act of 1927.

127. DILL, *supra* note 123, at 89.

128. Radio Act of 1927, ch. 169, 44 Stat. 1162 (repealed 1934). The Interstate Commerce Commission retained authority over common carrier use of radio spectrum.

129. Communications Act of 1934, ch. 652, 48 Stat. 1064 (codified as amended at 47 U.S.C. §§ 151–615b (2006)). The Communications Act of 1934 and its amendments, including those in the 1996 Telecommunications Act, are posted by the FCC at http://www.fcc.gov/Reports/1934new.pdf.

over the operations of licensees, including the ability to influence program content.[130] And because license holders would be restricted to explicitly authorized activities, competition between licensees could be limited. New services, technologies, and business models were barred—forming, effectively, a government-enforced cartel. Later, law and economics scholars would characterize the general arrangement as "Taxation by Regulation."[131] Reducing competition via legal barriers increased profits; regulators then redirected some of the economic gains towards "public interest" expenditures.

FCC spectrum allocation proved bureaucratically tidy but economically inflexible. First, the FCC would zone the real estate. It assigned large blocks of spectrum for particular uses such as AM radio or Very High Frequency TV ("VHF TV"). Then it sliced each block into smaller licenses and assigned them to individual firms.[132] Licensees got no "property rights" in spectrum;[133] licenses typically expired after eight years,[134] and they could not be transferred without Commission approval.[135] Between 1927 and the early 1970s, the FCC promulgated a dense web of rules governing license retention and alienability, transmission and programming rights, signal privacy, and content—rules like the "fairness doctrine" or an obligation to air programs deemed educational for children.[136] Broadcasters, for example, were barred early on from using their main frequencies and facilities to transmit private, addressed messages to specific receivers, essentially telephone or telegraph services.[137]

The traditional licensing approach results in vast waste of bandwidth precisely because licensees are given no ownership in the underlying spectrum but are instead restricted to specific uses of a radio technology as defined by the regulator. Without spectrum ownership, private parties cannot transact to make more valuable use of idle frequencies. The tragedy is not the

130. Hazlett, *supra* note 121, at 158–63.

131. Richard A. Posner, *Taxation by Regulation*, 2 BELL J. ECON. & MGMT. SCI. 22 (1971).

132. Hazlett, *supra* note 66, at 337–50. *See generally* John O. Robinson, *Spectrum Management in the United States: An Historical Account* (Fed. Commc'ns Comm'n, O.P.P. Working Paper No. 15, 1985).

133. *See* 47 U.S.C. §§ 301, 304, 309(h)(1) (2006); *see also* Radio Act of 1927 § 1.

134. 47 U.S.C. § 307(c).

135. *Id.* § 310(d); *see also* Radio Act of 1927 § 12.

136. *See generally* Glen O. Robinson, *The Federal Communications Commission: An Essay on Regulatory Watchdogs*, 64 VA. L. REV. 169 (1978).

137. *See* Scroggin & Co. Bank, 1 F.C.C. 194 (1935); Bremer Broad. Co., 2 F.C.C. 79 (1935).

overuse of the commons, but its underuse.[138] In theory, the regulatory agency could prevent this, but it is not vested with real ownership, is unable effectively to finance productive, spectrum-enhancing investments, and receives no economic reward for generating extra social value. These misaligned incentives create "non-market failure."[139]

Shifting to liberal licenses—granting wireless service providers exclusive rights and broad flexibility to use allocated spectrum—remedies this tragedy, thereby creating the legal institutions to support Coasean contracting. These unleash incentives to invest in complementary assets that improve the productivity of airwaves, creating value for wireless users, some of which can then be captured by spectrum owners. This property structure enables complex organizational efforts, including those involving billions of dollars in risk capital dependent on the actions of millions of customers far into the future.

In sum, the case for un-licensing spectrum today is based on the deficiencies of a broadcast licensing policy established in the 1920s, a framework largely repudiated by spectrum rights that emerged in the United States and elsewhere in the 1980s and 1990s. Broadcasting is no longer dominant, having been eclipsed in value terms by the world of mobile communications. The licenses that enable those emerging markets are sufficiently liberal as to resemble de facto ownership of radio spectrum. The traditional licensing regime, despite the continued support of survivalists in politics and broadcasting, is universally recognized as obsolete. Grounded in a critique of the Radio Act of 1927, the case for un-licensing spectrum targets a corpse, oblivious to the thundering herd of exclusive spectrum rights now dominating communications markets.

V. WIRELESS CARRIERS AND LIBERAL LICENSES

A. THE NEXTEL "REALLOCATION"

One group of "broadcasters" has already completed a very successful transition from the old ways of licensing to the new. Radio dispatch services—used by taxicab companies, for example—once operated much like radio stations, under licenses that narrowly specified the service to be

138. Or, some call it a tragedy of the anticommons. *See* MICHAEL HELLER, THE GRIDLOCK ECONOMY (2008).

139. CHARLES WOLF, JR., MARKETS OR GOVERNMENTS: CHOOSING BETWEEN IMPERFECT ALTERNATIVES 57 (1988) (describing "non-market failure" as a situation in which the incentives of government policymakers do not reliably produce efficient outcomes, analogous to "market failure").

provided and the technology to be used. In 1987, a former FCC lawyer named Morgan O'Brien teamed up with an investment banker and began buying up dispatch companies—and their Specialized Mobile Radio ("SMR") licenses—across the country.[140] Their company, FleetCall, then put forward a plan asking regulators to approve a technical upgrade for SMR licenses: they sought the right to deploy digital instead of analog radios.[141]

The license modification also sought permission to use the extra capacity made possible by the technical upgrade for cellular phone calls. Such requests are met with strong opposition from established interests and are typically deterred: why spend scarce resources on such low-probability payoffs? But Morgan O'Brien believed that his knowledge of the Commission and the timing of this proposal—with the explosion of cellular use and the drift in deregulatory philosophy—brightened FleetCall's prospects. Some years and $2 million in legal fees later (as against an estimated $25 million in fees for opponents of the requested rule changes),[142] the underdog received approval.[143]

FleetCall adopted a new Motorola technology (iDEN, based on Time Division Multiple Access or "TDMA"), greatly expanding network capacity.[144] In March 1993, the "taxi dispatch" company renamed itself Nextel.[145] In 1995, wireless pioneer Craig McCaw invested $1.1 billion.[146] Motorola improved iDEN to enable data and fax communications in addition to voice, along with two-way dispatch and paging applications.[147] By 2003, FleetCall had acquired spectrum rights in the 700, 800, and 900 MHz bands[148] and was operating one of the largest digital networks in the country.[149] FleetCall executed a "spectrum swap" with the FCC, reducing

140. *See* K. Maney, *Nextel's Morgan O'Brien Kept the Faith and, Boy, Has It Paid Off*, USA TODAY, Dec. 15, 2004, at 3B; *see also* O. CASEY CORR, MONEY FROM THIN AIR: THE STORY OF CRAIG MCCAW 235–48 (2000).

141. The technological innovation would have violated license terms—hence the petition for license modification.

142. Hazlett, *supra* note 66, at 388.

143. *See* Request of Fleet Call, Inc. for Waiver and Other Relief To Permit Creation of Enhanced Specialized Mobile Radio Systems in Six Markets, 6 FCC Rcd. 1533, ¶ 36 (1991).

144. *New Motorola Digital Technology Increases Channel Capacity as Much as Six Times, Promises Enhanced Services to Thousands of Customers*, PR NEWSWIRE, Sept. 20, 1991.

145. *See Nextel History*, SPRINT, http://shop2.sprint.com/en/about/corporateinfo/company_history5.shtml (last visited Apr. 7, 2011).

146. *See Motorola Licenses Radio System, McCaw Invests in Nextel*, NEWSBYTES, Apr. 5, 1995.

147. *Motorola Announces Commercial Availability of iDEN Technology Enhancement*, PR NEWSWIRE, June 17, 1996.

148. *See* Nextel Commc'ns, Inc., Annual Report (Form 10-K), at 6 (Mar. 27, 2003).

149. Press Release, Nextel, Nextel Completes Another Industry First (July 29, 2003), http://findarticles.com/p/articles/mi_m0EIN/is_2003_July_29/ai_105986364.

interference that its phones caused with adjacent public service (fire, police, etc.) frequencies, receiving spectrum holdings in the 1.9 GHz band.[150] It then acquired several 2.1 and 2.5 GHz licenses and established partnerships and roaming agreements, giving it national coverage.[151] In 2005, Nextel—providing push-to-talk walkie-talkie service, wireless data services, wireless internet access, and short messaging to roughly seventeen million customers[152]—was sold to Sprint for an acquisition price of $35 billion.[153]

B. FCC REFORMS

Other wireless carriers acquired their spectrum assets more directly. Verizon Wireless,[154] AT&T Mobility,[155] T-Mobile,[156] and other mobile networks received cellular licenses—initially issued by the government in lotteries[157]—largely through secondary market purchases. The early cellular networks were required to build systems incorporating the analog Advanced Mobile Phone Systems ("AMPS") standard. In 1988, the FCC relaxed this requirement, allowing operators to upgrade to a digital standard of their choosing, though still requiring that they maintain the old AMPS system as

150. Roy Mark, *Nextel Finalizes Spectrum Swap*, WI-FI PLANET (Feb. 7, 2005), http://www.wi-fiplanet.com/news/article.php/3469601; Improving Public Safety Communications in the 800 MHz Band, 19 FCC Rcd. 14,969, ¶ 61 (2004).

151. Nextel Commc'ns, Inc., Annual Report (Form 10-K), at 5 (Mar. 15, 2005).

152. *See* Press Release, Nextel, Nextel Reports Strong Results (July 21, 2005), http://www.thefreelibrary.com/Nextel+Reports+Strong+Results.-a0134233871.

153. *Sprint, Nextel Complete Merger*, DIGITAL TRENDS (Aug. 12, 2005), http://www.digitaltrends.com/mobile/sprint-nextel-complete-merger/.

154. Verizon was formed from the merger of Bell Atlantic and GTE. Bell Atlantic had previously merged with NYNEX and AirTouch (originally the wireless arm of Pacific Bell). Verizon owns fifty-five percent of Verizon Wireless; Vodafone, a global mobile carrier based in the United Kingdom, owns the other forty-five percent. Bill Greenberg, *History of the Big Four US Carriers*, PHONESCHOLAR.COM (Feb. 7, 2011), http://www.phonescholar.com/verizon/a-history-of-the-big-four-us-carriers.

155. SBC (which had previously merged with Ameritech and Pacific Telesis) jointly owned Cingular with BellSouth (with SBC owning sixty percent and BellSouth forty percent). In October 2004, Cingular merged with AT&T Wireless; by this time, SBC had acquired the long distance operator AT&T. Following the merger, SBC changed its name to AT&T. In 2006, AT&T acquired BellSouth. *Id.*

156. T-Mobile was created by the 2000 purchase of U.S. carrier VoiceStream by Deutsche Telekom, a German telecommunications provider spun off from the former state monopoly. *VoiceStream, Deutsche Telekom Seal $50.7B Deal*, REUTERS, July 24, 2000, *available at* http://www.crn.com/news/channel-programs/18809269/voicestream-deutsche-telekom-seal-50-7b-deal-reuters.htm.

157. Thomas W. Hazlett & Robert J. Michaels, *The Cost of Rent-Seeking: Evidence from the Cellular Telephone License Lotteries*, 39 S. ECON. J. 425 (1993).

well.[158] That was an important policy pivot, which pointed the way to further liberalization.

In 1993, with the FCC getting ready to assign new PCS licenses, adding competitors to cellular, Congress ordered that all wireless phone rivals be regulated under a unified Commercial Mobile Radio Service ("CMRS") designation.[159] Cellular, PCS, and SMR—the reinvented taxi dispatcher—would operate as direct rivals with broad, flexible-use spectrum rights.[160] This codified what the FCC had already begun to implement, pre-empted state rate regulation, and for the first time permitted licenses to be assigned through competitive bidding.[161] Chairman Reed Hundt remarked: "[W]e had totally deregulated the wireless industry."[162]

C. MARKETPLACE SPECTRUM ALLOCATION

Carriers seized the opportunity to deploy a range of digital voice technologies—not only TDMA, but also GSM (the standard used most widely in the rest of world), and Code Division Multiple Access ("CDMA") (a rival standard developed by San Diego-based Qualcomm). These technologies have permitted aggressive upgrades over the years to third and fourth generation ("3G" and "4G" in industry parlance) systems that have paved the way for innovative wireless services and devices. GSM networks, which AT&T and T-Mobile deploy, have evolved from EDGE to UMTS to HSDPA technology. CDMA networks, which Verizon and Sprint have deployed, evolved from IS-95 to CDMA2000 1x to EV-DO to EV-DO Rev A. All four wireless carriers have recently begun yet another new upgrade—to Long Term Evolution ("LTE") technology, which offers download speeds to mobile handsets of up to 100 Mbps.

This progression occurs seamlessly, without disturbing network users, but requires vast resources: over $24 billion annually in network capital expenditures.[163] Customers and operators spend billions more on handsets,

158. Amendment of Parts 2 and 22 of the Commission's Rules To Permit Liberalization of Technology and Auxiliary Service Offerings in the Domestic Public Cellular Radio Telecommunications Service, 3 FCC Rcd. 7033 (1988).

159. Omnibus Budget Reconciliation Act of 1993, Pub. L. No. 103-66, Title VI, § 6002(b)(2)(A), 107 Stat. 312, 393.

160. The FCC found that SMR systems providing interconnected service should be classified as CMRS providers. Implementation of Sections 3(n) and 332 of the Communications Act; Regulatory Treatment of Mobile Services, 9 FCC Rcd. 1411, ¶¶ 90–92 (1994).

161. *See* 47 C.F.R. § 24.301 (2009).

162. REED HUNDT, YOU SAY YOU WANT A REVOLUTION 98 (1999).

163. CTIA, *supra* note 2.

an estimated $22.2 billion in 2009.[164] With licensees given wide latitude to choose the technologies deployed and the services offered, firms compete vigorously to improve services, upgrade architectures, cut prices, and provide popular platforms for third party content. Market forces compel efficiency of spectrum use in these bands.[165]

Wireless license auctions have likewise been a success. Since competitive bidding began in 1994, the government has realized $52.6 billion in receipts.[166] These numbers confirm the obvious: wireless service providers will pay substantial sums to avoid having to operate in a "spectrum commons." Any firm has the choice to do otherwise and deploy state-of-the-art radios, spread spectrum devices, mesh Wi-Fi networks, or array antennae—all put forward as exhibits for the proposition that exclusive spectrum rights have outlived their usefulness. Yet, in March 2008, telecommunications firms shelled out $19 billion to acquire exclusive access to 52 MHz of prime frequencies—paying far higher prices than in previous FCC auctions on an adjusted "per MHz-pop" basis.[167] That firms reject available "free" spectrum in unlicensed bands and instead bid aggressively to acquire liberal licenses suggests that firms expect exclusive spectrum ownership to offer productive efficiencies.

The standard explanation of private property rights in the economics literature is that a grant of exclusive control creates incentives for resource conservation and improvement.[168] Liberal spectrum licenses promote precisely these outcomes. Between June 1985 and June 2009, cellular networks built approximately 241,000 cell sites,[169] investing some $274 billion

164. CEA Database, *supra* note 4.

165. Of course, this "property rights" framework is limited both in the extent of the liberalization and in its scope (i.e., allocated spectrum). *See* Kwerel & Williams, *supra* note 8.

166. Press Release, Fed. Commc'ns Comm'n, Statement by FCC Chairman Kevin J. Martin (Mar. 18, 2008).

167. Price comparisons are generally made by adjusting for the bandwidth allocated the license (MHz) and the population in the coverage area of the license (pop). The 700 MHz auction brought an average winning bid of $1.20 per MHz-pop, while the 2006 AWS auction averaged 51 cents. The largest previous auction for broadband PCS licenses, in terms of bandwidth assigned, was the A-B auction concluded in March 1995. It generated total bids of $7.7 billion and an average price equal to 51 cents per MHz-pop. Thomas W. Hazlett, David Porter & Vernon Smith, *Radio Spectrum and the Disruptive Clarity of Ronald Coase*, 54 J.L. & ECON. (forthcoming Nov. 2011) (manuscript at v, 1), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1583098; *see also* Jeremy Bulow, Jonathan Levin & Paul Milgrom, *Winning Play in Spectrum Auctions* (NBER, Working Paper No. 14765, 2009), *available at* http://www.nber.org/papers/w14765.

168. *See generally* Demsetz, *supra* note 52.

169. CTIA, *supra* note 89.

in the process.[170] CMRS network and end-user equipment evolved rapidly, smoothly transitioning from analog to digital service. Advanced compression technologies and smart antennas have been widely deployed, allowing carriers to pack more and more traffic into given bandwidth. For example, CDMA handsets are programmed to check 800 times per second for the lowest possible power level that maintains a link to the base station; GSM over 1,000 times. Dynamic power adjustment reduces spillovers, allowing more phone calls to be made.[171] These innovations and a host of others have resulted in increasingly intense use of frequency space. In its 2004 Comments to the FCC, CTIA noted: "[I]n 1993 a 10 km cell would have averaged fewer than seven subscribers per MHz. In 2003, in the same 10 km area, wireless averaged just under 500 subscribers per MHz."[172]

D. OVERLAYS

Thus, over two decades, great progress has been made in creating exclusive, flexible-use, geographically defined spectrum licenses. CMRS operators have freedom to choose what kind of network equipment to deploy and what services to offer. The approach has also given rise to the regulatory innovation of "overlay" rights. Bands littered with incumbent users, but containing substantially under-utilized "white spaces," are shared with new, encumbered licenses. The new licensee can emit, with broad flexibility, in the allotted frequencies, while respecting the operations of existing users who are grandfathered to continue transmitting. Incumbents are free to "sell out" to the new licensee, moving operations to other bands, fixed links, or ceasing altogether, with the overlay licensee capturing the benefits created.

Overlays effectively cede the task of spectrum reallocation to markets, endowing the overlay licensee with property rights to the fruits of whatever new, innovative band uses it can create. For example, new PCS licensees relocated 4,500 microwave incumbents, paying their moving costs, in order to take full advantage of their spectrum.[173] Likewise, in the 700 MHz band, Qualcomm paid dozens of TV stations on Channels 54, 55, and 56 to accept interference so that Qualcomm's new mobile TV application, MediaFlo,

170. CTIA, CTIA's WIRELESS INDUSTRY INDICES 133, 153 (2009).

171. *Time for Plan B*, ECONOMIST (Sept. 26, 2002), http://www.economist.com/business/displayStory.cfm?story_id=1353050 (discussing how to make 3G work in Europe).

172. Comments of Cellular Internet & Telecomms. Ass'n at 5, *Interference Temperature Metric*, ET Docket No. 03-237 (Fed. Commc'ns Comm'n Apr. 5, 2004), *available at* http://files.ctia.org/pdf/filings/ITEMPcomments.pdf.

173. Cramton, Kwerel & Williams, *supra* note 32, at 668.

could launch nationwide service in 2007.[174] Similar overlays are also used in the AWS frequencies, where licenses were auctioned in 2006.[175]

No alternative property regime allows such efficiencies. In unlicensed bands, the dealmakers necessary for efficient coordination do not exist, which is exactly why such "commons" must rely on spectrum allocations, power limits, and technology mandates pre-set by regulators. Commons advocates have argued that abundant economic unlicensed activity now takes place in what were formerly considered "junk" or "garbage bands," useless for productive activity.[176] But liberally licensed bands have likewise often started with "garbage" and then struck gold by moving thousands of polluters out. The unlicensed bands host valuable applications, but only by powering down, accepting intermittent interference, and living amongst the "garbage." These constraints severely limit the potential of such bands and deter certain unlicensed frequencies, including those allocated to U-PCS and to the 3650–3700 MHz band, from providing net social value. Hamstrung by FCC rules and assigned to no owner, these bands are stuck in the bowels of administrative process.

VI. FROM LMDS TO WIMAX, A REGULATORY ODYSSEY

Consider the evolution of Wireless Metropolitan Area Networks ("WMANs"). These technologies support wireless links that can span distances of up to thirty miles; they can be used to backhaul traffic from local points of aggregation, or to provide last-mile broadband connectivity directly to customers.[177] Both the first-generation WMAN technologies (LMDS and MMDS, for example) and the second (WiMax and Mobile Fi) were designed to operate principally in licensed bands. Unlicensed bands are viewed as viable alternatives only in sparsely populated, rural locations. The treatment of the rival spectrum models, both in regulatory proceedings and in the marketplace, evinces underlying realities about the scarcity of spectrum and socially useful methods to organize radio access.

---

174. Thomas W. Hazlett, *A Law and Economics Approach to Spectrum Property Rights: A Response to Weiser and Hatfield*, 15 GEO. MASON L. REV. 975, 1000–04 (2008).

175. *See, e.g.*, Marketing Materials, Comsearch, Spectrum Sharing and Incumbent Relocation Services for AWS Licensees (2009), *available at* http://docs.commscope.com/Public/Spectrum_Sharing_and_Relocation.pdf.

176. Philip J. Weiser & Dale N. Hatfield, *Policing the Spectrum Commons*, 74 FORDHAM L. REV. 663, 663 (2005).

177. FCC Task Force Paper, *supra* note 17, at 20 ("Wireless Metropolitan Area Networks (WMANs) are point-to-point or point-to-multipoint networks with individual links that not only can span distances of up to 30 miles, which is important for backhaul applications, but also can provide last-mile connectivity in metropolitan environments.").

A. EARLY WMAN INVESTMENTS AND THE EMERGENCE OF FIXED WIRELESS TECHNOLOGIES

The first major wave of investment in WMAN technologies occurred shortly after passage of the Telecommunications Act of 1996.[178] Between 1997 and 2000, the FCC licensed several large parcels of spectrum for these services—the 24 GHz band allocated for Digital Electronic Messaging Service ("DEMS"),[179] the 27–31 GHz band allocated to Local Multipoint Distribution System,[180] and the 39 GHz band.[181] These high frequencies provide very generous bandwidth, but signals carry only a couple of miles and require a clear line of sight.[182] Companies like Teligent, WinStar, and NextLink (later XO) bid aggressively to acquire licenses.[183] By year-end 2001, these companies had spent over $10 billion building out their networks.[184] But the radios capable of handling these high frequencies were very

178. Pub. L. No. 104-104, 110 Stat. 56 (1996).

179. DEMS services were initially allocated spectrum in the 18 GHz band. This allocation was then moved to a different section of the same 18 GHz band. DEMS was finally relocated to the 24 GHz band in 1997. *See* Amendments to Parts 1, 2, 87 and 101 of the Commission's Rules to License Fixed Services at 24 GHz, 15 FCC Rcd. 16,934, ¶¶ 3–4 (2000).

180. *See Factsheet for Auction 17: Local Multipoint Distribution System (LMDS)*, FED. COMMC'NS COMM'N (Nov. 29, 2007), http://wireless.fcc.gov/auctions/default.htm?job=auction_factsheet&id=17. There are two LMDS licenses issued in each of 493 Basic Trading Areas (BTAs). Frequency Block A licenses are for 1.15 GHz in the 27.5–28.35 GHz, 29.1–29.25 GHz, and 31.075–31.225 GHz bands; Frequency Block B licenses are for 150 MHz in the 31–31.075 GHz and 31.225–31.300 GHz bands. *Id.*

181. *See Factsheet for Auction 30 (39 GHz)*, FED. COMMC'NS COMM'N (July 27, 2006), http://wireless.fcc.gov/auctions/default.htm?job=auction_factsheet&id=30 ("39 GHz licensees may provide fixed communications including point-to-point and point-to-multipoint communications.").

182. Presentation, K. Wanichkorn, Thailand Rural Wireless Broadband Access Initiative (APT Regional Forum for ICT Experts in South-East Asia, Feb. 4–5, 2004) (on file with authors) ("High Microwave Frequencies (>10GHz)" work at "[s]hort propagation distances (3–5km) and require line-of-sight"); Angela Langowski, *LMDS Hits the Spot?*, CED MAGAZINE, Feb. 1, 2001, *available at* http://www.cedmagazine.com/ced/2001/0201/id2.htm ("[LMDS] systems operate only over short distances.").

183. Langowski, *supra* note 182.

184. *See, e.g.*, Neil Weinberg & Michael Maiello, *Malone Clone*, FORBES (Apr. 15, 2002), http://www.forbes.com/forbes/2002/0415/082.html ("[Howard Jonas] bought Winstar Communications out of bankruptcy for $42 million, gaining control of a broadband wireless network that had cost $5 billion to build."); Jim Barthold, *Teligent Faces New Sober Reality*, CONNECTED PLANET (Sept. 16, 2002, 12:00 PM), http://connectedplanetonline.com/mag/telecom_teligent_faces_new/ (quoting Teligent Marketing VP, Denisse Goldbarg, as saying, "We had $2 billion to $3 billion thrown at us."); Press Release, Infospace, Inc., Infospace Adds to Senior Management Team (Apr. 2, 2003), http://phx.corporate-ir.net/phoenix.zhtml?c=119056&p=irol-newsArticle&ID=397450 (noting that NEXTLINK raised more than $4 billion in capital).

expensive and much less reliable than wireline alternatives.[185] These first-generation services all failed when Wall Street's dot-com bubble collapsed.[186]

As markets recovered and technologies for utilizing these bands improved,[187] the licenses returned to investors' radar screens.[188] Firms with new business models acquired the licenses in the 24, 27–31, and 39 GHz bands. Companies like First Avenue Networks[189] and IDT[190] began leasing this spectrum wholesale to providers of high-speed internet access, mobile carriers providing backhaul services, and wireline carriers building wireless extensions to their fiber-optic networks. This reallocation of spectrum to new and hopefully more valuable employments is an automatic function of exclusive rights, even through the disruption of bankruptcy. Indeed, trial and error is a socially useful discovery process when incentives guide investors to place the best bets, generating "creative destruction" that iterates on new efficiencies and produces technological disruptions of its own.

---

185. *See, e.g.*, *Intel and Clearwire Forge WiMAX Alliance*, THE REGISTER (Oct. 29, 2004, 11:08 AM), http://www.theregister.co.uk/2004/10/29/intel_clearwire_wimax/ ("The most critical element in the bursting of the last BWA bubble was the cost of subscriber equipment. . . . [T]he critical stumbling block for Winstar [was] the expensive, proprietary subscriber equipment.").

186. M. MCCORMACK & PHIL CUSICK, BEAR STEARNS, WIRELESS BROADBAND: THE IMPACT OF 802 TECHNOLOGY 17, ex. 7 (June 2004) (on file with authors).

187. *Id.* at 25 (noting that prices for CPE are still high, but "[t]he consensus from our interviews is that prices will be halved over the next two years, reaching a $100 price point within five years"); *id.* at 26 (remarking that the current generation of wireless broadband equipment is "greatly improved over equipment from the 1990s").

188. *See, e.g.*, Press Release, First Ave. Networks, First Avenue Networks Closes Acquisition of Teligent Assets, PR NEWSWIRE (Jan. 18, 2005), http://www.thefreelibrary.com/First+Avenue+Networks+Closes+Acquisition+of+Teligent+Assets.a0132593914; M. Dano, *Nextel Adds to MMDS Spectrum*, RCR WIRELESS NEWS (Nov. 10, 2003, 6:00 AM), http://www.rcrwireless.com/article/20031110/SUB/311100739/-1/nextel-adds-to-mmds-spectrum. In 2006, First Avenue was acquired by FiberTower. Press Release, First Ave. Networks, FiberTower Corporation and First Avenue Networks Announce Merger Agreement (May 15, 2006), http://www.fibertower.com/corp/downloads/investors/FT_FAN_Release.pdf.

189. *See* Press Release, Covad, First Avenue Networks and Covad Announce Metro Wireless Reseller Agreement (June 5, 2006), http://www.covadwireless.com/pdf_files/06_05_06_FirstAveNetworks.pdf ("First Avenue provides wireless backhaul and Carrier Ethernet services over its licensed spectrum footprint, which includes nationwide 24 GHz and 39 GHz holdings."); Press Release, IDT Corp., IDT Spectrum, Inc. Names Peter B. Atwal, President, Engineering and Operations (Apr. 28, 2005), http://www.idt.net/about/press/story.aspx?id=9336 ("IDT Spectrum, Inc. is the recently formed subsidiary of IDT Corporation that operates and markets wireless spectrum products and solutions.").

190. Press Release, IDT Corp., IDT Corp. Announces the Acquisition of Winstar Communications, Inc. (Dec. 20, 2001), http://www.idteurope.com/corporate/press/releases/246.asp.

A new family of fixed wireless technologies—commonly known as WiMax—emerged in parallel. The original WiMax standard (IEEE 802.16) was developed to operate in licensed bands between 10 and 66 GHz.[191] Recent versions of the standard support both licensed and unlicensed operations between 2 and 11 GHz,[192] including licensed bands at 2.5 GHz and 3.5 GHz, and the unlicensed 5.8 GHz band that Wi-Fi radios often use.[193]

Operators that will use licensed spectrum, however, dominate the new investment; equipment manufacturers and service providers have both concluded that this is where the most promising opportunities lie. Clearwire has aggregated licenses in the 2.5 GHz band pursuant to a series of FCC rulings expanding licensee rights to allow two-way, cellularized, broadband services.[194] By year-end 2010, it was offering wireless broadband service in at least sixty-eight markets, including Anchorage, Jacksonville, Waco, Reno, and Rochester, New York,[195] and deploying 4G networks in New York City, Seattle, Washington, D.C., Los Angeles, Cleveland, and Cincinnati, among other cities.[196] And it served 4.4 million subscribers,[197] at least *100 times* any WISP relying on access to unlicensed airwaves.[198]

191. Roger Marks et al., *IEEE 802.16 Backgrounder*, IEEE (May 24, 2002), http://ieee802.org/16/pub/backgrounder.html.

192. *Id.*

193. *See* Intel, *Understanding Wi-Fi and WiMAX as Metro-Access Solutions* 10 (2004), http://www.rclient.com/PDFs/IntelPaper.pdf.

194. For the long regulatory history of 2.5 GHz licenses, see Hazlett, *supra* note 28.

195. *See Clearwire Markets*, CLEARWIRE CORP., http://media.corporate-ir.net/media_files/irol/21/214419/mediakit/Market_List_12309.pdf (last visited May 3, 2011).

196. Don Reisinger, *Clearwire Extends 4G Rollout to Five More Markets*, CNET NEWS (Nov. 29, 2010), http://news.cnet.com/8301-13506_3-20023955-17.html.

197. Dan Meyer, *Clearwire Posts Strong 4Q Wholesale Growth, Cost Conservation Remains Intact*, RCR WIRELESS NEWS (Feb. 17, 2011), *available at* http://www.rcrwireless.com/article/20110217/CARRIERS/110219933/1097.

198. In its last ranking of Top Wireless ISPs, Broadband Wireless Exchange Magazine listed SpeedNet the third largest WISP with 15,000 customers. The two higher ranking ISPs (one of which was Clearwire) both use licensed spectrum in the 2.5 GHz band; it was not specified whether SpeedNet used licensed, unlicensed, or both. *See "Top Ten" Wireless Internet Service Providers*, BROADBAND WIRELESS EXCH. MAGAZINE (2007), http://www.bbwexchange.com/wireless_isp/; Kenyon Commc'ns Holdings, Inc., Quarterly Report (Form 10-Q), at 13 (June 30, 2009).

C. CARRIERS ADVOCATE FOR WIMAX

Carriers that have used unlicensed spectrum have done so predominantly in rural areas.[199] Intel, often a strong advocate of unlicensed spectrum, has concluded that, to supply "carrier class QoS [Quality of Service]," WiMax requires licensed spectrum.[200] "In general," Intel concluded, "unlicensed bands can be subject to [service quality] issues because deployment is open to anyone."[201] In a 2007 paper, Intel outlined the differences between Wi-Fi and WiMax this way:

> **Licensed to Thrill**. . . . Unlicensed spectrum is open to any users, which raises the possibility of interference from other devices. Wi-Fi networks use unlicensed spectrum. Wimax service providers use licensed spectrum which allows exclusive rights to its use for more predictability and stability.[202]

These very factors are what yield incentives for firms to sink substantial capital in creating spectrum complements—namely, wide area wireless networks. Clearwire itself began, in 1999, as a WISP operating "in the unlicensed 2.4 GHz frequency, which was subject to interference." It found that it could lease licensed bandwidth from educational institutions with Instructional Television Fixed Service ("ITFS") licenses, upgrading its service and allowing it to craft an ambitious service strategy.[203] Liberalization of the rights permitted in the ITFS licenses made this set of transactions possible.[204]

In that pursuit Clearwire has attracted powerful economic support, registering over $4 billion in capital infusions from Intel, Motorola, Bell Canada, Google, Sprint, Comcast, Time Warner, Bright House Networks, and thousands of equity investors buying the firm's shares in its 2007 IPO.[205] The financial meltdown of 2008–2009 adversely impacted share prices; much

---

199. *See* INTEL CORP., DEPLOYING LICENSE-EXEMPT WIMAX SOLUTIONS 6 (2005), *available at* http://empoweringsoho.files.wordpress.com/2007/07/deployingwimaxlicense exemptband.pdf ("[L]icense-exempt WiMAX solutions are focused on rural areas . . . .").

200. *Id.* at 6–7; *see also* MCCORMACK & CUSICK, *supra* note 186, at 8 ("In our view, the use of licensed spectrum is necessary to guarantee a level of service and availability for the paying portable user. The use of unlicensed spectrum could lead to wide disparities in quality of service, bottlenecks, and security issues.").

201. Intel, *supra* note 193, at 3.

202. Intel Corp., *Welcome to Your Internet Future* 15 (2007), http://download.intel.com/network/connectivity/products/wireless/welcome-to-your-internet-future.pdf.

203. *Clearwire, Inc.*, FUNDING UNIVERSE, http://www.fundinguniverse.com/company-histories/Clearwire-Inc-Company-History.html (last visited Jan 21, 2011).

204. *See generally* Hazlett, *supra* note 28.

205. *Id.* at 256.

of Clearwire's strategic investment was written off in early 2009.[206] This is entirely consistent with the social efficiencies created by such investment incentives. Only when firms suffer the adverse consequences of the risks they take will their incentives be fully aligned with the interests of consumers. When, conversely, Apple Computer lobbied the FCC for an U-PCS frequency allocation in the early 1990s, prompting the Commission to allocate 30 MHz for U-PCS that has gone essentially unused for over a decade, the loss was socialized. Apple internalized only the cost of its lobbying.[207]

D. FCC'S REGULATION VIA REGISTRATION HAS CREATED INEFFICIENCIES THAT LIBERAL LICENSES WOULD NOT

The FCC's recent proceeding to allocate the 3.65 GHz band for non-exclusive access rights was initiated at the behest of wireless WISPs who "expressed a clear need for additional spectrum for broadband use . . . especially in rural areas."[208] The FCC was persuaded, stating that allocating the band for unlicensed devices "would be the most beneficial approach."[209] Still, rural WISPs were concerned that "intense use of spectrum by a variety of devices under a traditional unlicensed approach could result in mutual interference, thereby reducing the utility of this band."[210] In 2005, the Commission ultimately adopted a regime that eschews assigning any spectrum "for the exclusive use of any licensee"[211] and instead directs all licensees "to cooperate and avoid harmful interference to one another."[212]

206. *See, e.g.*, Stacey Higginbotham, *Intel Writes Almost $1B Off Clearwire Investment*, GIGA OM (Jan. 7, 2009), http://gigaom.com/2009/01/07/intel-writes-almost-1b-off-clearwire-investment/. Overall, capital losses are approximated by comparing the capitalization estimated by Wall Street analysts when the company was acquiring investment partners (and public investors in an IPO), to the current enterprise value. The former valuation was $14.5 billion, the latter is $3.5 billion. Erick Schonfeld, *$3.2 Billion Wimax Deal Goes Through. Take Cover*, TECHCRUNCH (May 6, 2008), http://techcrunch.com/2008/05/06/32-billion-wimax-deal-goes-through-take-cover/; *Key Statistics for Clearwire (CLWR)*, YAHOO! FINANCE, http://finance.yahoo.com/q/ks?s=CLWR+Key+Statistics (last visited Feb. 27, 2011) (listing enterprise value at $3.54 billion as of Feb. 27, 2011).

207. Thomas W. Hazlett, *The Spectrum-Allocation Debate: An Analysis*, 10 IEEE INTERNET COMPUTING 68, 73–74 (Sept.–Oct. 2006).

208. *Wireless Operations in the 3650–3700 MHz Band*, 20 FCC Rcd. 6502, ¶ 13 (2005).

209. *Id.*

210. *Id.* ¶ 14.

211. *Id.* app. A.

212. *Id.* ¶ 16. All licensees are required "to cooperate and avoid harmful interference to one another." *Id.* All WiMAX radios must incorporate protocols to determine who gets priority when "when two or more devices attempt to simultaneously access the same channel." *Id.* These same protocols will "establish[] rules by which each device is provided a

Everyone operating in these bands must register "their fixed and base stations in a common database."[213] And any interference between these high-power transmitters "will be addressed by the process we adopt to register fixed and base stations."[214] Registration rules require base stations to "operate at locations and with technical parameters that will minimize the potential for interference between stations."[215]

The better this registration policy works in the short term, the faster its fundamental shortcomings will become apparent. Many companies certainly want to offer WiMax service, competing in the $32 billion per year U.S. market for DSL and cable modem services.[216] Spectrum-sharing protocols in WiMax radios can provide for orderly, reasonable sharing of spectrum, under certain circumstances. But they do not limit how many radios are deployed to share it.

Just as power limits are an implicit exclusionary device, a registration policy is likewise an acknowledgement of the benefits of coordinating spectrum users and a rejection of "open access." The unlicensed space is in some sense licensed, but without the benefit of de facto (or de jure) spectrum ownership. Under this hybrid approach, there are no claimants to seek gains from efficient spectrum reallocation—for example, to clear the 3650 MHz band of the satellite operations that could impinge on the use of WiMax devices. In a liberal license regime, licensees could pay satellite operators to move to alternative bands, thus expanding opportunities for WiMax. Under the hybrid approach, it is regulators, not market forces, that set sharing rules.

What advantage does such state control afford society? The reflex response—that it protects spectrum access on a no-fee basis—is strictly correct. But this regime can render "free" access a high-cost failure. If the coordination supplied by exclusive rights owners results in networks and applications that are superior from the viewpoint of consumers, but fail to materialize here, then the incremental social benefits of the non-exclusive rights are negative.

U.S. regulators, as Coase deduced from basic economic theory, fail to properly account for the opportunity costs of alternative allocations when

reasonable opportunity to operate." *Id.* By "control[ling] access to spectrum, terrestrial operations will avoid interference that could result from co-frequency operations." *Id.*

213. *Id.*

214. *Id.*

215. *Id.*

216. Ray Le Maistre, *US Tops Broadband Revenues Chart*, LIGHT READING (Jan. 2, 2009), http://www.lightreading.com/document.asp?doc_id=169812 ("According to data compiled for the report, Global Fixed and Mobile Broadband Outlook, the United States generated more than $32 billion in broadband revenues in 2008.").

zoning spectrum. The results are starkly on display in the 3650 MHz proceeding. What is called the 3.5 GHz band includes the 3650–3700 MHz frequencies; it is the most popular location for *licensed* WiMax networks internationally.[217] Indeed, the IEEE 802.16 (WiMax) technology protocol at 3.5 GHz has been written for exclusive rights holders. U.S. regulators chose to deviate from world markets, a decision strongly opposed by Intel and Alvarion, firms with substantial sales of wireless hardware using the 2.4 GHz and 5 GHz unlicensed bands. Yet, the companies jointly urged the FCC to license 3650 MHz spectrum in the top fifty U.S. markets—areas with the lion's share of potential customers and those in which potential airwave conflicts are most intense.[218] They lost that battle,[219] and the only large-scale WiMax developments observed in the U.S. market today are those in the licensed 2.5 GHz band.

VII. THE QUIET PAST AND THE NOISY FUTURE

There was no demand for spectrum before 1895 because nobody yet knew how to build a radio. Guglielmo Marconi demonstrated the basic engineering that year, but it would take another thirty years to develop products and business models to move hundreds of broadcast stations and millions of receivers into the mass market—enough radio hardware to spawn the conflicts that impelled the creation of a legal regime to police spectrum access.[220]

To frame this history in terms relevant to the current debate, licenses matter when radio technologies and markets evolve to the point where things get crowded. The commons advocates insist that when the technology is smart enough, things *never* get crowded. That story is exactly backwards. Setting aside regulatory barriers, it is the lack of technology that has left some bands relatively empty. Bands that were empty a decade ago are crowded today in large measure because affordable new products have arrived to fill them. In our frame of experience, technology is not the solution to spectrum scarcity, but its cause. As wireless technologies become smart, cheap, and

217. Sam Churchill, *Italy Opens 3.5 GHz for WiMax*, DAILY WIRELESS (Dec. 29, 2006, 10:07 AM), http://www.dailywireless.org/2006/12/29/italy-opens-35-ghz-for-wimax/.

218. Petition of Intel Corp. et al., Petition for Reconsideration of Intel Corporation, Redline Communications, Inc., Alvarion, Inc., ET Docket No. 04-151 (Fed. Comm'ns Comm'n June 10, 2005) (arguing that the FCC's refusal to license 3650 MHz spectrum would lead to squatting, inefficiency, and delay).

219. Jerry Brito, *The Spectrum Commons in Theory and Practice*, 2007 STAN. TECH. L. REV. 1, ¶¶ 75–78.

220. On the origins of these conflicts and the policy course taken, see Hazlett, *supra* note 121.

ubiquitous, the social value of property rights—helping to create the platforms on which such burgeoning economic activity can be best accommodated—rises.

We face a world of spectrum scarcity looking forward. A radio's power amplifier is the toughest part to build, and the higher the frequency, the tougher it gets. Twenty years ago, no one knew how to build the high-frequency gallium arsenide chip-scale amplifiers that now power many Wi-Fi radios. Wi-Fi radios operate at 2.4 and 5 GHz, and cheap, compact chip-scale amplifiers capable of handling such high frequencies had not been developed until around that time. Military research funded the development of this exotic semiconductor,[221] and licensed wireless carriers created the mass-market demand by selling millions of cell phones.[222] In that sense, Wi-Fi took a free ride on technological innovation in licensed bands.

Unsurprisingly, the history of FCC licensing has tracked the progressive march of radio-amplifier technology up the frequency ladder—albeit with significant, wealth-destroying lags.[223] And the FCC's allocation of spectrum for mass-market licensed services has largely tracked the frequency-climbing evolution of radios.[224] The amount of usable spectrum continues to expand because engineers continue to push radio frequencies up into higher bands, and radio costs down. The spectrum always looks uncrowded to pioneers at the very top of the ladder. Then, when costs drop and regulatory barriers fall, crowds follow.

---

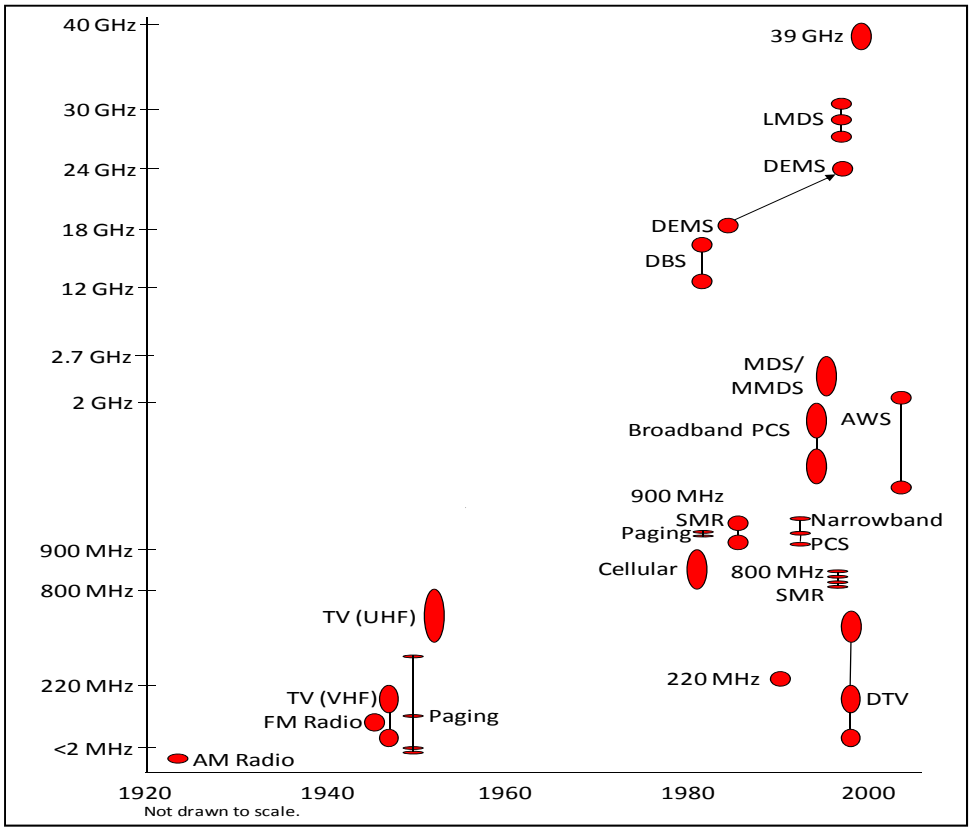
221. *See* Press Release, TRW, TRW To Fabricate Advanced Integrated Circuits for RF Micro Devices (Dec. 7, 1993) ("Much of TRW's [gallium arsenide (GaAs)] chip manufacturing expertise was developed as part of the Microwave and Millimeter Wave Monolithic Integrated Circuits (MIMIC) program sponsored by the Department of Defense's Advanced Research Projects Agency (ARPA). . . . The MIMIC program was begun in 1987 by ARPA to make GaAs integrated circuits producible, affordable and applicable to a wide range of critical defense system needs.").

222. *See, e.g.*, Tim Whitaker, *SiGe and CMOS Target GaAs Dominance of Cellular PA Slots*, COMPOUND SEMICONDUCTOR (May 4, 2004), http://compoundsemiconductor.net/csc/features-details.php?id=19397 ("[I]n the largest market segment, [power amplifiers] for cellular handsets, GaAs enjoys almost total dominance, and accounted for well over 95% of handset [power amplifier] revenues [in 2003], according to market research firm ABI Research.").

223. Cellular phone allocations were delayed from the late 1940s, when cellular technology was developed by Bell Labs, until licenses were distributed in the 1980s—one notable example of regulatory lag. *See* GEORGE CALHOUN, DIGITAL CELLULAR RADIO 39–49 (1988).

224. *See infra* Figure 2.

Figure 2: Progressive Allocation of Spectrum for Commercial Mass-Market Wireless Services



The bands occupied today have an additional feature: many high-frequency transmissions are easily blocked by physical obstacles, rain, and so forth. DBS satellites, for example, broadcast at 12 GHz. At this frequency, signals are blocked by foliage, so that pizza-sized receiving antennas require a treeless line of sight to the southern sky. Heavy rain can also block these signals. By contrast, the Navy's ultra-low-frequency radios can communicate with submerged submarines. That high-frequency signals are easily blocked is a schizophrenic blessing, as the Wi-Fi experience again teaches. One Wi-Fi radio is unlikely to interfere with another when shielding supplied by walls or windows sharply limits its range. But that same limit is what makes it relatively costly to scale Wi-Fi deployments for many applications larger than a Starbucks.

A. CONFLICTS IN RADIO COMMUNICATIONS

Smart radios have overcome the problem of interference, commons advocates allege. Unlicensed wireless devices may not have legal protection against interference, but they do not need any. It is possible that licensing

spectrum may occasionally be useful, just as "toll roads or paid carpool lanes" sometimes make sense "in some predictably congestion-prone roads."[225] But mostly we are dealing here with "city streets and sidewalks, dirt roads, or highways at nighttime."[226] Smart radios are like ships traveling on an ocean of spectrum—not infinite in size, perhaps, but so vast that vessels can simply "be trusted to navigate around one another."[227]

These conclusions are said to flow from the laws of physics: "[I]nterference is a consequence of system design, rather than an inherent property of the radio spectrum."[228] Interference is not "physical" but "inherently a legal construct."[229] "Interference is a metaphor that paints an old limitation of technology as a fact of nature."[230] "Spectrum is not scarce. We're talking about radio waves. Radio waves run through one another."[231] "More than one service can occupy the 'same' spectrum, in the same place, at the same time."[232] "The electromagnetic waves do not actually bounce off each other or 'fall' to the ground before reaching the receiver's antenna. 'Interference' describes the condition of a stupid lone receiver faced with multiple sources of radiation that it is trying to decode but, in its simplicity, cannot."[233] "Radio waves do not . . . cancel each other out."[234]

But, in fact, they do.

1. Noise and Interference

Thomas Young's double slit experiment, first conducted in 1801 and repeated in high-school physics classes to this day, establishes that electromagnetic waves interact, interfere, amplify, and obliterate each other like waves on the surface of a pond. Two clean signals superposed become one messy one; throw in more and you end up with a cacophony of pure

225. Benkler, *supra* note 17, at 69.

226. *Id.* at 69, 32–33 (stating that property rights and pricing mechanisms are "useful only occasionally, at peak utilization moments").

227. WERBACH, *supra* note 17, at 20.

228. *Id.* at 14.

229. *Id.*

230. David Weinberger, *The Myth of Interference*, SALON (Mar. 12, 2003), http://dir.salon.com/story/tech/feature/2003/03/12/spectrum/index.html? (quoting David Reed).

231. Heath Row, *The Open Spectrum Revolution*, FAST COMPANY (Mar. 13, 2004), http://www.fastcompany.com/blog/heath-row/open-spectrum-revolution (quoting Kevin Werbach).

232. WERBACH, *supra* note 17, at 3.

233. Benkler, *supra* note 17, at 39.

234. WERBACH, *supra* note 17, at 5.

noise. No amount of additional intelligence embedded in the receiver can reverse the process when interference transforms information into chaos.[235]

In fact, radios dispatch streams of energy from their antennas, and that energy propagates through the surroundings at the speed of light. These fluxes are not legal constructs,[236] but physical things. In a microwave oven, they heat soup. When they strike a silicon-crystal solar cell, electromagnetic energy at a slightly higher frequency can generate electricity. Suitably synchronized, flukes like these become a maser or laser that can cut through steel. And these same energy fluxes—streams of photons of dual wave-like and particle-like nature—interact, deflecting flows and destroying communications.

Thus, for example, microwave ovens cause "noticeable" interference with Bluetooth devices operating nearby.[237] Bluetooth devices interfere with each other.[238] Cell phone jammers are readily available (though illegal[239]); it is equally easy to jam Wi-Fi nodes. The most common form of interference

235. The concession is made that interference among smart radios is a "realistic possibility" when "very large numbers of such devices operate in the same location." *Id.* at 17. That this obviates the claim that interference is a myth, or that it fully supports the economic analysis of spectrum scarcity is, however, lost.

236. Were radio signals "only" legal constructs, the argument over how best to assign property rights would not be decided, of course. Intangible property rights, including those created in contract law, tend to dominate ownership institutions in advanced economies. More generally, all property rights are legal constructs and govern not things but relations between people.

237. T.W. Rondeau, M.F. D'Souza & D.G. Sweeney, *Residential Microwave Oven Interference on Bluetooth Data Performance*, 50 IEEE TRANSACTIONS ON CONSUMER ELECS. 856, 863 (2004).

238. J.E. Ballagh, T.W. Rondeau & D.G. Sweeney, *Bluetooth Frequency Hop Selection Kernel Impact on 'Inter-Piconet' Interference*, IEEE TRANSACTIONS ON CONSUMER ELECS. (forthcoming); *see also* J. Lipman, *The "Other" Wireless Technology Is Alive and Kicking*, EE TIMES (Oct. 14, 2003), http://www.techonline.com/community/ed_resource/feature_article/28419 ("[S]imilar to devices operating in the 900 MHz spectrum interfering with other like-frequency devices, Bluetooth is more prone to interference from other Bluetooth devices, cell phones, microwave ovens, and other equipment operating at 2.4 GHz."); Jonathan Miller, *Tips on Using WiFi in RV Parks*, RVERS ONLINE (Mar. 2005), http://www.rversonline.org/RVWiFi.html ("From 4:30 to 6:00 pm each night at every RV resort in the country the WiFi signal is impeded the most. This is due to the concentration of RVs in a relatively small area preparing dinner and using Microwave ovens. The same is true in apartment complexes.").

239. The manufacture, importation, sale, and operation of transmitters designed to jam or block signals is a violation of 47 U.S.C. §§ 301, 302(a), 333 (2006), and is subject to severe penalties. "Fines for a first offense can range as high as $11,000 for each violation or imprisonment for up to one year." *Cellular Services, Operations: Blocking & Jamming*, FED. COMMC'NS COMM'N, http://wireless.fcc.gov/services/cellular/operations/blockingjamming.html (last updated Nov. 19, 2002).

arises when an emission from a single transmitter interferes with itself. This can occur when part of a signal travels directly from the tower to the television, and part travels indirectly, reflecting off (say) a nearby skyscraper. Two different electromagnetic signals of the same frequency cannot in fact coexist at exactly the same place and time.

All else equal, the noisier the electromagnetic environment, the longer it takes to transmit information through it without distortion. This too is a fundamental law of physics[240] that engineering cannot repeal. To state the same law another way: to get through at all, a radio transmission has to be powerful enough to penetrate the ambient noise.

The "ambient noise" itself is highly variable. It is composed of all the radio transmitters—"intentional emitters," in FCC jargon—and all the "incidental" and "unintentional" emitters, including virtually every device that runs on AC electric power. It is therefore impossible to know precisely how noisy things will be along any given pathway, at any given point in time; getting a signal through is a fundamentally chancy business.[241] One can improve the odds of getting a signal through by raising the power of one's radio, or lowering the power of other radios transmitting on the same frequencies, or shutting down competing radios altogether. One can switch one's radio to a different band, which may be quieter. And one can transmit the same message more than once, or simultaneously on multiple bands.

But none of these strategies eliminates fundamental economic trade-offs. Quite the contrary—*every* strategy that boosts the chances of punching your own signal through the airwaves either adds to the expense of the communications conducted, lowers the odds for every other radio that is trying to do the same, or both. It is suggested that by searching for unused gaps in the airwaves, "agile radios can in effect *manufacture* new spectrum."[242] "In effect," correct, but *in fact* radios transmit radio signals, and when they do that, they do not produce spectrum, they consume it.

The much-heralded smartness of Wi-Fi radios provides a case in point. When two Wi-Fi radios use the same (limited) channels in close proximity, each device detects the presence of a competing transmitter and adjusts by transmitting more slowly and using more power to send each bit. As Hewlett-Packard describes it, Wi-Fi radios "fail gracefully in the presence of interference"—the "result of increasing levels of interference is almost

240. *See* C.E. Shannon, *A Mathematical Theory of Communication*, 27 BELL SYS. TECH. J. 379 (1948), *available at* http://www.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-379.pdf.

241. Weiser & Hatfield, *supra* note 176.

242. WERBACH, *supra* note 17, at 19 (emphasis added).

always confined to a slowing of the data rate as more packets need to be present."[243] Wi-Fi does not eliminate the interference problem but degrades performance to accommodate it. This is more acceptable for certain types of data transmission where it does not matter too much how fast the traffic gets through, such as web browsing, while it is less acceptable for voice and other interactive applications that require steady throughput.

These limitations are costly. Moreover, the coordination between users that may mitigate these costs is difficult to achieve. Wi-Fi radio users, particularly those who attempt outdoor deployments, are advised to seek out other Wi-Fi users and gain their cooperation, using different channels and placing facilities in complementary locations.[244] Indeed, WISPs may find themselves in "broadcast wars," where transmissions to occupy Wi-Fi channels, along with the use of higher power levels, are strategic tools used to lower rivals' quality of service.[245] Degraded performance is only the tip of the iceberg; networks never deployed due to the costs of coordination in this space constitute the largest losses. The tragedy of the commons is the unobserved counter-factual.

2. Physics and Architecture

Radio waves are real things transmitted with real energy.[246] Potential conflicts depend on the separation, if any, between the band being used by the interfering transmitter, and the band that the unwitting receiver is trying to decipher. It depends on the power of the rival transmitter, its proximity,

243. HEWLETT-PACKARD, WI-FI AND BLUETOOTH—INTERFERENCE ISSUES 1 (2002), *available at* http://www.hp.com/rnd/library/pdf/WiFi_Bluetooth_coexistance.pdf.

244. Tim Pozar, Regulations Affecting 802.11 Deployments, Version 1.5 (Mar. 10, 2004) (unpublished manuscript), *available at* http://www.lns.com/papers/part15/Regulations_Affecting_802_11.pdf.

245. Christian Sandvig, The Return of the Broadcast War 23 (Sept. 23, 2005) (unpublished manuscript), *available at* http://www.communication.illinois.edu/csandvig/research/Broadcast_War.pdf (quoting one WISP operator as saying, "The more channels I grab means the less competition").

246. Although the concepts are often used interchangeably, it is here important to distinguish between "spectrum," which itself is not a physical thing, and the radio waves or signals that are transmitted through space. *See* Howard A. Shelanski & Peter W. Huber, *Administrative Creation of Property Rights to Radio Spectrum*, 41 J.L. & ECON. 581, 584 (1998) ("There is no such thing as 'spectrum' out there, any more so than there was 'ether,' to be bottled by the Commission or anyone else. 'Spectrum' is composed entirely of the engineering characteristics of transmitters and receivers. Those characteristics are defined, in turn, by power, sensitivity, and modulation parameters in a fuzzy and permeable zone of space."). Thus, when we talk about the scarcity of spectrum, this is shorthand for the conflicts resulting from the transmission of radio waves within defined frequency spaces.

and on directional antennas mounted (or not mounted) on the transmitter, the unwitting receiver, or both.

The interference problem is defined by the *aggregate* of all competing transmitters, i.e., by the transmission frequencies, power, proximity, and antenna configurations of *all* intentional, incidental, and unintentional interfering transmitters in the band, and all the buildings, foliage, fog, meteor trails, and extraterrestrial radiation belts, that may reflect their signals (usually aggravating the problem) or attenuate them (usually mitigating it). A licensee with liberal spectrum rights uniquely possesses the information, economic incentive, and financial ability to optimize the architecture to make the best possible use of such a complex, turbulent resource.

As Ronald Coase noted decades ago, the appropriate quest is not to eliminate interference between wireless users, which is overly protective, but to achieve just the proper amount.[247] How much interference is tolerable can never be defined by the interfering transmitter—it must, self-evidently, be defined by the receiver. Firefighters groping their way through an inferno tolerate less interference than teenagers text-messaging at the mall. "Tolerable" itself is inevitably defined in statistical terms: how often will noise levels rise high enough to block a transmission, and what blocking probability is acceptable given the cost of reducing the odds?

The argument that radio waves do not interfere, that spectrum is not scarce, implies that the only quality of service issue lies in the design of the individual receiver. A race for better radios then becomes the alternative to network coordination.[248] When interference is explained as a problem of insufficient processing power in the radio receiver, individual users are portrayed as efficiently pursuing—with help from equipment vendors, who profit from selling better radios—the optimal level of reception (the reciprocal of interference).

This mischaracterizes the economic problem, which centers on how to entice all productive contributions where benefits exceed costs. Whatever the incentive of an individual to buy a radio that punches through the din, and the vendor to profit from selling that device, the resulting transactions will not take account of the costs transmissions impose on other users. This loads all interference mitigation on the individual user's radio and government rules (such as power limits), with decisions taken in isolation from other sources of conflict. While mobile carriers, as spectrum owners, aggressively

247. "It is sometimes implied that the aim of regulation in the radio industry should be to minimize interference. But this would be wrong. The aim should be to maximize output." Coase, *supra* note 21, at 27.

248. *See* WERBACH, *supra* note 17, at 37; Benkler, *supra* note 17.

deploy systems that make radios quieter and spillovers smaller, unlicensed device users predictably pollute the airwaves by blasting signals at higher power than necessary so as to guarantee safe passage for their communications.[249] In short, each user competes to claim control over airspace, insensitive to the costs imposed on others. This is standard tragedy of the commons,[250] and there is considerable concern among engineers about how to deter it in unlicensed bands.[251] This inevitably involves some level of enhanced coordination, public or private.[252]

---

249. Radio engineers who work with Wi-Fi commonly acknowledge costs of non-exclusivity. Using 5.8 GHz frequencies to bring broadband connections to a housing development in a low-income part of San Francisco, Tim Pozar notes that "[f]requency coordination is a constant problem," as new radios disrupt existing links and there is "no way to encourage or to enforce coordination." Presentation, Tim Pozar, A Sample Wireless Broadband Deployment—City of San Francisco Housing Projects, at slide 26 (Apr. 4, 2008), http://www.iep.gmu.edu/documents/GMU-Pozar-20080404.pdf. Pozar provides a useful basic description of unlicensed usage:

> As 802.11 [Wi-Fi] is designed for short-range use, such as in offices and homes, it is limited to very low power. Ideally, a well-engineered path will have just the amount of power required to get from point 'A' to point 'B' with good reliability. Good engineering will limit the signal to only the area being served, which both reduces interference and provides a more efficient use of the spectrum. Using too much power would cover more area than is needed, and also has the potential to wreak havoc on other users of the band.

Pozar, *supra* note 244, at 3. More generally, Jon Peha writes:

> It has been shown that devices in unlicensed spectrum are likely to transmit for greater duration and at greater power than is necessary, as this will advance other design goals. This phenomenon must be addressed if spectrum is to be used efficiently. The alternative is to allocate excessive spectrum so that contention is rare.

Jon M. Peha, *Wireless Communications and Coexistence for Smart Environments*, 7 IEEE PERS. COMM., Oct. 2000, at 66, 66.

250. *See* Workshop Paper, Jon M. Peha, Emerging Technology and Spectrum Policy Reform 5 (Int'l Telecomms. Union Workshop on Market Mechanisms for Spectrum Management, Jan. 2007), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.543&rep=rep1&type=pdf ("When releasing unlicensed spectrum, regulators must guard against two related sources of inefficiency. One is that unlicensed spectrum will attract applications that would operate more effectively and efficiently in licensed spectrum. The other is that engineers will design 'greedy' devices, i.e. those that transmit with greater power, duration, or bandwidth than necessary, because they have little incentive to conserve spectrum that is shared. In the extreme, greedy devices can lead to a tragedy of the commons, where many devices are greedy, and all devices in the band experience inadequate performance as a result.") (footnote omitted).

251. Hyun Jin Kim & Jon M. Peha, *Detecting Selfish Behavior in a Cooperative Commons*, IEEE DYSPAN, 1 (2008), *available at* http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4658233.

252. Ad hoc mesh networks that create "cooperation gain" in unlicensed bands are touted as mechanisms for expanding the capacity of radios using non-exclusive spectrum

The contrast with liberal licensed bands could not be sharper. Private spectrum owners will internalize the cost of spillovers—emissions that conflict with rival wireless users—and invest heavily to reduce them. Such entities have a readily available efficiency metric to guide this calculus: undertake only those interference-reducing outlays where expected benefits exceed expected costs.

To be rude and to chew up bandwidth on such networks is to pay extra. Pricing schedules steer subscribers towards reducing costly interference, charging higher rates (or, equivalently, limited bucket minutes) for peak time calls. But network efforts to conserve spectrum go much further. Dynamic power control features in cellular phones provide one example of how private networks seek to reduce emissions, making their exclusively controlled bandwidth quieter, expanding valuable opportunities. Courteous protocols, often put forth as a way to share unlicensed spectrum,[253] are actually hardwired into mobile handsets accessing licensed spectrum. Thus, handsets are developed and programmed to be polite emitters, and manufacturers—to sell to carriers, or be certified to access their spectrum in sales directly to end users—continually press to increase performance at lower power levels, reducing demands on shared frequencies. As Charles Jackson has written, "handsets are part of the network,"[254] and carriers are careful to protect their spectrum by promoting (and often subsidizing) radios that behave in friendly fashion.

3. Smart Radios, Dumb Crowds

As noted in Section VII.A.1, *supra*, interference is determined by the total electromagnetic din created by the sum total of all the radios trying to transmit in a band, together with other incidental and inadvertent sources of

rights. But the costs of organizing such solutions are relatively high, which is why the model has yet to be embraced in any appreciable volume. Indeed, the need for network coordination is itself hampered by the lack of spectrum ownership. This is why Jon Peha recommends licensed spectrum to effect the ad hoc mesh solution: "[A] 'spectrum commons' could be created by a license-holder instead of the regulator. Rather than using unlicensed spectrum, a private entity might obtain a license, establish its own operating rules, and allow devices to operate in its spectrum. The latter approach is particularly appropriate for a cooperative system," such as ad hoc meshes. Peha, *supra* note 250, at 7 (footnote omitted).

253. LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE 184 (1999).

254. Charles L. Jackson, Wireless Handsets Are Part of the Network (Apr. 27, 2007) (filed as Attachment C to Opposition of CTIA—The Wireless Association, Skype Communications S.A.R.L: Petition to Confirm a Consumer's Right to Use Internet Communications Software and Attach Devices to Wireless Networks, RM-11361 (Fed. Commc'ns Comm'n Apr. 30, 2007)), *available at* http://files.ctia.org/pdf/filings/Skype_Opposition_Final_Attachments_04302007.pdf.

noise. In the total-din calculus, the number of radios is important, as is their power. But no one directly controls the total number of transmitters in an unlicensed band.

End users compound the problem when they hang on to obsolete transmitters long after more spectrum-efficient technologies have been developed. Countless radios that rank as "low-power" and "smart" today may still be transmitting with impunity in unlicensed bands ten years hence, occupying spectrum that could be used much more efficiently by much smarter, lower-power technology developed in the interim. Arguments for more unlicensed bandwidth assert that competition between technology suppliers will produce devices that effectively limit interference. This fails to engage the central problem posed by unlicensed wireless users whose economic interest is to keep operating their old radios no matter the spillovers they cause.

Exclusive airwave rights help address both problems. The licensee controls the total number of transmitters (base stations, handsets, M2M radios,[255] and so forth). It can (and does) hardwire cooperation into such devices, promoting spectrum-saving devices. It also orchestrates orderly transitions from old technology to new. The cellular networks of the early 1980s, for example, were analog by FCC mandate. After the FCC authorized digital upgrades, carriers effected a seamless transition, in part by giving their customers new, less polluting, subsidized phones.[256]

Wi-Fi standards, commons advocates claim, have dealt with the problem of "orphan" technology much better than broadcasters have managed to make the transition from analog to digital television.[257] This evinces the confusion discussed earlier over the distinction between traditional licenses and liberal licenses. Coordination problems plague the TV market precisely because TV broadcasters do not own their spectrum, cannot transact to rearrange it, and cannot control the radio receivers using it. TV sets are, in other words, *unlicensed devices*. These receivers are made to government specifications, and the program content they access is transmitted via government mandates.

---

255. Machine-to-machine wireless devices are a burgeoning part of the mobile network landscape. *See* Mayo & Wallsten, *supra* note 75, at 65–67.

256. The providers had an economic incentive to retain existing customers. For the rules governing the transition, see Year 2000 Biennial Regulatory Review—Amendment of Part 22 of the Commission's Rules to Modify or Eliminate Outdated Rules Affecting the Cellular Radiotelephone Service and Other Commercial Mobile Radio Services, 16 FCC Rcd. 11,169, ¶¶ 18–20 (2001).

257. WERBACH, *supra* note 17, at 23–24.

In the cellular market, with its liberal licenses, networks control their airspace and manage the subscriber interface: the leaps in technology—from the flip phones of ten years ago to the iPhones and broadband datacards of today—are large and continuous. No one seems much concerned about the "digital cellular transition" or the "EV-DO transition," despite the fact that the economic consequences for actual consumers are far more profound, remembering that over ninety percent of TV viewing takes place via cable and satellite, unaffected by the digital TV transition—completed in 2009, and officially *twenty-two* years in the making.[258]

Perhaps the most popular metaphor for the view that smart radios obviate the utility of a controlled spectrum space is the cocktail party tale. The party venue may fill up and the din of the crowd increases to a dull roar. But the human ear is adroit at focusing on just one conversation in the mix. Sometimes so adroit that one can eavesdrop on particularly juicy chatter being conducted halfway across the room.[259] The moral of the tale is that good receivers can beat a noisy roar. And science has provided just the technologies to make those good receivers.

The cocktail party metaphor popped up long ago in the classroom of Claude Shannon, author of Shannon's Law and a towering figure in modern radio frequency engineering.[260] His vision inspired his M.I.T. student, Irwin Jacobs, to implement the insight. Jacobs, with fellow scientist Andrew Viterbi, developed the "spread spectrum" technology that unraveled so many garbled sounds into intelligible conversations and then applied that technology to mass communications. The company they founded, Qualcomm, pioneered advanced wireless networks based not on unlicensed, but licensed, spectrum.[261] Hence, even the iconic spectrum-sharing technological twist is nested in a globally successful application using

258. The FCC's Advanced Television proceeding actually launched in 1987. *See* Hazlett, *Transition to Yesterday*, *supra* note 19.

259. *See On the Same Wavelength*, ECONOMIST, Aug. 14, 2004 ("A well-attended cocktail party has a din of many voices speaking at once and on similar frequencies. But it is still possible for party-goers to have conversations and pick out individual voices—ie, sound waves—from the din, because our brains are equipped with powerful software for this task. There is no limitation in the spectrum of sound waves, only in the refinement of the human ear. The same can be true in the electromagnetic spectrum.").

260. *See* DAVE MOCK, THE QUALCOMM EQUATION 70–72 (2005).

261. U.S. wireless carriers Sprint and Verizon use Qualcomm's CDMA technology. More generally, all 3G technologies are built around CDMA algorithms. "In order for a mobile phone to function on a CDMA-based network (which includes all 3G networks), handset makers must pay royalties to Qualcomm." Brian Colello, CPA, *Qualcomm, Inc.*, MORNINGSTAR, http://quicktake.morningstar.com/s/stock/analysis.aspx?t=QCOM (last visited Apr. 24, 2011).

exclusive spectrum rights which, among other things, allowed spread spectrum methods to work by limiting the number of conversations and the overall level of noise—necessary for successful communications even with the most advanced science.[262]

4. Physical Separation

When an antenna transmits equally in all directions, the power of the signal falls with the square of the distance. Obstacles—buildings, trees, and the air itself—cause additional attenuation. The earth's curvature blocks signals from transmitters situated over the horizon. But here too, physical separation involves trade-offs and judgment calls based on the cost and quality of the transmitters and receivers. The ability to squeeze more radios within a given space is at the heart of advanced communications. "Father of the Cellphone," Martin Cooper, estimates that it accounts for the lion's share of a *one-million-fold increase in spectrum capacity* every fifty years, a relationship now known as Cooper's Law.[263]

With broadcasting, the FCC's historical policy was wide physical spacing to accommodate cheap receivers.[264] This approach devoted most channels to "taboo" fillers, guard bands left idle to absorb stray interference. As technology improved, the Commission—with long lags—has required less

262. Mock explains:

> At the code-division party . . . more people are allowed to flood into any one space at the same time to hold discussions. At this party, though, each pair speaks in a different language . . . . Because you would be keying in on the nuances of your partner's spoken language, all the other languages would just sound like gibberish in the background. Even though thousands of dialects were available for couples to speak, there was still a limit on how many could be spoken at the same time—basically the limit that resulted from the overall noise. If there were fifty couples, all talking different languages, it could get too noisy for anyone to hear. So at the CDMA party it was important to regulate how loudly everyone spoke, to make sure that the maximum number of couples could be heard.

MOCK, *supra* note 260, at 71.

263. Martin Cooper, *Internet: A Life Changing Experience*, IEEE MULTIMEDIA, Apr.–June 2001, at 11, 14.

264. *See, e.g.*, FCC Notice, *900 MHz and 3 GHz*, *supra* note 61, ¶¶ 9, 10 (2002) (describing Commission rule requiring distance separations between channel-adjacent frequencies); J. Walker, *Don't Touch That Dial: Free Radio Berkeley Takes on the FCC and Official History*, REASON, Oct. 1995, at 30 (quoting San Francisco radio station KQED engineering supervisor Fred Crock as stating that the shortcomings of inexpensive receiver design may have been basis of U.S. broadcast allocations).

spacing.[265] The stultification associated with rigid administrative controls is evident. In 1952, four national commercial TV networks existed—ABC, CBS, DuMont, and NBC. This set defined the scope of America's programming choice. By 1984, that number had sunk to three. The elimination of rules blocking cable and satellite TV competition then permitted the number of national and local broadcast networks to skyrocket.[266] But for a generation, regulation put a stranglehold on competitive progress.

The administrative approach to spectrum resulted in massive under-utilization of radio,[267] television,[268] and other allocated bands.[269] What appeared to be a cheap way to assure high-quality reception, turned out to be anything but. The opportunity costs—what valuable wireless stuff could have used the taboo channels—have easily dominated any improvement in off-air viewing they provide.

With flexible-use licenses for mobile services, the Commission has abandoned "site licensing" for "geographic licensing." The new approach delegates physical-spacing decisions to wireless operators. The core engineering concept behind "cellular" phone service is that the same frequencies can be reused again and again when arrays of potentially conflicting transmitters are suitably deployed. Wireless carriers are in charge of making these architectural calls. They generally add new cells as they add new subscribers. Wireless carriers thus control the spacing of their emissions.[270] Neither the FCC nor the licensees can easily control the location of the peripatetic transmitters in cell phones, PDAs, and netbooks,

265. For example, on average, the minimum separation distance for the same channel in neighboring regions is significantly less for DTV compared to typical analog signals. *Compare* 47 C.F.R. § 73.610(b) (2010) *with id.* § 73.623(d).

266. *See, e.g.*, W. KIP VISCUSI, JOSEPH E. HARRINGTON, JR. & JOHN M. VERNON, ECONOMICS OF REGULATION AND ANTITRUST 480–81 (4th ed. 2005).

267. *See* Thomas W. Hazlett & Bruno E. Viani, *Legislators vs. Regulators: The Case of Low Power FM Radio?*, 7 BUS. & POL., no. 1, art. 1 (2005); Thomas Hazlett, *Who Killed Micro Radio?*, ZDNet (Apr. 17, 2001, 12:00 AM), http://www.zdnet.com/news/who-killed-micro-radio/115396.

268. *See* Hazlett, *U.S. Digital TV Transition*, *supra* note 19.

269. *See* Thomas W. Hazlett, *Optimal Abolition of FCC Spectrum Allocation*, 22 J. ECON. PERSP. 103, 109 tbl.2 (2008).

270. *See* 47 C.F.R. § 22.907 (2010) ("Licensees in the Cellular Radiotelephone Service must coordinate, with the appropriate parties, channel usage at each transmitter location within [75 miles] of any transmitter locations authorized to other licensees or proposed by tentative selectees or other applicants, except those with mutually exclusive applications."); *id.* § 24.134 ("A co-channel separation distance is not required for the base stations of the same licensee or when the affected parties have agreed to other co-channel separation distances.").

so carriers follow their customers, building out facilities, splitting cells to increase capacity, executing roaming agreements for seamless "out of market" service, and pricing calls to limit crowding.[271]

Recently, carriers have followed their customers all the way to their homes, offering to better extend the reach of their WANs. T-Mobile offers a device that routes in-home calls to broadband modems (connected to DSL, fiber, or cable data networks) for VoIP connections using Wi-Fi links, with calls flipping over to mobile networks when the subscriber hops in her car and drives off.[272] Alternatively, carriers supply subscribers with "femtocells," miniature base stations that extend "four bars" to the individual subscriber's home.[273] These transceivers may utilize the carrier's licensed frequencies within the home, but then route outgoing traffic via (fixed) broadband links.[274]

Advanced wireless technologies are alleged to put the role of geographical separation in network planning to an end. Radios in close proximity are now easily coordinated by smart technologies that use simple etiquettes to coordinate emissions. Mesh networks use Wi-Fi radios as links in a chain over which they hop across the Internet, creating capacity with the additional user. Instead of limiting capacity, the more the merrier: extra radios will mean more coverage, and more total wireless capacity, too.[275] Where scarcity and conflicts once reigned, abundance now flowers.[276]

---

271. There is no irony in the fact that most customer minutes are "free," in the sense of being off-peak, on-net, or within the allotted "bucket." So long as network "members" support the jointly shared facilities (including spectrum) with monthly subscription fees, the carrier enthusiastically extends access rights (which, of course, is why subscribers join). Metering peak minutes that run past the bucket limits congestion.

272. Andrew Berg, *T-Mobile Intros Wi-Fi Plan for Enterprise*, WIRELESS WEEK (May 7, 2009), http://www.wirelessweek.com/News-T-Mobile-Wi-Fi-Plan-Enterprise-050709.aspx.

273. Derek Kerton, *Pre-Brief of the Upcoming CTIA Conference*, TECHDIRT (Mar. 31, 2009, 2:59 AM), http://www.techdirt.com/articles/20090330/2030174313.shtml.

274. Glenn Fleishman, *Verizon Getting on Femtocell Bandwagon with Sprint, AT&T*, ARS TECHNICA (Jan. 19, 2009, 1:20 PM), http://arstechnica.com/telecom/news/2009/01/verizon-getting-on-femtocell-bandwagon-with-sprint-att.ars.

275. Benkler, *supra* note 17, at 45 ("[A]dding *users* with the right kind of equipment to an open wireless network can *add capacity*, not only demand.").

276. Gregory Staple & Kevin Werbach, *The End of Spectrum Scarcity*, IEEE SPECTRUM (Mar. 2004), http://www.spectrum.ieee.org/mar04/3811/ ("[E]very new device uses some of the network's capacity but also adds capacity back. Because a device in a mesh no longer needs to send information all the way to its ultimate destination (such as a cell tower), it can use less power. That allows the network to add more devices without any noticeable increase in interference.").

But wireless meshes, which have been available for over a quarter century,[277] are no free radio spectrum lunch. Costs of coordinating the mesh are substantial—so high, relative to the alternatives, that ad hoc meshes are virtually non-existent. Some agencies, including the U.S. military, do use mesh networks, but these are engineered top-down, not spontaneously, and the network providers that create such systems could purchase spectrum inputs just as mobile carriers do. Indeed, mesh networks are built using licensed spectrum.[278] The promise of meshes as disruptive innovations was that they would eliminate the network coordination function altogether.[279] That promise has not been realized.[280] ArrayComm, a wireless technology firm pioneering the development of "smart antennas," offers a general explanation:

> The final and most confused argument against concerns about spectrum availability rests on the belief that "technology" will solve the problem by enabling through cognitive radios and other concepts the peaceful cohabitation of spectrum by formerly interfering applications. While there are large tracts of allocated but currently underutilized spectrum in the mobile-device sweet spot, especially for public safety and military applications, and while it is true in principle that continuing advances in signal processing technologies will *eventually* make the collaborating-radios vision feasible, the predominant view is that the long timeline for its realization does not make this argument relevant for current business planning purposes. In the meantime, the industry must

277. Kristin Masters, *Mesh Networks*, POWERSOURCE ONLINE (Mar. 2010), http://www.powersourceonline.com/magazine/2010/03/mesh-networks; *see also, e.g.*, D. J. Baker, J. Wieselthier & A. Ephremides, *A Distributed Algorithm for Scheduling the Activation of Links in a Self-Organizing, Mobile, Radio Network*, *in* 1 IEEE 1982 INTERNATIONAL COMMUNICATIONS CONFERENCE, at 2F.6.1 (1982); N. F. Maxemchuk, *Regular Mesh Topologies in Local and Metropolitan Area Networks*, 64 AT&T TECH. J. 1659 (1985). *See generally* Tim B. Lee, *Multi-Hop Matters: The State of Wireless Mesh Networking,* ARS TECHNICA (Dec. 1, 2009), http://arstechnica.com/tech-policy/news/2009/12/mesh-networks-come-of-age.ars.

278. The 4.9 MHz band, allocated for licensed use by public safety organizations, hosts significant mesh deployments. *See* FARPOINT GRP., INTERFERENCE AND METRO-SCALE WI-FI MESH NETWORKS 4 (Jan. 2008), *available at* http://www.ict-partner.net/en/US/solutions/collateral/ns340/ns394/ns348/ns736/net_implementation_white_paper0900aecd805eb886.pdf.

279. Benkler, *supra* note 17, at 47.

280. R. P. Karrer, A. Pescapé & T. Huehn, *Challenges in Second-Generation Wireless Mesh Networks*, 2008 EURASIP J. WIRELESS COMM. & NETWORKING (Aug. 2008), *available at* http://www.hindawi.com/journals/wcn/2008/274790.html ("[T]he performance of the [mesh] networks is dismal; experience shows that the throughput is limited, and unfairness and throughput degradations of multihop communication impose severe limitations. Moreover, from an economical perspective, subscription rates to city-wide meshes, such as in San Francisco, are dismal.") (footnote omitted).

> maintain its focus on more efficient policies for licensed spectrum use . . . .[281]

Tellingly, what have been deployed are network-centric meshes, where a carrier distributes devices across an area, hardwires them to coordinate their use, and then links them to a broadband connection (over privately owned spectrum) to the Internet. This mimics the structural model of the cellular operator.

One difference, of course, is found in performance. Whereas mesh networks are designed for limited applications and have great difficulty handling mobile communications, cellular networks provide ubiquitous mobile coverage and scale to handle billions of minutes of use. The market test has been run. Were wireless meshes to render spectrum scarcity moot, the unlicensed bands would have displaced the cellular bands and the $160 billion per year mobile industry would have been displaced by mesh carriers using "free" airwaves.

B. TECHNOLOGICAL INNOVATION AND MARKET EFFICIENCY

When it liberally licenses spectrum, the FCC largely deregulates hardware. When it unlicenses spectrum, the FCC necessarily regulates hardware.

In the CMRS bands, where spectrum is licensed for flexible use, the licensee has wide discretion to decide what mix of low-power and high-power radios to deploy, selecting a mix to maximize profits. When spectrum is unlicensed, the regulator must—and does—regulate radio hardware power and standards.[282] "Low-power," "smart," and "non-interfering threshold" are not technical terms, but regulatory constructs.

Thus, however smart they may be, all "intentional radiators" are regulated under Part 15, Subpart C, of the FCC rules, the first section of which establishes an "equipment authorization requirement." This Subpart goes on to regulate antennas, power amplifiers, and bands of operation, and then it sets out detailed "radiated emission" limits, band by band.[283] The radios used in licensed bands require FCC approval, too. But Part 15 regulation of devices that operate in unlicensed bands is much more intrusive

---

281. Marketing Materials, ArrayComm, Wireless Isn't Broadband Without Us: Navigating the Harsh Realities of Broadband Wireless Network Economics 9 (2004), http://www.arraycomm.com/docs/ArrayCommonMBWAecons.pdf (footnote omitted).

282. This is true for every other developed country, as well. The "open access" free-for-all that is sometimes advanced as a regulatory alternative is simply utopian. The only country reported not to regulate unlicensed radio devices is Haiti, but little use is made of the policy. Jon M. Peha, *Lessons from Haiti's Internet Development*, 42 COMM. ACM, no. 6, June 1999, at 67, 71 (1999).

283. *See* 47 C.F.R. § 15.201 (2010).

because it strictly regulates power, one of the two most important technical characteristics (alongside operating frequency) of every radio.

There is no law of engineering that says everyone will always be better served by low-power, short-range radios. Transmitting from 23,000 miles in space and covering half the continent, direct broadcast satellites deliver competitive digital television signals to over one hundred million households, including those in rural communities that have no prospect of getting comparable video service from unlicensed wireless devices any time soon. Mobile carriers built out their networks efficiently by starting with fewer, larger, higher-power cell sites; more cell sites and lower power radios followed in step with growth in subscribers and minutes of use. Here too, the economical roll-out of service in rural areas has often depended on using taller masts and higher-power transmitters to provide large service footprints across thinly populated areas. Where people are sparse, simpler, cheaper radios that "waste" spectrum are more efficient, because there is spectrum to spare.[284]

These are exactly the tradeoffs that liberal license holders routinely discover and efficiently exploit. The FCC's settled policy is now to give licensees "flexibility to determine the types of services and the technologies and technical implementation designs used to provide those services."[285] Technology choices have thus emerged as a key dimension of competition among service providers.

Because about half of all transmissions originate in the hands of end users,[286] wireless carriers have also invested heavily to get high-performance, feature-rich, spectrum-efficient wireless devices into their customers' hands. Carriers paid about three-quarters of the approximately $14 billion spent on mobile phones in the United States in 2003.[287] Such subsidies reflect carriers' propertied interest in their spectrum. Newer phones embed more advanced technologies, reducing spectrum spillovers (interference) and offering greater functionality for users. Both are of value to the network over and above the gains delivered to the individual adopter. Hence, handset subsidies are higher for newer models and have been particularly important in spreading 3G

284. Ellen P. Goodman, *Spectrum Rights in the Telecosm To Come*, 41 SAN DIEGO L. REV. 269, 308–09 (2004).

285. *Interference Temperature Metric*, 18 FCC Rcd. 25,309, ¶ 6 (2003).

286. *See* Jackson, *supra* note 254.

287. Mike Dano, *Phone Subsidies Alive and Well*, RCR WIRELESS NEWS, Jan. 5, 2004, at 1.

technology.[288] Many smaller networks have more aggressively supplied subsidies, underscoring the competitive role of such vertical promotions.[289]

Unlicensed bands accommodate non-exclusive use rights but require highly exclusionary public policies. The technologies authorized in unlicensed bands effectively prohibit most wireless options, on the expectation that they would cause harmful interference. When the FCC unlicenses spectrum, carriers and consumers must choose Intel's Centrino chips over Qualcomm's CDMA chips and Wi-Fi access points over data networks provided by GSM UMTS/HSDPA, CDMA 1xEV-DV, or WiMax optimized for licensed radio spectrum.[290]

The fact that the approved low-power technology choices sometimes result in widespread adoption is evidence that not all social value is eliminated by the allocation policy. It does not, however, prove that the resulting wireless activity is the optimum. A government-managed band does not contain the requisite feedback mechanisms to reveal and then adjust to the most efficient spectrum sharing arrangements. It is always possible that an alternative set of rules would generate greater gains. But whereas owners of licensed spectrum have profit incentives and the financial ability (i.e., access to capital markets) to arrange positive-sum transactions to make such transitions as present themselves as good candidates for superior results, spectrum regulators do not.

Some unlicensed bands appear to have performed well, like the 900 MHz and 2.4 GHz bands, while others, like U-PCS and the 3650 MHz band, appear to constitute allocation failures. But regulators have no reliable way to optimize any particular band because the relevant counter-factual obviously cannot be observed. More basically, even after the fact—when the FCC observes "tremendous success" or some other categorically salubrious outcome—it is under no pressure to perform the relevant social welfare calculation: was the FCC's allocation, and the associated rules (including adoption of the unlicensed access model and the restrictions on radio usage

288. *See* Marko Repo, *Regulation of Wireless Stakeholders* 4 (Seminar on Networking Business, Helsinki Univ. of Tech., Paper, Oct. 2006), http://www.netlab.tkk.fi/opetus/s383042/2006/papers_pdf/B2.pdf; Ville Saarikoski, The Odyssey of the Mobile Internet 7 (2006) (unpublished Ph.D. thesis, University of Oulo, Finland).

289. *See* Pedro Pita Barros, Handset Subsidies—An Empirical Investigation 3 (Aug. 14, 2008) (unpublished manuscript), *available at* http://www.anacom.pt/streaming/est22112006.pdf?categoryId=218502&contentId=424227&field=ATTACHED_FILE.

290. This set of technologies appears to include WiMax, by many accounts the most advanced technology thus far emerging from the Wi-Fi family. *Wi-Fi? How About Way Far? WiMax Delivers High-Speed Wireless Internet Service as Far as 35 Miles Away*, HARTFORD COURANT, Mar. 25, 2004, at D3.

attendant to that approach), the most efficient way to use this particular bandwidth? Could markets have supplied more creative, lower-cost ways to accommodate the services obtained, while stimulating additional wireless services of value to consumers? That the FCC does not have the institutional ability to make such judgments with the reliability of alternative institutions, notably profit-maximizing capital owners in competitive markets, is obvious. Perhaps most striking is that the FCC does not even, as a pro forma matter, deem it necessary to *ask the question*.

Government employees are constrained to be disinterested. They do not generate information as to profitable opportunities for spectrum re-use but, rather, maximize utility under the political realities governing their agency. Whereas entrepreneurs scour the marketplace looking for assets that are undervalued, searching for the financial means to divert such resources to more productive employments, the bureaucratic goal is to preserve an existing governing equilibrium.

It is unknowable within an administrative allocation system whether a given unlicensed block—even where it is certifiably creating more benefit than cost in aggregate—is of the right size. Perhaps the 2.4 GHz ISM band would optimally be half, or twice, its present scope. Without ownership, there are no transactions, and without transactions, no market values. This leaves regulators guessing about where to draw lines.

Suppose that a firm like Apple, instead of arguing for the government to allocate more bandwidth to license-exempt use, were to buy liberal licenses and then provide a "spectrum commons" on its own.[291] It could then set frequency sharing rules, including technology formats and power limits, becoming a private FCC. In fact, this is the vertical structure of the mobile carrier space that already exists, the twist being a business model that rides by contract on the carrier's network infrastructure and yet "opens" the licensed airwaves to radio devices and applications provided by hundreds of third party providers.

Under this approach, competitive market forces regulate the performance of Apple in providing the "commons." Just as carriers are rewarded for providing customers with superior user experiences, the "unlicensed" space would become profitable to the degree that sharing rules were properly balanced. Not only is such a model already well established, if not ubiquitous,

291. This option has been suggested by many analysts, including professional staff at the FCC. *See* Kwerel & Williams, *supra* note 8, at 31.

in the licensing of intellectual property rights,[292] it mirrors the system in place for mobile carriers. There, device makers contract to gain access to carriers' networks and spectrum assets, passing the rights to customers who "play" their radios right out of the box. Mobile handsets, including those like TracFone (a virtual mobile operator that enables wireless phone calls using access rights purchased in wholesale markets) or Kindle (a book reader, sold by Amazon, that downloads content via Amazon's contract with Sprint) embed seamless spectrum access rights for end users.

The advantages of such an approach over administrative allocation are manifest. But such an outcome is displaced by mandated allocations to unlicensed allocations. Hence, public policy drove Apple to request that the government place spectrum resources at risk, inviting its lobbyists to steer allocations rules in the direction sought by the firm—no payment required. Whatever Apple has gained from unlicensed band set asides, the regulatory process has muddled transparency, undermined productive efficiencies, and blocked information-rich feedback loops. It should be noted, of course, that Apple has since emerged as a powerhouse in the wireless sector by abandoning its quest for unlicensed allocations in favor of contracts with carriers owning exclusive spectrum rights.[293]

VIII. CONCLUSION

Administrative allocation of radio spectrum has historically been plagued by overly conservative policies that unduly limited competition and blocked technological innovation. But as broadcasting was eclipsed by mobile telephony, a subtle shift in policy took hold that in hindsight amounted to a policy revolution. In the modern marketplace, liberal spectrum licenses host extremely complex economic structures, allowing millions of customers to buy wireless services using advanced technologies, while facing a continually increasing number of access options and applications. This regime generates intense sharing of spectrum by rival networks and by mass-market

292. This analogy is nicely developed in WILLIAM J. BAUMOL & DOROTHY ROBYN, TOWARD AN EVOLUTIONARY REGIME FOR SPECTRUM GOVERNANCE: LICENSING OR UNRESTRICTED ENTRY? (2006).

293. Apple's lobbying for U-PCS in the 1990s was driven by a strategy to secure wireless access for its early PDA, the Newton. That the Newton flopped may or may not be associated with the limitations of the regulatory process generally, or the U-PCS allocation specifically. That the iPhone's success following its 2007 launch is associated with liberal licenses and the wide area networks they enable is, however, crystal clear. No other spectrum regime could provide the level of social organization necessary to accommodate the access services embedded in the iPhone.

subscribers to voice and data services. In delegating choices to competitive firms, de facto airwave ownership rationalizes spectrum use.

The suggestion that exclusive rights to radio spectrum are made obsolete by advancing technology has the basic economic coordination problem backwards. The advanced wireless devices are today superior, in their ability to send and receive data, to previous generations. But they cannot be fruitfully deployed without some form of social control over the airwaves they access. The alternative to competitive ownership is the imposition of behavioral constraints by regulators. Power limits and technology restrictions are inevitably applied to license-exempt spectrum to limit conflicts. Smart radios do not portend the "end of scarcity" but constitute yet another ascending pathway on the mountain wireless entrepreneurs have been climbing since Marconi blazed a trail for wireless innovations in 1895.

That seminal technological breakthrough triggered a chain of events that created spectrum scarcity. Contrary to the view that advanced devices solve such mundane matters of economic organization and obviate the value of exclusive spectrum rights, each new and improved radio actually triggers more demand for airwave access, *increasing* potential spectrum conflicts. Intensifying calls for more unlicensed spectrum in more desirable bands reflects just this scarcity, with rival claims made in the political marketplace. The pro-consumer policy unleashing the social value of wireless would shift such competitive bidding from the political marketplace to the economic realm.

This approach carries great promise and little risk. There is nothing that "spectrum inventories" held by government can achieve. The most reliable way to destroy valuable spectrum, in fact, is not to use it. The second most reliable way is to distribute a massive number of tiny, overlapping rights that cannot be usefully re-aggregated—the Humpty Dumpty approach that can easily result from unlicensed allocations.[294] By allocating liberal licenses to competitors, however, market forces will divert spectrum resources to where demands are highest. If regulators in the future ascertain that important demands are unmet, it will be free to acquire frequencies for the task and to know the market price of doing so.

The 700 MHz licenses auctioned in March 2008 raised $19 billion for the U.S. Treasury. Those bids reflect future anticipated profits (now transferred to the government) available to firms that control the resource rights

294. On tragedy of the anti-commons generally, see HELLER, *supra* note 138. On such tragedy with respect to unlicensed use of television band "white spaces," see Thomas W. Hazlett, *Tragedy T.V.: Rights Fragmentation and the Junk Band Problem*, 53 ARIZ. L. REV. 83 (2011).

conveyed. Nothing prevented bidders from purchasing such bandwidth and deploying it as a "spectrum commons"; indeed, that approach would seem an attractive alternative to spending the many billions of dollars on network infrastructure required in executing the mobile carriers' network-centric model—were customers willing to pay for the services therein offered.

Marketplace rejection of the "commons" model does not constitute market failure, but a competitive equilibrium. It reveals that there are superior choices, given the available technologies and the associated consumer demands, for using valuable radio frequencies. That is the rationality supplied by liberal spectrum licenses.

COMBATING CYBER-VICTIMIZATION

Jacqueline D. Lipton[†]

TABLE OF CONTENTS

---

† Professor of Law and Associate Dean for Faculty Development and Research; Co-Director, Center for Law, Technology and the Arts; Associate Director, Frederick K. Cox International Law Center, Case Western Reserve University School of Law, 11075 East Boulevard, Cleveland, OH, 44106, Email: JDL14@case.edu, Telephone: (216) 368-3303. The author would like to thank Professor Lyrissa Barnett Lidsky, Professor Elizabeth Rowe, and Professor Ann Bartow for comments on an earlier draft of this Article. Additionally, the author is extremely grateful for comments from participants at the 3rd Annual Privacy Law Scholars' Conference at The George Washington University Law School, Washington, D.C., June 3, 2010, including comments from Mr. David Thompson, Professor Mary Fan, Professor Bruce Boyden, Professor Danielle Keats Citron, Mr. Ryan Calo, Professor Jon Mills, Mr. Avner Levin, Mr. Doug Curling, Ms. Eileen Ridley, Mr. Stefaan Verhulst, and Professor Joel Reidenberg. In particular, Professor Raphael Cohen-Almagor was extremely generous with his time and comments. All mistakes and omissions are, of course, my own.

I. INTRODUCTION

> "Once, reputation was hard-earned and carefully guarded. Today, your reputation can be created or destroyed in just a few clicks."[1]

Words can hurt. Whether true or false, whether spoken by friend or frenemy,[2] the cyber-pen is mightier than the sword.[3] In today's networked society, abusive online conduct such as cyber-bullying and cyber-harassment can cause serious damage, including severe emotional distress,[4] loss of

1. MICHAEL FERTIK & DAVID THOMPSON, WILD WEST 2.0: HOW TO PROTECT AND RESTORE YOUR ONLINE REPUTATION ON THE UNTAMED SOCIAL FRONTIER 2 (2010).

2. "Frenemy" has been defined as "a person who pretends to be a friend but is actually an enemy; a rival with which one maintains friendly relations." *Frenemy Definition*, DICTIONARY.COM, http://dictionary.reference.com/browse/frenemy (last visited June 6, 2010).

3. *See* Raphael Cohen-Almagor, *Responsibility of Net Users*, *in* 2 THE HANDBOOK OF GLOBAL COMMUNICATION AND MEDIA ETHICS 415, 419 (Mark Fackler & Robert S. Fortner eds., 2011) ("Words can wound. Words can hurt. Words can move people to action.").

4. *See, e.g.*, Jacqueline Lipton, *"We, the Paparazzi": Developing a Privacy Paradigm for Digital Video*, 95 IOWA L. REV. 919, 921–22 (2010) (discussing the "Star Wars Kid" incident in which a Canadian teenager filmed himself mimicking the use of a light saber, which prompted the creation of humiliating mash-up videos that resulted in him dropping out of school and requiring psychiatric care).

employment,[5] and even physical violence[6] or death.[7] Thirteen-year-old Megan Meier, who believed she had found a soul mate in the fictional "Josh Evans" on MySpace, was driven to suicide by his spurning words.[8] This is but one of an increasing number of examples of abusive online conduct.[9] Almost one in four teenagers reportedly experiences cyber-bullying,[10] and approximately sixty-five percent of children know someone who has been the victim of cyber-bullying.[11] Furthermore, a 2006 Pew Internet study found that one-third of online teenagers had been victims of online harassment and that thirty-nine percent of social network users have been cyber-bullied.[12] Online abuses—cyber-bullying, cyber-stalking, and cyber-harassment—

5. Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 64 (2009) [hereinafter Citron, *Cyber Civil Rights*] ("Victims who stop blogging or writing under their own names lose the chance to build robust online reputations that could generate online and offline career opportunities.").

6. *See, e.g.*, Danielle Keats Citron, *Law's Expressive Value in Combating Cyber Gender Harassment*, 108 MICH. L. REV. 373, 396–97 (2009) [hereinafter Citron, *Harassment*] ("The online abuse inflicts significant economic, emotional, and physical harm on women in much the same way that workplace sexual harassment does."); *see also* Kara Carnley-Murrhee, *Cyberbullying: Hot Air or Harmful Speech? Legislation Grapples with Preventing Cyberbullying Without Squelching Students' Free Speech*, U. FLA. L. MAG., Winter 2010, at 17, 18 (describing the case of thirteen-year-old Hope Witsell, who committed suicide after being the victim of a "sexting" campaign—a variation of cyber-bullying in which sexually explicit images of the victim or sexually explicit messages about the victim are disseminated over digital communications services); *Cyber Bullies Target Girl*, BBC NEWS (May 24, 2003), http://news.bbc.co.uk/2/hi/uk_news/england/nottinghamshire/2933894.stm ("[The victim's] family says there has been a two-year campaign of intimidation and she has twice been attacked in school.").

7. Danielle Keats Citron, *Mainstreaming Privacy Torts*, 99 CALIF. L. REV. 1805, 1817 (2011) [hereinafter Citron, *Mainstreaming*] ("Today, the physical harm associated with information disclosures can become as serious as murder.").

8. Gordon Tokumatsu & Jonathan Lloyd, *MySpace Case: "You're the Kind of Boy a Girl Would Kill Herself Over,"* NBC LOS ANGELES (Jan. 26, 2009), http://www.nbclosangeles.com/news/local-beat/Woman-Testifies-About-Final-Message-Sent-to-Teen.html (describing the last electronic message sent by Megan Meier, the teenage victim of an infamous online cyber-bullying incident, before she committed suicide by hanging herself in her closet). *See also* Cohen-Almagor, *supra* note 3, at 421–24 (discussing the Meier incident); Lyrissa Barnett Lidsky, *Anonymity in Cyberspace: What Can We Learn from John Doe?*, 50 B.C. L. REV. 1373, 1386 (2009) (describing the Megan Meier incident and the legal responses to it).

9. For more examples of cyber-bullying conduct involving school-age children, see Drew Jackson, *Examples of Cyberbullying*, http://www.slais.ubc.ca/courses/libr500/04-05-wt2/www/D_Jackson/examples.htm (last updated Apr. 18, 2005); *see also* Citron, *Mainstreaming*, *supra* note 7, at 10–11 (giving examples of high-profile cases of online abuses).

10. Cohen-Almagor, *supra* note 3, at 423.

11. *Id.* at 22–23.

12. *Id.* at 23.

disproportionately affect "traditionally subordinated groups,"[13] which include women,[14] children,[15] and minorities.[16] The prevalence of this conduct suggests that more effective means are necessary to redress online wrongs and to protect victims' reputations, but action against cyber-abusers has posed significant challenges for the legal system. Because of the global and largely anonymous nature of the Internet, reliance on the law tends to be time-consuming and expensive for victims. In the United States, many potential legal solutions will also face First Amendment hurdles.[17]

Unlike previous writing in this area, this Article considers legal solutions within a broader context, including alternative approaches to regulating online conduct such as public education and the more effective use of commercial reputation protection services. The Article makes specific suggestions for reform of tort and criminal laws, but more importantly it places the legal debate into a larger multi-modal framework aimed at protecting online reputations. This new framework combines specific legal reforms with extra-legal regulatory approaches, many of which will prove more affordable and effective for victims of online wrongs. The principal issue addressed in this Article is how best to enable victims to combat harms and protect their own reputations. Part II explores the categories of abusive online conduct that require regulatory attention—cyber-bullying, cyber-harassment, and cyber-stalking—and contrasts these categories with their offline counterparts. Part III identifies gaps in the current law as applied to

13. Citron, *Cyber Civil Rights*, *supra* note 5, at 65–66 (citing statistics from 2006, evidencing that cyber-harassment is concentrated on women and also to some extent "people of color, religious minorities, gays, and lesbians").

14. Ann Bartow, *Internet Defamation as Profit Center: The Monetization of Online Harassment*, 32 HARV. J.L. & GENDER 383, 392 (2009) (citing Ellen Nakashima, *Sexual Threats Stifle Some Female Bloggers*, WASH. POST, Apr. 30, 2007, at A1) (explaining that "[a]s women gain visibility in the blogosphere, they are the targets of sexual harassment and threats" and tend to be "singled out in more starkly sexually threatening terms" than men, which is parallel to the treatment of women in chat rooms beginning in the early 1990s); *id.* at 394 ("Self-identifying as a woman online can substantially increase the risk of Internet harassment."); *see* Citron, *Harassment*, *supra* note 6, at 378 ("While cyber attackers target men, more often their victims are female.") (footnote omitted). But note that some victims of online harassment are men, such as male doctors. Citron, *Mainstreaming*, supra note 7, at 1817 (describing physical assaults and murders of doctors who perform abortions, where an online list of these doctors was involved in identifying them).

15. Citron, *Harassment*, *supra* note 6, at 398 (noting that younger individuals are particularly impacted by online abuses because their lives are "inextricably tied to the net").

16. Citron, *Cyber Civil Rights*, supra note 5, at 65–66.

17. For example, a number of laws directed at online speech have run afoul of the First Amendment. *See, e.g.*, Ashcroft v. ACLU, 542 U.S. 656 (2004) (invalidating content-regulating sections of the Child Online Protection Act); Reno v. ACLU, 521 U.S. 844 (1997) (invalidating content-regulating sections of the Communications Decency Act of 1996).

abusive online conduct, focusing on remedies found in criminal law, tort law, and, to a lesser extent, civil rights law. As part of this examination, it suggests ways in which current laws could be updated to more effectively combat online wrongs.

Part IV proposes extra-legal regulatory mechanisms that might better protect individual reputations online by surveying currently available options, such as commercial reputation management services,[18] and discussing their shortcomings. This Part advocates developing educational programs to empower victims of online abuses to utilize currently available legal and technological means for protecting their online reputations. Furthermore, it also suggests an increased role for reporting hotlines, evolving social norms, and industry self-regulation through codes of conduct and "naming and shaming" programs.

Part V concludes this Article by suggesting future directions in the regulation of online abuses, with a focus on extra-legal solutions. The advantages of developing these extra-legal approaches relate to easing the time and cost burdens on victims and avoiding some of the First Amendment concerns raised by legislated solutions. Additionally, development of these extra-legal avenues will ultimately change the climate of online discourse and facilitate a more civil and accountable global online society where internet service providers[19] play a more active role in monitoring and enforcing norms of accountability.

II. CATEGORIZING ABUSIVE ONLINE CONDUCT

A. DELINEATING THE BOUNDARIES OF ONLINE ABUSES

> "The Internet has turned reputation on its head. What was once private is now public. What was once local is now global. What was once fleeting is now permanent. And what was once trustworthy is now unreliable."[20]

Recent literature describes online abuse predominantly in terms of cyber-stalking, cyber-harassment, and cyber-bullying. However, none of these

18. *See, e.g.*, REPUTATION.COM, http://www.reputationdefender.com/ (last visited Apr. 14, 2010); REPUTATION HAWK, http://www.reputationhawk.com (last visited June 6, 2010); UDILIGENCE, http://www.udiligence.com (last visited May 20, 2010) (service for student-athletes); YOUDILIGENCE, http://www.youdiligence.com (last visited May 20, 2010) (service for children).

19. Internet service providers include social networking sites such as Facebook and MySpace as well as other online services that enable people to connect with each other via digital devices.

20. FERTIK & THOMPSON, *supra* note 1, at 44.

terms has achieved a universally accepted definition, and there are significant areas of overlap between them. Some authors have coined umbrella terms such as cyber-victimization[21] and cyber-targeting[22] to encompass all of these categories of conduct. These commentators have avoided individual terms for different cyber-wrongs on the basis that overlaps between the classes of wrongs might "thwart clear analysis and the creation of successful solutions."[23] There is some merit to the view that an umbrella term—such as online abuses, cyber-abuses or cyber-wrongs—is more effective than categorizing individual sub-classes of conduct, although there will be some circumstances in which the individual classifications are important.[24]

Nevertheless, a brief consideration of the kinds of conduct described in recent years as cyber-bullying, cyber-harassment, and cyber-stalking is a useful background to understanding cyber-victimization as a whole. These terms are derived from their offline counterparts—bullying, harassment, and stalking. As much current law, particularly state criminal law, is focused specifically on bullying, harassment, and stalking, it is also necessary to understand the terms in order to appreciate the gaps in the current legal system that become apparent when pursuing a cyber-victimization case.

1. Cyber-bullying

> "Bullying is an attempt to raise oneself up by directly demeaning others; the attacker hopes to improve his social status or self-esteem by putting others down."[25]

The term cyber-bullying typically refers to online abuses involving juveniles or students.[26] While it is possible that in any given instance of cyber-

21. Kate E. Schwartz, *Criminal Liability for Internet Culprits: The Need for Updated State Laws Covering the Full Spectrum of Cyber Victimization*, 87 WASH. U. L. REV. 407 (2009).

22. David A. Myers, *Defamation and the Quiescent Anarchy of the Internet: A Case Study of Cyber Targeting*, 110 PENN ST. L. REV. 667, 667–68 (2006).

23. Schwartz, *supra* note 21, at 409.

24. For example, cyber-harassment laws usually require a credible threat of immediate physical harm to a victim and thus are less likely to be successfully challenged under the First Amendment, as threats are generally not protected speech. *See* Planned Parenthood of the Columbia/Willamette, Inc. v. Am. Coal. of Life Activists, 23 F. Supp. 2d 1182, 1188–89 (D. Or. 1998) (holding that threatening speech is not protected by the First Amendment); *see also* Cohen-Almagor, *supra* note 3, at 416.

25. FERTIK & THOMPSON, *supra* note 1, at 105.

26. Schwartz, *supra* note 21, at 410–11 (explaining that cyber-bulling, which has been described as the online counterpart to traditional playground bullying, usually refers both the victim and the victimizer as being minors or students, but can also encompass situations in which the culprit is a juvenile and the victim is an adult).

bullying at least one of the parties may not be a youth,[27] discussions about cyber-bullying generally revolve around school-age children and often call on schools to address the issue.[28] The term bullying in the physical world has tended to describe conduct that occurs "when someone takes repeated action in order to control another person."[29] It can involve tormenting, threatening, harassing, humiliating, embarrassing, or otherwise targeting a victim.[30]

In recent years, the term has also been increasingly used in the employment context to describe hostile or threatening conduct in the workplace.[31] In this context, bullying is differentiated from other offensive conduct, such as harassment, on the basis that bullying tends to be targeted at a particular person for reasons other than the person's gender or race, which is the typical focus of harassment laws.[32] Targets of workplace bullying are often perceived as a threat to the bully in some way.[33] This notion of

27. *See, e.g.*, Tokumatsu & Lloyd, *supra* note 8 (reporting that the bully was the mother of a school mate of thirteen-year-old victim of cyber-bullying).

28. *See, e.g.*, CAL. EDUC. CODE § 32261(d) (2009) (setting forth the legislative intent of fostering "interagency strategies, in-service training programs, and activities that will improve school attendance and reduce school crime[,] violence, . . . [and] bullying, *including bullying committed personally or by means of an electronic act*") (emphasis added); Citron, *Harassment*, supra note 6, at 410 ("[P]arents and educators have an important responsibility to teach the young about cyber harassment's harms because the longer we trivialize cyber gender harassment, the more difficult it will become to eradicate."); Andrew M. Henderson, *High-Tech Words Do Hurt: A Modern Makeover Expands Missouri's Harassment Law To Include Electronic Communications*, 74 MO. L. REV. 379, 381 (2009) (explaining that cyber-bullying typically involves "a child, preteen or teen [being] tormented, threatened, harassed, humiliated, embarrassed or otherwise targeted by another child, preteen or teen using the Internet, interactive and digital technologies, or mobile phones" but that adults can also be involved).

29. Henderson, *supra* note 28, at 381.

30. *See id.* (explaining that these types of behavior occur in the cyber-bullying context, but equally occur in face-to-face bullying).

31. *See, e.g.*, *Bullies in the Office: Bullying Worse Than Sexual Harassment*, ABC NEWS, abcnews.go.com/index/playerindex?id=4527601 (last visited May 18, 2010); *Bullying: What Is It?*, BULLY ONLINE, http://www.bullyonline.org/workbully/bully.htm#Why (last visited May 20, 2010) ("Bullying is persistent unwelcome behaviour, mostly using unwarranted or invalid criticism, nit-picking, fault-finding, also exclusion, isolation, being singled out and treated differently, being shouted at, humiliated, excessive monitoring, having verbal and written warnings imposed, and much more.").

32. *See* FED. COMMC'NS COMM'N (FCC), UNDERSTANDING WORKPLACE HARASSMENT, http://www.fcc.gov/owd/understanding-harassment.html (last updated Jan. 8, 2008) (noting that harassment occurs in cases of "unwelcome verbal or physical conduct based on race, color, religion, sex (whether or not of a sexual nature and including same-gender harassment and gender identity harassment), national origin, age (40 and over), disability (mental or physical), sexual orientation . . .").

33. *Bullying: What Is It?*, *supra* note 31, at 7 ("Jealousy (of relationships and perceived exclusion therefrom) and envy (of talents, abilities, circumstances or possessions) are strong motivators of bullying.").

bullying would cover the Megan Meier scenario where Lori Drew—the perpetrator of the "Josh Evans" scam—perceived Meier as a potential threat to her own daughter, one of Meier's classmates.[34] Drew targeted Meier because she was concerned that Meier was saying, or would say, unpleasant things about Drew's daughter online, rather than because of Meier's gender or race. Drew took on the false digital identity of a fictional young man "Josh Evans" in order to see what Meier might say about her classmates to a digital "friend."

2. Cyber-harassment

Like harassment in the physical world, cyber-harassment should technically be limited to targeting people by virtue of their membership in a protected class such as race or gender.[35] In cyberspace, as in the offline world, the distinctions between bullying and harassment tend to blur. Much conduct that has been described as cyber-harassment involves mobbing behavior aimed at silencing women and racial minorities,[36] which seems to cross the line between bullying and harassment. While it is directed at a protected class, mobbing is typical of bullying[37] and the aim of driving subjugated groups offline seems more about control than possession—typical characteristics of bullying as opposed to harassment.[38]

Because of the overlaps between bullying and harassment and the fine distinctions between them, it may be appropriate, at least in the early days of online regulation, to address cyber-harms more universally and to worry about the distinctions later. In fact, new distinctions between classes of conduct may emerge that are more appropriate in the digital age than some of the existing distinctions. For example, regulators may choose to distinguish between communications specifically directed to an individual and general communications about an individual on the basis that the former conduct may be more immediately threatening or frightening to the victim. If direct communications contain threats, such conduct may be easier to regulate through legislation than general online communications directed to

34. *See* Cohen-Almagor, *supra* note 3, at 422 (noting that Lori Drew had suggested talking to Megan Meier via the Internet to find out what Meier was saying online about Drew's daughter).

35. *See* FCC, *supra* note 32, at 1 (discussing workplace harassment).

36. Citron, *Cyber Civil Rights*, *supra* note 5, at 4 (discussing "the growth of anonymous online mobs that attack women, people of color, religious minorities, gays, and lesbians").

37. *See Bullying: What Is It?*, *supra* note 31 (describing "gang" or "group" bullying, also known as "mobbing").

38. Presentation, Erica Merritt, Workplace Bullying (Case Western Reserve Univ., May 18, 2010) (session notes and PowerPoint slides on file with author).

an audience at large. Where an immediate threat of harm is involved, speech is less likely to be protected by the First Amendment than general speech directed to the world at large.[39]

3. Cyber-stalking

Cyber-stalking involves conduct directed at a victim, rather than general communications about a victim. At least in some jurisdictions, cyber-stalking legislation requires a credible threat to the victim for there to be a violation of law.[40] Because some commentators have described cyber-stalking as a direct online analog to the offline crime of stalking, cyber-stalking may thus be defined as "the use of the Internet, e-mail, or other means of electronic communication to stalk or harass another individual."[41] Similarly, stalking has typically been defined as involving "repeated harassing or threatening behavior."[42] The goal of the traditional stalker is to exert control over a victim by instilling fear into her.[43] In the physical world, as in cyberspace, stalking can lead to actual physical harm.[44]

While cyber-bullying and cyber-harassment may damage an individual's reputation or livelihood, cyber-stalking is more likely to result in severe and immediate emotional or physical harm. Thus, legislation aimed at redressing cyber-stalking may be able to stand up to First Amendment scrutiny more easily than legislation aimed at other kinds of online abuses.[45] While the First Amendment may protect the ability to say something unpleasant about another online—subject to defamation and privacy law—it is much less likely to protect the ability to send threatening e-mails.

---

39. For a fuller discussion of Congress's attempts to pass regulations that do not violate the First Amendment, see *infra* Part III.

40. Schwartz, *supra* note 21, at 411 ("[O]ne commentator states that cyberstalking is distinct from cyberbullying because cyberstalking involves credible threats.").

41. Naomi Harlin Goodno, *Cyberstalking, a New Crime: Evaluating the Effectiveness of Current State and Federal Laws*, 72 MO. L. REV. 125, 126 (2007); *see also* Shonah Jefferson & Richard Shafritz, *A Survey of Cyberstalking Legislation*, 32 UWLA L. REV. 323, 323 (2001).

42. *See* Goodno, *supra* note 41, at 128.

43. *Id.* at 127.

44. *See id.* at 128 ("[C]yberstalking involves repeated harassing or threatening behavior, which is often a prelude to more serious behavior."); Citron, *Mainstreaming*, *supra* note 7, at 1817 (describing a case in which online stalking led to the murder of the victim by the stalker).

45. *See* Myers, *supra* note 22, at 675.

B. COMPARING ONLINE AND OFFLINE ABUSES

> "[T]hanks to the power of the Internet, attackers and gossipmongers enjoy instant global audiences and powerful anonymity."[46]

Laws targeted at real world activities often do not translate well when applied to cyberspace. Despite facial similarities between physical abuses and cyber-abuses,[47] there are significant underlying differences. Cyber-attackers can utilize the Internet to harass their victims on a scale never before possible because of both the immediate effect of their conduct, and the speed and ease of the global dissemination of online information.[48] This immediate dissemination is inexpensive for the abuser and is not particularly time-consuming.[49] Online postings have a *constant* effect on the victim, as opposed to more transient conduct in the physical world.[50] Even where information about a victim is removed from one website, it may be cached and copied on other websites.[51] Online communications therefore have a permanent quality that real world conduct lacks.[52] Compounding the permanence effect is the fact that online information is easily searchable through Google and other popular search engines.[53] Thus, damaging information is more readily accessible to those who may be looking for it. It is also extremely difficult to redress problems relating to the availability of the damaging information. Attempts to publish corrective information may suffer from being uninteresting to many readers and thus may be de-

46. FERTIK & THOMPSON, *supra* note 1, at 6.

47. Goodno, *supra* note 41, at 128 (discussing the similarities of cyber-stalking with off-line stalking as being "a desire to exert control over the victim; and . . . repeated harassing or threatening behavior, which is often a prelude to more serious behavior").

48. *Id.* at 128–29.

49. *Id.* at 129.

50. *Id.*; Citron, *Mainstreaming*, supra note 7, at 1813 ("While public disclosures of the past were eventually forgotten, memory decay has largely disappeared. Because search engines reproduce information cached online, people cannot depend upon time's passage to alleviate reputational and emotional damage.").

51. FERTIK & THOMPSON, *supra* note 1, at 54–55 (discussing the impact of the Internet Archive on the permanent quality of online information); Citron, *Mainstreaming*, *supra* note 7, at 1813; Lipton, *supra* note 4, at 977 ("[W]ith projects such as the Internet Archive, many images will continue to be available in some form even after all 'live' images have been removed from relevant websites.").

52. Citron, *Mainstreaming*, *supra* note 7, at 1813.

53. FERTIK & THOMPSON, *supra* note 1, at 53–54 (illuminating the fact that conversations and notes amoung friends were formerly more private and lacked permanence, "[b]ut many of those same conversations are now conducted online in a blog or chat room, in full view of the world, automatically indexed by Google, and broadcast to an audience of millions").

prioritized in search results when search engines are designed to focus on popularity of information. Blogs also tend to list comments in order of posting, thus making a rebuttal comment by a victim difficult to find as compared with the original damaging posting.

A cyber-attacker can also be physically removed from the victim. He may be across the state, across the country, or even across the globe.[54] The unlimited reach of the Internet differentiates online abuse from its offline counterparts in three important respects. First, online abusers can initiate and pursue their wrongful act inexpensively and easily from anywhere in the world.[55] Second, there is a sinister element in the secrecy of the attacker's location—the victim is constantly left wondering whether the attacker is in the next house or at some far away location.[56] Finally, the global reach of the Internet leads to jurisdictional problems in enforcing laws against wrongdoers both in terms of law enforcement and in terms of gathering evidence.[57]

One might argue that online abuses are actually less serious than their offline analogs because the victim has the option of simply turning off the computer and walking away. However, in today's interconnected world that is not a viable option, as people who are forced offline forgo important personal and professional opportunities.[58] Also, if a victim moves offline, this does not stop others from posting harmful things about her that may continue to harm her personal and professional development despite her own choice not to read the postings. In many ways, it is better for a victim to know what is being said about her so she can take steps to combat the abuses.

The anonymity of online abusers also distinguishes them from their offline counterparts. While one might assume that online conduct is less harmful than the offline equivalent because it does not involve immediate

54. FERTIK & THOMPSON, *supra* note 1, at 61–62 ("Online, it is often impossible to know if the person you're chatting with is half a block or half a world away. The owner of a website might be your neighbor, or it might be someone in Azerbaijan."); Goodno, *supra* note 41, at 129 ("Cyberstalkers can be physically far removed from their victim.").

55. Goodno, *supra* note 41, at 129.

56. *Id.*

57. *Id.* at 129–30; *see infra* Section II.B.

58. MARY MADDEN & AARON SMITH, PEW RESEARCH CTR., REPUTATION MANAGEMENT AND SOCIAL MEDIA: HOW PEOPLE MONITOR THEIR IDENTITY AND SEARCH FOR OTHERS ONLINE 3 (2010) ("12% of employed adults say they need to market themselves online as part of their job."); Citron, *Harassment*, supra note 6, at 398 (explaining that women who are victims of cyber-victimization miss out on opportunities if they choose to "close their blogs, disengage from online communities, and assume pseudonyms").

physical contact, the opposite may be true.[59] The anonymity provided by the Internet may increase the volume of abusive conduct because it may encourage individuals who would not engage in such conduct offline to do so in the anonymous virtual forum provided by the Internet[60]—people are less inhibited when faced with a computer terminal than when faced with a live person.[61] Cyberspace also enables perpetrators of online abuses to spy on their victims in virtual space for extended periods of time without ever being detected.[62] Furthermore, anonymity naturally makes it more difficult for victims and law enforcement officers to identify and locate cyber-wrongdoers.[63]

Cyberspace also enables perpetrators to manipulate the victim's identity online.[64] Cyber-abusers can both impersonate their victims and can manipulate others' reactions to their victims.[65] They may pretend to be their victims and send inflammatory messages to online discussion groups or social networks under the guise of the victim.[66] Wrongdoers may also engage in identity theft for financial purposes.[67] Additionally, retaliation against the

59. *See generally* Schwartz, *supra* note 21, at 412, 414–15.

60. *Id.* at 414–15.

61. *See* FERTIK & THOMPSON, *supra* note 1, at 76–78 (describing psychological studies and theories on dis-inhibition effects when perpetrators of harm are physically removed from their victims); Cohen-Almagor, *supra* note 3, at 418 ("The Internet has a dis-inhibition effect. The freedom allows language one would dread to use in real life, words one need not abide by, imagination that trumps conventional norms and standards."); Lidsky, *supra* note 8, at 1383 ("Anonymity frees speakers from inhibitions both good and bad. Anonymity makes public discussion more uninhibited, robust, and wide-open than ever before, but it also opens the door to more trivial, abusive, libelous, and fraudulent speech."); Lyrissa Barnett Lidsky & Thomas Cotter, *Authorship, Audiences, and Anonymous Speech*, 82 NOTRE DAME L. REV. 1537, 1575 (2007) ("The technology separates the speaker from the immediate consequences of her speech, perhaps (falsely) lulling her to believe that there will be no consequences. Since the Internet magnifies the number of anonymous speakers, it also magnifies the likelihood of false and abusive speech."); Schwartz, *supra* note 21, at 414–15 ("[A]nonymity makes it easier for the perpetrator to overcome person inhibitions that might have deterred him from carrying out the victimizing behavior if he were confronting his victim face-to-face.").

62. *See generally* Schwartz, *supra* note 21, at 415–16.

63. Goodno, *supra* note 41, at 131.

64. FERTIK & THOMPSON, *supra* note 1, at 78–79 (explaining that such "[a]ttacks by impersonation can be particularly harmful: How do you prove that you didn't really make an offensive comment that appears to be posted under your name? How do you show that it wasn't really you who engaged in a juvenile spat online?").

65. *See* Schwartz, *supra* note 21, at 413–16.

66. *Id.* at 1815.

67. Citron, *Mainstreaming, supra* note 7, at 1817–18 ("Identity thieves use SSNs and biometric data [obtained through data leaks] to empty bank accounts, exhaust others' credit card limits, secure loans, and flip property," which may cause the victim to "face financial ruin.").

victim often follows. It might include the victim being banned from certain websites, being threatened by those who perceive her conduct as inappropriate, or being propositioned by people who have been misled into thinking that she is interested in engaging in unorthodox sexual activities.[68]

Thus, the conduct of a cyber-abuser may be differentiated from that of a physical world wrongdoer in that the online abuser does not necessarily communicate a direct threat to the victim. Instead, he can use general online communications not specifically directed to his victim in order to incite others to directly threaten or harm the victim. In many cases these puppet actors used by the original attacker will not even be aware that their activities are unwelcome or threatening in any way. This may occur where, for example, a puppet believes that the victim harbors rape fantasies and thinks he is merely playing out those fantasies rather than scaring or harming the victim. In several cases involving the popular website Craigslist, bad actors posted messages giving personal details of intended victims, including their home addresses, and saying that the victims harbored rape fantasies.[69] In at least one case, the intended victim was actually raped by a third party who claimed he acted at the victim's invitation and that he was merely fulfilling what he thought was her rape fantasy.[70]

In practice, it is very difficult for victims of these kinds of impersonation attacks to effectively fight back. Because identities are extremely difficult to verify online, it can be almost impossible for a victim to establish that she was not in fact the person who posted the comments in question.[71] It is very difficult for the victim to prove a negative; that is, the "I didn't do it" part of the equation.[72] Even if she can, the victim's revocation may attract more attention to the original content and ultimately make the damage worse.[73] Additionally, even if the victim has a way of proving the negative, it may be extremely difficult for her to connect with the appropriate audience for her rebuttal of the perpetrator's conduct. Many websites, like blogs, will list comments in order of posting. Thus, a rebuttal by the victim may be deprioritized at the end of a comment list where few readers are likely to see

68. *Id.*

69. *Id.*

70. *Id.*

71. *See* FERTIK & THOMPSON, *supra* note 1, at 78–79 (noting the difficulty of showing that it was the abuser, not the victim, who made the online comments).

72. *Id.*

73. *Id.* at 144 ("Replying [to false-flag attacks] often draws more attention to the original content, making the damage worse.").

it.[74] As noted by the founder and the general counsel of ReputationDefender, "many victims feel completely helpless when faced with an anonymous impersonator."[75]

Overall, while online attackers will engage in a variety of damaging types of conduct—harassment, bullying, and stalking—the conduct will have different effects online than its analogs in the physical world. The reputation-damaging information will have a permanent and global quality, and rebuttals by the victim may be difficult to find in practice. Much online conduct will damage a victim's reputation permanently with little recourse because many laws are focused on physical world conduct rather than online communications. The resulting limitations of legal solutions are considered in the following Part.

III. REDRESSING ONLINE WRONGS: GAPS IN THE EXISTING LEGAL FRAMEWORK

This Part examines the currently existing matrix of state and federal laws, both civil and criminal, that potentially could be applied to the kinds of conduct described in this Article. It identifies serious limitations with these laws, particularly those existing in federal criminal laws. While some of these limitations could potentially be redressed through legislative amendments to bring laws into the digital age, many of the limitations do not have feasible legislated solutions. As noted in the Introduction, law alone cannot be the sole answer to problems of cyber-victimization, as laws have jurisdictional and First Amendment limitations. This is not to say that the law should not be an important part of the regulatory matrix aimed at protecting victims against cyber-victimization. However, those making and enforcing laws should be aware of the limitations of legal solutions and should ensure that laws work in tandem with non-legal solutions, such as public education about self-protection online. Furthermore, laws also serve an important expressive function about acceptable modes of online behavior even in situations where their enforcement may be limited by a variety of the factors discussed below.

74. *Id.* ("And a repudiation [of a false flag comment] might never be seen: because some websites list their comments in order by the date they were submitted, a late repudiation may show up far down the page and thus be practically invisible.").

75. *Id.*

A. CRIMINAL LAW

1. Criminal Law Versus Civil Law

Current criminal laws, including those targeted specifically at online conduct, fail to comprehensively deal with today's cyber-abuses. Existing disharmonized state laws cannot effectively deter conduct that typically crosses state or national borders. Criminal law shares with civil law the shortcoming that victims are forced to relive on the public judicial record the humiliation, embarrassment, shame, and fear attached to the defendant's conduct.[76] Therefore, closed criminal trials may be preferable in particularly sensitive cases.[77] However, closed criminal trials raise constitutionality concerns and have been difficult to achieve in practice in other contexts. In addition, absent effective privacy protections, victims of online abuses may be reticent to make complaints or to give evidence in court.[78]

Unlike civil law, criminal law does not typically require a victim to shoulder the costs of a lawyer or the associated costs of litigation. However, effective criminal law does require prosecutors and police to be sufficiently well-versed in the law and in the related online conduct to make a credible case against the abuser.[79] The current lack of reliable data on the prevalence of cyber-stalking, for example, might be attributable to both the failure of victims to bring complaints and the lack of adequate training and funding for police and prosecutors to effectively deal with online abuses.[80]

Despite its shortcomings, criminal law may be a better option than civil law for redressing many online wrongs. Criminal law seeks to punish and deter wrongdoing, while civil law seeks to provide remedies that make a plaintiff whole.[81] Where the concern is with deterring and punishing aberrant conduct, criminal law will be an important part of the regulatory matrix.

---

76. Lipton, *supra* note 4, at 961.

77. *See, e.g.*, Press Enter. Co. v. Super. Ct. of Cal., 478 U.S. 1 (1986) (reversing the order sealing the transcript of lower court proceedings on First Amendment grounds).

78. *See id.* at 3.

79. FERTIK & THOMPSON, *supra* note 1, at 6 ("Many victims of 'routine' online attacks cannot obtain help from the legal system . . . because local courts and lawyers simply don't know how to deal with complex online attacks that might have come from the far side of the world."); Citron, *Harassment*, supra note 6, at 402–03 ("[Police officers] are often either incapable of properly investigating harassment or unwilling to do so until it has traveled offline. Officers often advise victims to ignore the cyber harassment until that time.").

80. *See* Goodno, *supra* note 41, at 156 (discussing the fact that few government officials have had training in cyber-stalking issues).

81. Schwartz, *supra* note 21, at 427 ("[C]yber victimization is better suited to prosecution under criminal law, which seeks to punish and deter wrongdoing, than liability under civil law, which seeks to make a person whole.").

Because of their importance to the regulatory matrix, criminal laws should be better harmonized and specifically targeted to today's most prevalent online abuses. The following examination of current federal and state criminal laws identifies existing gaps in these laws in the online context and makes suggestions for reform.

2. Federal Criminal Law

a) Interstate Communications Act

Federal legislation contains many gaps and inconsistencies when applied to online abuses.[82] The federal laws that are most relevant to online wrongs are mainly found in those sections of the United States Code that deal with electronic communications and computer systems. The Interstate Communications Act, for example, provides that "[w]hoever transmits in interstate or foreign commerce any communication containing any threat to kidnap any person or any threat to injure the person of another, shall be fined under this title or imprisoned not more than five years, or both."[83] This provision has limited application to online abuses because of its requirement of a threat of physical injury,[84] as many online abuses do not contain overt physical threats. In fact, many abusive communications are not specifically directed at their targets but rather are comments *about* their targeted victims on generally accessible websites.[85] The Interstate Communications Act will also not cover situations where a perpetrator poses as a victim online to incite third parties to harass or harm the victim. Thus, this legislation will have limited application to the types of cyber-victimization on which this Article focuses.

b) Telephone Harassment Act

Alternatively, the federal Telephone Harassment Act may have relevant applications to cyber-wrongs. As amended in 2006, the statute prohibits a person from making a telephone call or utilizing a communications device without disclosing his identity and "with intent to annoy, abuse, threaten, or harass any person at the called number or who receives the communications."[86] The revisions to the statute were intended to capture

82. Carnley-Murrhee, *supra* note 6, at 18 (describing federal legislation in the cyber-bullying area as being a "void").

83. 18 U.S.C. § 875(c) (2006).

84. Goodno, *supra* note 41, at 147–48; *see* United States v. Alkhabaz, 48 F.3d 1220 (6th Cir. 1995) (mentioning that an actual threat must be directed to the recipient of the communication).

85. Goodno, *supra* note 41, at 147–48.

86. 47 U.S.C. § 223(a)(1)(C) (2006).

harassing e-mails.[87] While the provision will cover some cyberspace abuses, particularly the sending of threatening or harassing e-mails, it has some limitations. For example, it is limited to acts "in interstate or foreign communications,"[88] but this may not be a very significant hurdle in practice. Courts may hold that any activities involving global communications devices, such as the Internet, occur in interstate or foreign communications.

More importantly, the statute will not cover situations where an Internet communication is not directed towards a particular recipient. The law will not apply to situations in which a perpetrator simply posts information about the victim on a website, or where he poses as the victim.[89] Another limitation of the statute is that it carves out situations where the perpetrator has not remained anonymous.[90] In order for the prohibition to apply, the perpetrator must have failed to disclose his identity and the victim cannot otherwise have gained knowledge of his identity.[91] Again, this statute is unlikely to have significant application in situations involving the kinds of cyber-abuses under discussion in this Article.

c) Interstate Stalking Punishment and Prevention Act

Another recently amended federal statute that may apply to online abuses is the Federal Interstate Stalking Punishment and Prevention Act (FISPPA). This statute prohibits harassment and intimidation in "interstate or foreign commerce" and now specifically extends to conduct that involves using "the mail, any interactive computer service, or any facility of interstate or foreign commerce to engage in a course of conduct that causes substantial emotional distress."[92] As with the Telephone Harassment Act, the extent to which the "interstate or foreign commerce" requirement will limit the potential application of the FISPPA is unclear.

However, the FISPPA improves on the Telephone Harassment Act to the extent that it does not require a communication to be specifically directed

87. Goodno, *supra* note 41, at 148–49.

88. 47 U.S.C. § 223(a)(1).

89. Goodno, *supra* note 41, at 150 (stating that because the Telephone Harassment Statute "applies only to direct communications between the stalker and victim, . . . the amended statute is inadequate to deal with behavior where the cyberstalker indirectly harasses or terrorizes his victim").

90. *Id.* ("It seems odd to only make cyberstalking a crime where the identity of the cyberstalker is unknown [as this] carves out a number of terrifying cases where the victim knows the identity of the cyberstalker.").

91. Lidsky & Cotter, *supra* note 61, at 1590 (explaining constitutional concerns about the validity of this statute on First Amendment grounds because the statute fails to protect constitutionally-protected values inherent in the defendant's anonymity).

92. 18 U.S.C. § 2261A(1), (2) (2006).

to a victim. The FISPPA instead focuses on conduct that utilizes an interactive computer service to create a state of emotional distress in the victim, regardless of whether any communications posted on the computer service were specifically directed to the victim as a recipient.[93] In addition, unlike the Telephone Harassment Act, the FISPPA will apply where the defendant is not anonymous.[94] Like the other federal legislation described above, the FISPPA does not expressly deal with situations where the perpetrator of the online abuse poses as the victim online. Therefore, the FISPPA will similarly be far from a comprehensive answer to cyber-victimization problems.

d) Computer Fraud and Abuse Act

One other federal criminal law that may be relevant to online abuse is the Computer Fraud and Abuse Act (CFAA).[95] This legislation was originally aimed at unauthorized hacking into computer systems and was not focused on personal attacks. However, prosecutors in *Drew* creatively utilized the CFAA to bring criminal proceedings against Drew, who had perpetrated a cyber-bullying attack resulting in the suicide of thirteen-year-old Meier.[96] Drew was the mother of a classmate of Meier and knew that Meier struggled with depression. On the popular social networking site, MySpace, Drew posed as a sixteen-year-old boy named Josh Evans who started a friendship with Meier and later sent her insulting and harassing messages, concluding with a message that the world would be better off without her.[97] Evans never really existed but was rather a fictional creation of Drew, who had developed the Evans persona to find out whether Meier "would say anything negative about Drew's daughter" online.[98]

93. Goodno, *supra* note 41, at 152 ("[T]he newly amended § 2261A addresses many of the shortcomings of the other federal statutes. It does not have a 'true/credible threat' requirement; but rather adopts a standard that measures the victim's 'reasonable fear' or 'substantial emotional distress.' ").

94. *Id.* ("[The FISPPA does not] limit coverage of the 'use' of the computer to only anonymous e-mail messages.").

95. 18 U.S.C. § 1030 (2006).

96. United States v. Drew, 259 F.R.D. 449 (C.D. Cal. 2009); Henderson, *supra* note 28, at 393 (explaining that, despite Drew's egregious acts, only through "creatively interpreting the federal Computer Fraud and Abuse Act [were] federal officials [able to] charge[] her with conspiracy and unauthorized access of a computer"); Lidsky, *supra* note 8, at 1386 (describing the legal action in the Lori Drew case).

97. Henderson, *supra* note 28, at 379.

98. *Id.* at 379–80.

Drew's conduct was not a criminal act under local Missouri law.[99] However, federal prosecutors charged Drew with unauthorized access to a computer under the CFAA. They utilized the criminal trespass provisions of the statute, arguing that Drew had infringed MySpace's terms of service by failing to provide accurate registration information, engaging in abusive conduct, and harassing other people.[100] During the initial trial, a jury found that Drew had infringed provisions of the CFAA relating to making unauthorized access to, or exceeding authorized access to, a computer.[101] However, on appeal, a motion by Drew to acquit and overturn the misdemeanor conviction was granted.[102] The court found that the CFAA would be void for vagueness if it imposed criminal liability on anyone who infringed a website's posted terms of service.[103] Thus, Drew's misuse of the MySpace website could not result in criminal liability under the CFAA. This is not a surprising outcome because, as with the other federal laws discussed in this Section, the CFAA was not enacted specifically to deal with cyber-victimization of this kind.

e) Megan Meier Cyberbullying Prevention Act

In the wake of the Meier incident, federal legislation was proposed that would be more clearly directed at cyber-bullying than any existing federal laws. The Megan Meier Cyberbullying Prevention Act[104] was introduced in 2008 but was never enacted. If it had been implemented, it would have prohibited transmitting a communication "with the intent to coerce, intimidate, harass, or cause substantial emotional distress to a person; using electronic means to support severe, repeated, and hostile behavior."[105] The definitions of "communication" and "electronic means" in the bill were fairly

99. *Id.* at 380.

100. *Id.* at 393.

101. *Drew*, 259 F.R.D. at 453 ("The [trial] jury did find Defendant 'guilty' 'of on the dates specified in the Indictment accessing a computer involved in interstate or foreign communication without authorization or in excess of authorization to obtain information in violation of Title 18, United States Code, Section 1030(a)(2)(C) and (c)(2)(A), a misdemeanor.' ") (internal alterations omitted).

102. *Id.* at 468.

103. *Id.* at 464 (holding that basing a CFAA violation on a terms of service agreement runs afoul of the void-for-vagueness doctrine not only "because of the absence of minimal guidelines to govern law enforcement, but also because of actual notice deficiencies"); *id.* at 467 (explaining that a contrary result would "afford[] too much discretion to the police and too little notice to citizens who wish to use the [Internet]").

104. H.R. 6123, 110th Cong. (2d Sess. 2008).

105. *Id.* § 3(a).

broad and would have encompassed modern Web 2.0 technologies such as blogs and online social networks.[106]

While this legislation would have been broad enough to cover much abusive online conduct, it is arguably overbroad for a variety of reasons. For one thing, it is not confined to a repeated course of conduct and so could inadvertently catch one-time situations where people have acted uncharacteristically out of anger in the heat of the moment.[107] Additionally, while aimed at the Meier incident and drafted with a view to protecting minors,[108] the text of the statute is not expressly limited to conduct involving minors. As a result, the bill may have been unconstitutional on First Amendment grounds because it may have inadvertently sanctioned constitutionally protected expression among adults.[109]

This survey of federal criminal legislation that might potentially apply to cyber-victimization evidences the fact that there are currently no federal laws that are aimed clearly at cyber-victimization. The pastiche of laws that may incidentally catch aspects of cyber-victimization is problematic in practice. The lack of clear federal legislation may not be so much of a problem if there were a series of harmonized state laws that dealt with the issue of cyber-victimization. However, as the following discussion demonstrates, state criminal laws are disharmonized and many have not been brought up to date to deal with challenges posed by the digital age and online social networking technologies.

3. State Criminal Law

One state criminal law that has been particularly well-developed with respect to cyber-victimization is the Missouri statute dealing with online and offline harassment. This law was updated in the wake of the Meier incident

106. *Id.* § 3(b).

107. ROBERT SUTTON, THE NO ASSHOLE RULE: BUILDING A CIVILIZED WORKPLACE AND SURVIVING ONE THAT ISN'T 11 (2007) ("Psychologists make the distinction between states (fleeting feelings, thoughts, and actions) and traits (enduring personality characteristics) by looking for consistency across places and times . . . .").

108. H.R. 6123 § 2 (contemplating that the purpose of the bill is to protect children aged from two to seventeen years old).

109. In the past, legislatures have had difficulty establishing that laws abridging online speech are sufficiently narrowly tailored to survive First Amendment scrutiny. *See, e.g.*, Ashcroft v. ACLU, 542 U.S. 656, 665, 670 (2004) (holding that a statute that imposed criminal penalties for posting content harmful to minors on the Internet was unconstitutional under the First Amendment); Reno v. ACLU, 521 U.S. 844, 849 (1997) (holding that a statute attempting to restrict minors' access to harmful material was unconstitutional under the First Amendment).

to ensure that online bullying would be effectively covered. As now drafted, the Missouri anti-harassment law provides that:

> A person commits the crime of harassment if he or she:
>
> . . . .
>
> (3) Knowingly frightens, intimidates, or causes emotional distress to another person by anonymously making a telephone call or any electronic communication; or
>
> (4) Knowingly communicates with another person who is, or who purports to be, seventeen years of age or younger and in so doing and without good cause recklessly frightens, intimidates, or causes emotional distress to such other person; or
>
> . . . .
>
> (6) Without good cause engages in any other act with the purpose to frighten, intimidate, or cause emotional distress to another person, cause such person to be frightened, intimidated, or emotionally distressed, and such person's response to the act is one of a person of average sensibilities considering the age of such person.[110]

This statute is a good model for legislating against abusive online conduct because it covers multiple communications media, including the Internet, and focuses on the victim's state of mind. While several of the sub-sections require the victim to actually be the recipient of the harasser's communications,[111] the final sub-section does not require a communication directed to the victim.[112] Thus, it could cover a situation where the harasser poses as the victim online and incites third parties to harass the victim. That sub-section also includes a reasonableness requirement with respect to the victim's response. For liability to attach, the victim's response should be appropriate to a person of "average sensibilities considering the age" of the victim.[113]

The statute is not limited to situations in which the harasser engages in a repetitive pattern of abusive conduct towards the victim. Thus, it might catch a one-time situation where a perpetrator acts out of character in the heat of the moment.[114] This may be a factor that courts should consider in applying

110. MO. REV. STAT. § 565.090(1) (2011).
111. § 565.090(1)(3), (4).
112. § 565.090(1)(6).
113. *Id.*
114. SUTTON, *supra* note 107, at 11.

the statute, even though the express words of the statute do not require the courts to identify a pattern of abusive conduct. Additionally, there is no express "legitimate expression" defense. Because of this, courts applying the statute may need to consider whether the defendant's speech should be protected on constitutional grounds.

In recent years a number of other states have also enacted laws targeted specifically at online conduct.[115] However, most states still rely on pre-Internet legislation.[116] Nebraska, for example, maintains stalking and harassment legislation that does not expressly contemplate electronic conduct. The Nebraska Revised Code states that "[a]ny person who willfully harasses another person . . . with the intent to injure, terrify, threaten, or intimidate commits the offense of stalking."[117] In this context, "harassment" is defined as "conduct directed at a specific person which seriously terrifies, threatens, or intimidates the person and which serves no legitimate purpose."[118] "Course of conduct" is defined as "a pattern of conduct composed of a series of acts over a period of time, however short, evidencing a continuity of purpose, including a series of acts of following, detaining, restraining the personal liberty of, or stalking the person or telephoning, contacting, or otherwise communicating with the person."[119]

This legislative approach fails to cover a number of prominent online abuses. Online conduct will not amount to "detaining" or "restraining the personal liberty of" the victim. Online conduct may not even comprise "following" a person if the term "following" is confined to its traditional physical meaning. Additionally, the statutory definition of "course of conduct" contemplates that the perpetrator must have directly targeted the victim. In its application to communications technologies, the statute requires a direct communication to the victim. This requirement does not fit the realities of cyber-victimization, because much online harassment involves the perpetrator posting online messages *about* the victim or even *in the guise of* the victim, rather than communications *directed to* the victim.

New Jersey previously maintained a stalking law similar to Nebraska's law, but legislators updated the New Jersey statute in 2009. The new statute defines "course of conduct" as:

115. Carnley-Murrhee, *supra* note 6, at 18 ("In the void of federal legislation, many states have enacted anti-cyberbullying laws. In the last decade, 19 states . . . have enacted laws that prohibit cyberbullying within state boundaries . . . .").

116. *Id.*

117. NEB. REV. STAT. § 28-311.03 (2010).

118. *Id.* § 28-311.02(2)(a).

119. § 28-311.02(2)(b).

> [R]epeatedly maintaining a visual or physical proximity to a person; directly, indirectly, or through third parties, by any action, method, device, or means, following, monitoring, observing, surveilling, threatening, or communicating to or about, a person . . . ; repeatedly committing harassment against a person; or repeatedly conveying, or causing to be conveyed, verbal or written threats or threats conveyed by any other means of communication or threats implied by conduct or a combination thereof directed at or toward a person.[120]

Unlike Nebraska's law, the New Jersey statute covers activities utilizing *any kind of device* for monitoring, observing, surveilling, threatening, or communicating *to or about* a victim. This is a better model for legislation aimed at online conduct. It clearly covers electronic communications devices as well as online conduct that involves posting messages about a victim, rather than directed to the victim. Nevertheless, it is unclear even under this model whether a perpetrator who disguises himself as the victim and posts messages under the victim's name would be covered. Consider, for example, the scenario where a perpetrator uses the victim's identity to make online comments suggesting that the victim wants to be raped and providing her personal contact details.[121]

It may be difficult for a prosecutor to convince a court that the perpetrator here is effectively "communicating about a person" for the purposes of the New Jersey statute. Where a perpetrator is pretending to *be* another person, he is in a sense communicating *about* that person because anything he does in the guise of the victim indirectly communicates his views—be they true or false—about the victim. However, this conduct is not the same as writing something about the victim in the third person. A court might hold that the legislative intent of the statute was limited to comments about the victim made by a person *other than the victim*, rather than comments made *in the guise of the victim*.

Even if the New Jersey statute is broad enough to cover incitement of third parties to harass the victim, many other state statutes, even relatively recent statutes aimed directly at online conduct, are not as broadly drafted. For example, Florida's relatively new cyber-stalking legislation defines cyber-stalking as engaging "in a course of conduct to communicate, or to cause to be communicated, words, images, or language by or through the use of electronic mail or electronic communication, directed at a specific person, causing substantial emotional distress to that person and serving no

120. N.J. REV. STAT. § 2C:12-10(a)(1) (2010).
121. Citron, *Mainstreaming*, *supra* note 7, at 1839 n.266.

legitimate purpose."[122] Under this provision, there seems to be little doubt that a perpetrator posing as a victim online would not be communicating information *directed at a specific person*.[123] It is at least arguable that posting damaging information *about* an individual is not the same as directing that communication *to* the individual in question. Thus, while the New Jersey statute may cover these kinds of scenarios, the Floridian statute may well not extend this far. The differences in drafting between the criminal laws in different states also cause significant lack of harmonization where abusive online conduct crosses state borders.

4. Suggestions for Drafting Effective Criminal Legislation

Criminal laws focused on online abuses need to deal with a number of issues that many state and federal laws are currently lacking. The laws need to remove requirements of proximity to the victim and requirements of a credible threat of physical harm in order to be effective in cyberspace.[124] On the other hand, legislators may want to retain some laws with a credible threat requirement because such laws may be less open to First Amendment challenge than laws of more general application. However, where legislators have focused on credible threat provisions, resulting laws will have to be supplemented with other regulatory approaches that remedy situations where there is no direct and immediate threat to a victim.[125]

Cyber-abuse laws might also usefully include a requirement of repetitive conduct to avoid catching situations where a person feeling unconstrained by the online medium acts in a one-time capacity without any ongoing intent to

122. FLA. STAT. § 748.048(1)(d) (2009).

123. *See* Goodno, *supra* note 41, at 145 ("Although [the] group of state laws which overtly deal with cyberstalking is clearly a step in the right direction, these statutes have gaps . . . . Few of them explicitly address situations where the cyberstalker dupes an 'innocent' third party to harass."); *id.* at 146 ("As of March 2007, only three states, Ohio, Rhode Island, and Washington, have statutes that explicitly address cases where third parties innocently harass the victim at the cyberstalker's bidding.").

124. *Id.* at 136 ("In cyberstalking cases, a statute with a credible threat requirement does not protect against electronic communications (such as thousands of e-mail messages) that are harassing, but do not include an actual threat."); *id.* at 138 (listing problems of a credible threat requirement as including proof of receipt by the victim, as "a cyberstalker can easily post terrifying messages without ever being in direct contact with the victim or without the victim ever personally receiving the message;" and the "require[ment] the victim to prove that the cyberstalker had the 'apparent ability' to carry out whatever he threatens, . . . [which] is onerous and unnecessary"); Schwartz, *supra* note 21, at 429 ("[N]one of the crimes should require an element of proximity to the victim, nor should they include an 'overt' or 'credible' threat requirement.").

125. *See* discussion *infra* Part IV.

threaten or harass another.[126] Of course, some of these one-time communications can lead to permanent and lasting damage because of the global and permanent nature of online information disclosures.[127] Legislators will need to strike a careful balance to ensure that trivial comments are not sanctioned while more damaging one-time activities can be appropriately deterred.

There may be a number of ways to achieve this balance. For instance, judges could be asked to focus on the substance of the online communication, determining whether the statements made by the perpetrator are likely to cause minor annoyance or major harm to the victim. A comment that someone is "not a nice person" is less egregious than a comment that someone is a "slut" or that she "wants to be raped." Legislation could be drafted to give judges discretion to punish one-time offenders in cases where their online communications are particularly egregious. Another approach would be for legislation to require that the proscribed conduct should *generally* be of a repetitive nature, while not expressly preventing a judge from sanctioning stand-alone communications in appropriate cases.

Criminal legislation aimed at online abuses should also maintain the mens rea requirements that currently exist in state legislation. For example, the Nebraska statute requires willful conduct on the part of the perpetrator.[128] Such a willfulness requirement may go some way towards mitigating any perceived harshness inherent in allowing judges to sanction one-time abuses.

Effective legislation should not require a communication to be sent directly to the victim.[129] Web technologies including blogs, online social networks, wikis, and other online discussion forums are extremely popular. However, they generally do not involve communications sent directly to another. Rather, communications are posted for the whole world to see, or, in a closed network for a particular community to see, such as a community

126. SUTTON, *supra* note 107, at 11; Schwartz, *supra* note 21, at 430 (emphasizing the importance of a requirement of repetitive conduct).

127. Citron, *Mainstreaming*, *supra* note 7, at 1813 (describing the permanence of information posted online); Lipton, *supra* note 4, at 977 (describing the use of internet archives to maintain permanent records of information posted online).

128. NEB. REV. STAT. § 28-311.03 (2010) ("Any person who willfully harasses another person or a family or household member of such person with the intent to injure, terrify, threaten, or intimidate commits the offense of stalking."). *See supra* Section III.A.3.

129. *See* Goodno, *supra* note 41, at 146 (noting problems with current anti-cyber-stalking statutes in Louisiana and North Carolina in that those statutes require harassing communications to be sent "to another").

of "Facebook friends."[130] Communications sent directly to another might merit special attention, particularly if they involve direct and credible threats of harm. However, direct threats are not the sum total of today's damaging online conduct.

Any attempt to legislate against online abuses must be sensitive to First Amendment concerns. The First Amendment protects the right to speak freely and also the right for others to receive speech against government intrusion. Thus, legislation that restricts communication can be problematic, particularly where the information communicated is not in a class of speech that the government has traditionally been able to restrict in certain cases, such as protecting children from pornography or protecting individuals' reputations from false and defamatory statements.

Legislation aimed at prohibiting immediate and credible threats is less likely to be unconstitutional than legislation of broader application. In the cases of broader legislation, the First Amendment might be accommodated by ensuring that the legislation specifies that the speech in question is not constitutionally protected.[131] While it may be difficult to perfectly accommodate the First Amendment, free speech concerns should not be used as an argument against protecting victims. In the physical world, statutes have successfully criminalized offline analogs to many of today's online wrongs.[132] There is no reason why judges cannot continue to draw lines between protected and prohibited speech in the online context.

Another factor that might usefully be incorporated into future legislation would be a reasonable person standard relating to the victim's state of mind.[133] If criminal liability only arises when a victim *reasonably* fears for his or her safety, this may protect expression that could not reasonably be regarded as creating fear or emotional distress in the victim's mind. Thus, unpleasant but predominantly harmless online gossip would be protected, but speech

130. Lipton, *supra* note 4, at 939–40 (describing the concept of "Facebook friends").

131. Schwartz, *supra* note 21, at 431–32; *see* FLA. STAT. § 784.048(1)(b) (2008) (" 'Course of conduct' means a pattern of conduct composed of a series of acts over a period of time, however short, evidencing a continuity of purpose. Constitutionally protected activity is not included within the meaning of 'course of conduct.' "); § 748.048(1)(d) (" 'Cyberstalk' means to engage in a course of conduct to communicate, or to cause to be communicated, words, images, or language by or through the use of electronic mail or electronic communication, directed at a specific person, causing substantial emotional distress to that person and serving no legitimate purpose.").

132. *See* statutes discussed *supra* Section II.A.2.

133. Goodno, *supra* note 41, at 139–40 ("Those stalking statutes that have a reasonable person standard provide the most successful way to prosecute cyberstalking . . . [because] the standard focuses on the victim and whether it is reasonable for her to fear for her safety because of the cyberstalker's conduct.").

that involves egregious damage to a victim's reputation would be sanctioned. The Missouri anti-harassment legislation passed in the wake of the Megan Meier incident is a good example of the incorporation of a concept of the victim's reasonable response to the perpetrator's actions.[134] While reasonable person standards can be difficult to apply in practice, they do give the courts some flexibility in deciding which conduct to sanction and which conduct should be excused.

B. TORT LAW

1. Online Abuses: Common Challenges for Tort Law

Cyberspace interactions could incite tort-based lawsuits, including defamation,[135] privacy torts,[136] and intentional infliction of emotional distress.[137] As with the federal criminal laws discussed above, many tort laws have not been developed specifically to address the kinds of cyber-abuses under consideration in this Article. Defamation and privacy torts are good examples. The common challenges to all of these torts include the ease with which a perpetrator can hide his identity by utilizing a pseudonym and anonymizing technologies, making it difficult to locate and identify him.[138] While it is possible to unmask anonymous actors online,[139] often much

134. *See* MO. REV. STAT. § 565.090(1) (2011); *see* discussion *supra* Section II.A.3.

135. Kara Carnley-Murrhee, *Sticks & Stones: When Online Anonymous Speech Turns Ugly*, U. FLA. L. MAG., Winter 2010, at 21, 22 (citing Lyrissa Lidsky describing the ease of bringing defamation actions for objectionable speech online); Citron, *Cyber Civil Rights*, *supra* note 5, at 87–88 ("Targeted individuals [of online abuses] could . . . pursue general tort claims, such as defamation. False statements and distorted pictures that disgrace plaintiffs or injure their careers constitute defamation per se, for which special damages need not be proven."); Lyrissa Lidsky, *Silencing John Doe: Defamation and Discourse in Cyberspace*, 49 DUKE L.J. 855, 888–92 (2000) (expressing concerns that defamation suits will be the obvious type of legal action to combat online abuses and such suits may stifle online discourse).

136. Carnley-Murrhee, *supra* note 6, at 19 (citing Scott Bauries, who notes that tort actions for invasion of privacy might be a useful approach to cyber-bullying).

137. Citron, *Cyber Civil Rights*, supra note 5, at 88 ("Many victims [of online abuses] may have actions for intentional infliction of emotional distress."); Lyrissa Lidsky, Comment to *New Cyberbullying Case:* D.C. v R.R., PRAWFSBLAWG (Mar. 18, 2010, 3:45 PM), http://prawfsblawg.blogs.com/prawfsblawg/2010/03/new-cyberbullying-case-dc-v-rr.html #comments (noting that intentional infliction of emotional distress is relevant to new cyber-bullying case).

138. For example, this problem includes the TOR anonymizing software. *See Tor: Overview*, TORPROJECT, http://www.torproject.org/overview.html.en (last visited April 14, 2010) ("Individuals use TOR to keep websites from tracking them . . . ."); *see also* FERTIK & THOMPSON, *supra* note 1, at 71 (discussing anonymizing technologies, including TOR).

139. For some examples of "unmasking" litigation, see *In re* Verizon Internet Services, 257 F. Supp. 2d 244 (D.D.C. 2003) (attempting to unmask anonymous online copyright infringers under subpoena provisions in 17 U.S.C. § 512); Columbia Ins. Co. v. Seescandy.com, 185 F.R.D. 573 (N.D. Cal. 1999) (attempting to identify anonymous domain

damage has been done by the time the actor is identified.[140] In addition, unmasking a perpetrator of an online abuse may require a court order.[141] This can be expensive and time consuming, outside the budget of many victims of cyber-abuses.[142] Additionally, as with any litigation, the judicial proceedings potentially bring more publicity to the situation. Non-legal approaches to reputation protection, on the other hand, may be superior in many circumstances involving cyber-victimization because they avoid this level of publicity. These non-legal approaches are discussed in more detail in the next Part.

Another practical problem hypothetically raised by anonymous and pseudonymous online communications is the fact that some plaintiffs may use tort law to unmask the author of defamatory comments not with a view to proceeding with the litigation, but rather with the intention of taking matters into their own hands. Thus, instead of the judicial system working to compensate the victim for the harm she suffered, it creates a platform for her to engage in a campaign of vigilante justice against the potential defendant. Even in situations where the victim herself does not intend to use the defendant's identity to retaliate, the unmasking could lead to others engaging in online attacks against the defendant. Any legal action used to identify anonymous speakers thus runs the practical risk of creating a backlash against the speaker, regardless of whether the speaker might have a valid defense to a tort action. Whether or not the action goes forward, both the plaintiff and the defendant face a potential barrage of new online attacks as a result of the public nature of the lawsuit.[143] Many of the extra-legal approaches to

name cyber-squatter); *In re* Subpoena Duces Tecum to Am. Online, 52 Va. Cir. 26 (2000) (attempting to unmask anonymous online defendants).

140. For example, in the Megan Meier case, the victim had already committed suicide by the time Lori Drew's actions were investigated. *See* discussion *supra* Section II.A.3.

141. *See* Doe I v. Individuals, Whose True Names Are Unknown, 561 F. Supp. 2d 249 (D. Conn. 2008) (seeking to identify anonymous posters on AutoAdmit bulletin board in order to proceed with civil action relating to a number of torts dealing with reputational harm caused to the plaintiffs); *In re Subpoena Duces Tecum*, 52 Va. Cir. 26 (prospective plaintiff sought a court order to unmask identifies of AOL subscribers so that they could be named as defendants in an action for defamation).

142. Lidsky, *supra* note 8, at 1387 (noting that many victims of online defamation lack the resources to bring suit).

143. Bartow, *supra* note 14, at 386–87 (clarifying that the students in the *AutoAdmit* case were first victimized by people they knew in real life, "[b]ut once the women were contextually framed as people who deserved to be mocked and punished (mostly because they objected to the ill treatment [by commencing litigation]) online strangers mobbed and besieged them as well"); *id.* at 399 ("The AutoAdmit administrators seemed to intentionally create a climate that encouraged angry, widespread flaming of anyone who complained about the way they were treated by posters at the AutoAdmit boards.").

protecting online reputations discussed in Part IV do not involve publicity of the original abusive incident and thus avoid the potential for retaliatory attacks against those involved in the original incident. Any tort-based litigation will also involve time and costs that an individual victim may not be in a position to bear.[144] Along with these burdens, a victim would have to relive the shame and humiliation of the abuse during the proceedings, which occur on the public record.[145] While attempting to punish the wrongdoer, the victim would effectively be drawing more attention to the harmful conduct.

Victims of online abuses also face jurisdictional hurdles. Even in cases where the victim knows or is able to ascertain the identity of the perpetrator, he may be in another jurisdiction. Courts in the victim's place of residence may not be able to assert jurisdiction over out-of-state defendants. The costs to the victim of establishing jurisdiction over the defendant, often coupled with the costs of identifying the defendant in the first place,[146] may be prohibitive. Even in cases where the victim is able to identify and assert jurisdiction over an out-of-state defendant, the enforcement of an award for damages or an injunction may be another matter. In many cases it will be impossible or impracticable to enforce a judgment against a remote or impecunious defendant. Additionally, many individual defendants may well be impecunious and, therefore, effectively judgment-proof. In any event, the plaintiff's desired remedy will often not be damages, but rather an injunction to remove a harmful online posting. Some online communications services, such as many common blogs, will not give the original comment poster the technical access to remove the harmful posting.

Another general limitation of tort law is the difficulty associated with attaching liability to parties who provide forums for posting damaging content. These parties are generally immune from liability for the speech of others under § 230 of the Communications Decency Act (CDA).[147] Section

144. *Id.*; Citron, *Cyber Civil Rights*, *supra* note 5, at 91 (noting that many plaintiffs in the cyber-harassment context cannot afford the high costs of litigation); Schwartz, *supra* note 21, at 427 (highlighting, through the perspective a victim of cyber-victimization, how the costs and difficulties of litigation deter victims from seeking civil redress).

145. Lidsky, *supra* note 8, at 1390 ("[S]uing often brings more attention to libelous statements."); Lipton, *supra* note 4, at 961 (noting that as part of the court proceedings, plaintiffs are put in the awkward position of having to relive the humiliation and embarrassment of the images as they are entered into the public record).

146. Lidsky, *supra* note 8, at 1385 (noting uncertain state of law applying to the unmasking of anonymous defendants, which would also add to the costs of unmasking defendants in interstate cases).

147. 47 U.S.C. § 230(c)(1) (2000) ("No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.").

230 immunizes providers and users of "interactive computer services" from liability for information "provided by another information content provider."[148] In other words, where an entity has provided a forum for online speech, that entity shall not be held liable for tortious speech of others who may use the forum for harmful purposes.[149]

Section 230 presents challenges for victims of online abuse both because it immunizes the most obvious party against whom an injunction could be enforced and because it has been very broadly interpreted by the courts.[150] For victims, online service providers are the most effective points in the chain of communications for victims to pursue. In general, the platforms provide the gateways for online discourse, allow victims of online abuses to easily identify them, possess the financial resources to compensate victims by way of damages, and have the technical capacity to remove abusive postings and block abusive posters.

However, under § 230, courts have immunized online service providers from defamation and associated liability for extremely egregious conduct, including comments posted by those with whom the ISP may have a close contractual relationship.[151] Further, the near-absolute immunity[152] of online service providers under § 230 has in practice prevented courts from engaging in meaningful discussions about the standard of care that might be expected

148. *Id.*

149. FERTIK & THOMPSON, *supra* note 1, at 6 ("A legal loophole in the Communications Decency Act makes it impossible to force a website to remove anonymous attacks, no matter how false and damaging they may be."); Citron, *Mainstreaming*, *supra* note 7, at 1839 ("Website operators will enjoy immunity from tort liability under section 230(c)(1) of the Communications Decency Act . . . . Section 230 generally frees online service providers from liability related to the postings of others.").

150. *See, e.g.*, Zeran v. Am. Online, Inc., 129 F.3d 327 (4th Cir. 1997) (immunizing Internet service provider from false and defamatory comments posted by others even in circumstances where it had knowledge of the postings and had not acted swiftly to remove them on the basis of a broad application of § 230 of the Communications Decency Act); Blumenthal v. Drudge, 992 F. Supp. 44 (D.D.C. 1998) (holding that America Online was not liable for comments posted by a commentator it had contracted with to make sensationalist comments on its services because of the application of § 230 of the Communications Decency Act).

151. *See Blumenthal*, 992 F. Supp. at 51–52.

152. Internet service provider immunity has not been absolute as a result of the application of § 230 of the Communications Decency Act. In *Fair Housing Council of San Fernando Valley v. Roommates.com*, 521 F.3d 1157 (9th Cir. 2008), an online service provider was held liable for information that it had created in part. *See also* Citron, *Mainstreaming*, *supra* note 7, at 1839 ("Section 230 generally frees online service providers from liability related to the postings of others. This safe harbor is inapplicable, however, if the website operator helps create the content enabling the criminal activity. The anti-abortion group running the *Nuremberg Files* site exemplifies a party with no immunity under section 230.").

of these service providers absent the statutory immunity.[153] While § 230 immunizes intermediaries and disincentivizes them from monitoring online postings, a victim may effectively have no legal remedy in cases where an anonymous poster cannot be found. There will be no action available against the intermediary and no way of bringing an action against the original poster of the abusive content.[154]

2. Defamation

Defamation law only protects victims against false statements[155] that harm their reputations.[156] Many online statements are true, even if unpleasant or embarrassing. Many are also statements of opinion, which are not typically actionable.[157] Even where the comments are true, the victim bringing an action puts the defendant to proof—on the public record—of the truth of the comments. In many cases this could be very awkward for the plaintiff. For example, a defendant may be required to prove that a plaintiff is, in fact, a "slut." Even evidence of more innocuous things, like proof that the plaintiff was overweight, could be highly embarrassing to the plaintiff.

Despite these practical limitations, defamation law, like all laws impacting social conduct, serves an important expressive function that helps to guide conduct between individuals online.[158] Thus, even the possibility of a small

153. Citron, *Cyber Civil Rights*, *supra* note 5, at 117 (citing "efforts to read a sweeping immunity into § 230 despite its language and purpose have prevented the courts from exploring what standard of care ought to apply to ISPs and website operators").

154. FERTIK & THOMPSON, *supra* note 1, at 65 ("[W]hen the original author cannot be found, the website's refusal to act leaves the victim without any remedy: the false content stays online, forever staining the victim's reputation.").

155. RESTATEMENT (SECOND) OF TORTS § 558(a) (1977) (requiring a "false and defamatory statement" as an element of a defamation action).

156. *Id.* § 559 ("A communication is defamatory if it tends so to harm the reputation of another as to lower him in the estimation of the community or to deter third persons from associating or dealing with him.").

157. *Id.* § 566 ("A defamatory communication may consist of a statement in the form of an opinion, but a statement of this nature is actionable only if it implies the allegation of undisclosed defamatory facts as the basis for the opinion."); Lidsky, *supra* note 8, at 1382 ("A statement can only be defamatory if it asserts or implies objective *facts* about the plaintiff; otherwise, it will be deemed constitutionally protected opinion.").

158. NEIL NETANEL, COPYRIGHT'S PARADOX, 104–05 (2008) (describing how laws often serve "an expressive or symbolic function above and beyond regulating or providing incentives for conduct. . . . Such laws give vent to and help crystallize collective understandings and norms. In turn, by giving legal imprimatur to certain values, they shape future perceptions and choices"); Lidsky, *supra* note 8, at 1390 (noting that a defamation action can serve the function of creating a fear of being unmasked in other potential defendants, and thus can impact online behaviors with respect to parties outside the litigation process).

volume of online defamation actions may serve a larger regulatory purpose in terms of expressing social values more broadly. If we remain aware of the limitations of defamation as an enforcement mechanism, we might nevertheless accept its important expressive functions.

3. Privacy Torts

The American privacy torts were developed at a time well before the age of electronic communications technologies.[159] The laws are focused largely on reasonable expectations of privacy drawn from paradigms involving physical space.[160] One may have a reasonable expectation of privacy behind a locked door but may not have such an expectation in a public street. In the electronic sphere, these expectations break down. Is a Facebook page more like a public forum or a private space? While a Facebook user may exert some control over who accesses her profile, surely more people will access her profile than her private house. An individual Facebook user may not know her Facebook "friends" as well as she knows people she invites into her own home. It is not clear how much privacy she actually expects from her online relationships.

Although different states vary on privacy protections, most maintain some variations on the four privacy torts identified by Dean Prosser in 1960.[161] These torts are: (a) intrusion into seclusion;[162] (b) public disclosure of private facts;[163] (c) false light publicity;[164] and (d) commercial

159. Citron, *Mainstreaming*, *supra* note 7, at 1807 ("Privacy tort law is a product of prior centuries' hazards. In the late nineteenth century, snap cameras and recording devices provided a cheap way to capture others' private moments without detection. The penny press profited from the publication of revealing photographs and gossip about people's personal lives.").

160. Patricia Sánchez Abril, *Recasting Privacy Torts in a Spaceless World*, 21 HARV. J.L. & TECH. 1, 2 (2007) ("[P]rivacy is usually a function of the physical space in which the purportedly private activity occurred . . . ."); *id.* at 3 ("Traditionally, privacy has been inextricably linked to physical space.").

161. William Prosser, *Privacy*, 48 CALIF. L. REV. 383 (1960).

162. RESTATEMENT (SECOND) OF TORTS § 652B (1977) ("One who intentionally intrudes, physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns, is subject to liability to the other for invasion of his privacy, if the intrusion would be highly offensive to a reasonable person.").

163. *Id.* § 652D ("One who gives publicity to a matter concerning the private life of another is subject to liability to the other for invasion of his privacy, if the matter publicized is of a kind that (a) would be highly offensive to a reasonable person, and (b) is not of legitimate concern to the public.").

164. *Id.* § 652E (making a person liable for false light when "(a) the false light in which the other was placed would be highly offensive to a reasonable person, and (b) the actor had knowledge of or acted in reckless disregard as to the falsity of the publicized matter and the false light in which the other would be placed").

misappropriation of name or likeness.[165] None of these torts are obvious matches for the kinds of conduct examined in this Article.

Unpleasant comments about another, whether directed to that other or directed to a general audience, will generally not be an intrusion into another's seclusion. The intrusion tort is based on notions of intrusion into a person's private physical space, rather than intrusions into a person's mental state.[166] The intrusion tort would generally cover cases where someone has entered another's private domain without invitation. It would be difficult to apply the concept to unpleasant comments made in online forums.[167]

While some commentators have argued that it would not be much of a stretch for courts to extend the tort to conduct like hacking people's password protected email accounts,[168] there is as yet no judicial authority on point.[169] Another potential limitation of the intrusion tort, even if it were extended to online conduct, is that it would likely apply only to intrusions into the plaintiff's own private online spaces, such as the plaintiff's email account or Facebook page. It would be difficult to argue that the plaintiff could make out an intrusion claim where the defendant had simply published unpleasant information about her online without specifically impacting some area of the plaintiff's own "online space."

The public disclosure of private facts tort is also problematic. This tort deals with the publication of private and non-newsworthy information, disclosure of which would be "highly offensive to a reasonable person."[170] This tort may apply to some online abuses, but it is not clear where the line would be drawn in terms of identifying sufficiently offensive information. Courts have generally set the bar relatively high and have imposed a

165. *Id.* § 652C ("One who appropriates to his own use or benefit the name or likeness of another is subject to liability to the other for invasion of his privacy.").

166. Citron, *Mainstreaming*, *supra* note 7, at 1828 ("[P]laintiffs cannot bring intrusion on seclusion claims . . . because online postings do not involve invasions of a place or information that society recognizes as private.").

167. *Id.*

168. Citron, *Cyber Civil Rights*, *supra* note 5, at 89 ("Online mobs could face intrusion claims for hacking into password protected e-mail accounts containing private correspondence and conducting denial-of-service attacks to shut down personal blogs and websites.").

169. In fact, Professor Citron cites a case of an intrusion claim involving a creditor making intrusive phone calls as an example of the extension of the tort away from activities by the defendant that involve the defendant's physical presence in the plaintiff's personal space. Citron, *Cyber Civil Rights*, *supra* note 5, at 89 n.204 (citing Donnel v. Lara, 703 S.W.2d 257, 260 (Tex. Ct. App. 1985)).

170. *Id.* at 89 (citing RESTATEMENT (SECOND) OF TORTS § 652B (1977)).

significant burden on plaintiffs to prove offense.[171] While some online communications may meet this test, others will not. For example, photographs of an individual in a sexually explicit and compromising situation may be highly offensive, while comments that a person is fat or slutty, or simply the posting of generally unflattering photographs with unpleasant commentary, may not be sufficiently offensive.

False light publicity is also problematic online.[172] It might be regarded as the little brother of defamation law in the sense that it proscribes publication of information that is not, strictly speaking, false, but that may present an individual in a false light. Litigants will be forced to argue on the public record about the truth or falsity of unpleasant comments and the extent to which recipients of the information formed a false impression of the plaintiff. As with the public disclosure tort, the false light publicity tort—when coupled with the other disadvantages of litigation—is only a limited answer to online abuse.

It is unlikely that the misappropriation tort would be applicable to online harassment because this tort requires the defendant to have made an unauthorized commercial profit from the plaintiff's name or likeness.[173] Most online abuse is non-commercial. It is possible that a plaintiff might bring an appropriation action against the operator of a web service that made money from encouraging personally hostile discourse. For example, a service like AutoAdmit or Juicy Campus[174]—if it adopted a commercial model based on advertising or membership fees and then facilitated abusive online discussions—might be said to be making a commercial profit from another's name or likeness. However, a court may require that the defendant itself be the person who appropriated the plaintiff's name or likeness. Where the defendant has instead provided a forum for others to appropriate names and

171. Lipton, *supra* note 4, at 932 ("The [public disclosure tort] also generally requires that the private facts in question be shameful by an objective standard that is often difficult to prove."); Jonathan B. Mintz, *The Remains of Privacy's Disclosure Tort: An Exploration of the Private Domain*, 55 MD. L. REV. 425, 439 (1996) ("Whether a fact is private by nature—that is, whether a reasonable person would feel seriously aggrieved by its disclosure—is the subject of some disagreement.").

172. *See* Citron, *Mainstreaming*, *supra* note 7, at 1827 ("False light claims require proof of a plaintiff's placement in a false light. [They] do not apply when . . . leaked information causes mischief because it is true.").

173. RESTATEMENT (SECOND) OF TORTS § 652C (1977) ("One who appropriates to his own use or benefit the name or likeness of another is subject to liability to the other for invasion of his privacy.").

174. Cohen-Almagor, *supra* note 3, at 418–420 (discussing the moral responsibility of services like Juicy Campus for harmful postings by their members). These services typically post gossip amongst college students and much of the gossip is extremely hurtful.

likenesses for abusive discourse and has profited from providing that forum, a court may hold that the elements of the tort are not satisfied.[175] In any event, § 230 of the CDA would immunize most providers of these forums from any such liability.

4. Intentional Infliction of Emotional Distress

The intentional infliction of emotional distress tort may be more promising than the other torts. This tort requires a finding of extreme or outrageous conduct on the part of the defendant that caused, or was intended to cause, severe emotional distress.[176] Some courts have been willing to find for plaintiffs where a defendant exploits a power disparity between the parties or otherwise takes advantage of a vulnerable plaintiff.[177] It may be easier to convince a court of such a power disparity or vulnerability in online abuse cases than to focus on the content of the communication, which is generally necessary in defamation and some of the privacy torts.[178]

While it is difficult to determine by contemporary social standards what might satisfy the extreme or outrageous conduct limb of the tort, many cases of cyber-bullying and cyber-harassment will have powerful emotional effects on their victims. For example, a recent online posting on the "Casual Encounters" board on Craigslist said that a teenager had rape fantasies and enjoyed pornography. As a result of the posting, the teenager was inundated with pornographic messages and confronted by men at her work.[179] Even though the perpetrator's conduct involved merely posting a message on Craigslist, his action—coupled with the substance of the message and the harmful results—may amount to extreme or outrageous conduct. Although the intentional infliction of emotional distress action may theoretically be a promising avenue for individuals harmed by cyber-abuses, this tort still suffers from the same practical limitations as the other torts in terms of time, cost, jurisdictional challenges, and potential increased public humiliation for either or both parties.

175. *But see* Citron, *Mainstreaming*, *supra* note 7, at 1836–43 (suggesting the development of an action for tortious enablement of criminal conduct or tortious conduct by website operators).

176. RESTATEMENT (SECOND) OF TORTS § 46(1) (1977); Citron, *Cyber Civil Rights*, *supra* note 5, at 88–89.

177. Citron, *Cyber Civil Rights*, *supra* note 5, at 88 ("Courts are more willing to consider conduct 'outrageous' if the defendant exploited an existing power disparity between the parties or knowingly took advantage of a vulnerable plaintiff.").

178. For example, defamation actions and false light publicity claims focus, at least in part, on the *content* of the communications made by the defendant about the plaintiff.

179. Citron, *Mainstreaming*, *supra* note 7, at 1818.

C. CIVIL RIGHTS LAW

Professor Citron has recently suggested that a civil rights agenda might expand to combat certain cyber-abuses.[180] Civil rights laws include doctrines against race discrimination that might interfere with a victim's ability to make a living and laws that criminalize threats of force designed to intimidate or interfere with a person's employment based on that person's race, religion, or national origin.[181] In other words, civil rights law addresses the kinds of conduct typically described as harassment in the sense that victims are targeted because of their membership in a particular protected class.[182] Title VII of the Civil Rights Act of 1964 prohibits gender discrimination as a result of intimidation, threats, or coercion aimed at interfering with employment opportunities.[183] While this law focuses on employment opportunities, many online abuses aimed at women and minorities do prevent members of those groups from engaging in employment or "making a living" because many people's businesses are now conducted wholly or partly online.[184] Additionally, many people's physical world employment opportunities may be affected negatively if employers have access to harmful information about an individual and decide not to hire or promote that individual. Additionally, employers may even terminate that individual's employment.

Civil rights suits entail some advantages, such as easing the costs of litigation for victims of online harassment[185] and reaching wrongs that would otherwise escape criminal or tort liability.[186] However, while Citron's suggested civil rights agenda is well reasoned, it remains untried. Adopting a broader civil rights agenda aimed at online abuses would confront many of the same problems as extending tort and criminal law to cover online abuses.

---

180. Citron, *Cyber Civil Rights*, *supra* note 5, at 89 ("A meaningful response to abusive online mobs would include the enforcement of existing civil rights laws . . . .").

181. *Id.* at 91–92 (citing 42 U.S.C. § 1981 (2006) and 18 U.S.C. § 245(b)(2)(C) (2006), respectively).

182. *See* discussion *supra* Section I.A.3.

183. Citron, *Cyber Civil Rights*, *supra* note 5, at 92–93 (describing how gender discrimination that "interferes with a person's ability to make a living can be pursued under Title VII of the Civil Rights Act of 1964").

184. MADDEN & SMITH, *supra* note 58, at 3 (noting that twelve percent of employed adults now report that they need to promote themselves online); Citron, *Cyber Civil Rights*, *supra* note 5, at 93 (describing how online attacks can be particularly intimidating for women and minorities, given that online attacks can often interfere with an individual's work).

185. Citron, *Cyber Civil Rights*, *supra* note 5, at 91 (noting that civil rights lawsuits have statutory damages that make pursuing such cases more affordable and attractive when compared to traditional tort suits).

186. *Id.* ("[C]ivil rights suits may reach wrongs that would otherwise escape liability. These include victims' rights to be free from economic intimidation and cyber harassment based on race and gender.").

Enforcing authorities, including judges and, in some cases, the United States Attorney General,[187] would have to be willing to act against online abusers. These authorities may be reticent to do so absent a clearer mandate. Additionally, civil rights laws, along with tort and criminal law, raise problems of identifying often anonymous defendants.

Civil rights law, if applied online, might help some groups targeted by online abusers, such as women and racial and religious minorities. However, other sets of common victims, such as children, are unlikely to be covered unless an individual victim also happens to fall into a statutorily protected class. In other words, civil rights law might provide some protections against cyber-harassment, but not necessarily against cyber-bullying. As noted above, cyber-bullies generally target individuals for reasons outside membership in a protected class.[188] Bullies may target people whom they perceive as a threat, or whom they regard as weak—potentially including people who are poor, inarticulate, overweight, or socially inept. None of these traits would fall within the umbrella of civil rights protection.

This survey of the limitations of criminal, tort, and civil rights laws evinces the fact that law in and of itself will never be a full solution to problems of cyber-victimization. While these laws serve an important expressive function and may apply to certain kinds of cyber-abuses, they need to be supplemented by extra-legal approaches to redressing online wrongs. Laws per se suffer from difficulties of identifying an anonymous or pseudonymous defendant and having effective jurisdictional reach over the defendant. This is particularly problematic in the case of disharmonized state criminal laws. Even if plaintiffs can identify their defendants—which may require an expensive and time-consuming court order—they are often judgment-proof. Attaching liability to online service providers for the comments of anonymous posters is also problematic because of the operation of § 230 of the Communications Decency Act. Many extra-legal approaches to cyber-abuse avoid these problems because they can protect a victim without requiring expensive and public litigation. Some of the more obvious of these approaches are discussed in the next Part.

IV. EXTRA-LEGAL APPROACHES TO ONLINE WRONGS

This Part examines several extra-legal regulatory approaches that could impact the ways in which people interact online. It focuses on regulatory

187. *Id.* at 93 (noting that the Attorney General can file civil suits for injunctive relief under Title VII of the Civil Rights Act of 1964).

188. *See* discussion *supra* Section I.A.2.

modalities that can empower victims to control their own reputations online. It also suggests ways in which public and private funding might be usefully funneled into educational initiatives to assist individuals in preventing online harms, using abuse reporting hotlines, and creating programs that facilitate relevant industry self-regulation. One advantage of focusing on extra-legal initiatives is that their development is less likely to be hindered by concerns about the First Amendment than legal developments. This is because private actors such as reputation management services and private education providers are not generally subject to First Amendment guarantees.[189]

A. THE NEED FOR A MULTI-MODAL APPROACH

Because of the limitations inherent in the legal system, a broader multi-modal regulatory approach is necessary to combat online abuses. The idea of combining regulatory modalities in cyberspace is not new.[190] However, web 2.0 technologies increase the need for a complex interplay of regulatory approaches in order to identify and facilitate the development of appropriate online behaviors.[191] Relevant regulatory modalities will likely include social norms,[192] system architecture,[193] market forces,[194] public education,[195] and the use of private institutions.[196]

---

189. U.S. CONST. amend. I ("Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.").

190. Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501, 507–08 (1999) (suggesting four regulatory modalities for cyberspace: legal rules, social norms, market forces, and system architecture); Lawrence Lessig, *The Architecture of Privacy*, 1 VAND. J. ENT. L. & PRAC. 56, 62–63 (1999) (suggesting the same four norms of regulation for online privacy); Lipton, *supra* note 4, at 925 ("[L]egal regulation alone is unlikely to solve society's video privacy problems." Rather, "a multi modal approach that combines [the following] six regulatory modalities" may be necessary: "legal rules, social norms, system architecture, market forces, public education, and private non-profit institutions.").

191. *See, e.g.*, LAWRENCE LESSIG, CODE: VERSION 2.0, at 5 (2006), *available at* http://codev2.cc/download+remix/Lessig-Codev2.pdf ("Cyberspace demands a new understanding of how regulation works. It compels us to look beyond the traditional lawyer's scope—beyond laws, or even norms. It requires a broader account of 'regulation,' and most importantly, the recognition of a newly salient regulator. That regulator is the obscurity in this book's title—Code.").

192. *See* Steven Hetcher, *Using Social Norms To Regulate Fan Fiction and Remix Culture*, 157 U. PA. L. REV. 1869 (2009) (discussing the role of norms in regulating online fan fiction and remix communities); Jacqueline D. Lipton, *Copyright's Twilight Zone: Digital Copyright Lessons from the Vampire Blogosphere*, 70 MD. L. REV. 1 (2010) (discussing the development of norms of authorship and fan use of copyright works online); Jacqueline D. Lipton, *What Blogging Might Teach About Cybernorms*, 4 AKRON INTELL. PROP. J. 239 (2010) (discussing the development and identification of norms in the blogosphere); Mark Schultz, *Fear and Norms and Rock & Roll: What Jambands Can Teach Us About Persuading People To Obey Copyright Law*, 21

In global online communities, laws must interact with other regulatory modalities to achieve a comprehensive approach to combating abuses. Legislators and judges will learn much from observing the development of market solutions,[197] technological solutions, and emerging social norms[198] that impact online behavior. Participants in online communities will also learn something from a legislature's willingness to proscribe certain conduct. Public education, through news stories, and publicly or privately funded education initiatives, is also an important part of the framework. Appropriately tailored educational initiatives will assist in the development of online norms.

B. EMPOWERING VICTIMS TO COMBAT ONLINE ABUSES

1. Reputation Management Techniques

"Your online reputation is your reputation. Period."[199]

A key to protecting individuals from online abuses is to empower those individuals to protect themselves without needing to resort to the legal system. This may be more complex than it sounds because protecting an individual's reputation involves both teaching an individual to be more careful about information she discloses about herself online and to attempt to monitor information that others disclose about her online. There are a variety of ways in which individuals can guard their own reputations online. Some methods involve learning to control information that an individual

BERKELEY TECH. L.J. 651 (2006) (discussing the role of norms in regulating copyrights in certain sectors of the music industry); Katherine Strandburg, *Privacy, Rationality, and Temptation: A Theory of Willpower Norms*, 57 RUTGERS L. REV. 1235, 1238 (2005) ("Social norms are primarily understood as means to coordinate the behavior of individuals in a social group. Thus, norms may help to solve coordination problems—by determining how pedestrians pass one another on the street—and collective action problems—by stigmatizing littering—when individually rational behavior leads to collectively undesirable results.").

193. LESSIG, *supra* note 191, at 5; Joel Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553 (1998) (describing how digital technology can be utilized as a regulatory mechanism for online conduct).

194. Ann Carlson, *Recycling Norms*, 89 CALIF. L. REV. 1231, 1253 (2001) ("Markets constrain behavior through price. If the price of gasoline rises dramatically, people will drive less.").

195. *See* Lipton, *supra* note 4, at 979–80.

196. *Id.* at 980–81.

197. For examples of private online reputation management services, see, e.g., REPUTATIONDEFENDER, http://www.reputationdefender.com (last visited May 20, 2010); YOUDILIGENCE, http://www.youdiligence.com (last visited May 20, 2010).

198. *See, e.g.*, OWN WHAT YOU THINK, http://www.ownwhatyouthink.com (last visited May 20, 2010) (campaign to promote more accountable and responsible online discourse).

199. FERTIK & THOMPSON, *supra* note 1, at 16.

releases about herself on the Internet—such as personal anecdotes and photographs. Educating individuals about the risks of disclosing private information online is an important aspect of protecting online reputation. For example, individuals can be encouraged to use maximum privacy protections on services such as Facebook[200] and to ensure that they have sufficient security measures installed on their personal computers to prevent others from accessing their personal information. Individuals can also be trained to build positive online content about themselves. This serves as a form of online insurance and potentially prevents negative content from making it onto the first page of search results about them.

It is also important to educate people about the risks inherent in releasing their personal information online. However, bigger problems occur when the individual's friends or acquaintances disseminate the harmful information. While a potential victim may secure her own computer and may be careful about what she discloses about herself online, she has very little control over what others disclose about her. She also has very little control over attacks directed specifically to her.

Individuals now have to be vigilant not only about what they disclose about themselves online, but also in monitoring what others may be disclosing about them.[201] Individuals may also need to be aware of currently available ways to combat damaging content about them. This may involve learning how to conduct a personal reputation audit[202] and asking providers of online forums to monitor, police, and remove damaging content.[203] It may also involve knowing how to use other online tools, such as astroturfing, and search engine optimization to repair damage.[204]

---

200. In fact, Facebook has recently simplified its privacy settings to better enable its users to make use of its privacy-protecting technologies. Mark Zuckerberg, *Making Control Simple*, THE FACEBOOK BLOG (May 26, 2010, 10:55am), http://blog.facebook.com/blog.php?post=391922327130.

201. MADDEN & SMITH, *supra* note 58, at 2–3 (noting that individuals are indeed becoming more vigilant over time about self-monitoring and observation of information available about others online); Robert McGarvey, *Is Bad Taste the New Taste? Social Media Is Changing Our Sense of What's Acceptable—And What's Not*, THINK, Spring/Summer 2010, at 24, 26–27 (2010) (describing a situation in which an Ohio executive found out that an old friend had posted online a photo of him in a drunken stupor from his youth, and the steps he attempted to take to have the photo de-tagged from social networking websites).

202. FERTIK & THOMPSON, *supra* note 1, at 162–87.

203. Bartow, *supra* note 14, at 415 (noting that some people who run online forums do a lot of policing on their own initiative).

204. *Id.* at 426–27 (describing the use of astroturfing by reputation management services such as ReputationDefender). Astroturfing involves seeding the Internet with positive or neutral content generated by the individual herself in an attempt to drown out the abusive content. The term "astroturfing" has arguably begun to take on negative connotations in the

Search engine optimization techniques involve the manipulation of search engine results so that positive or neutral information is prioritized in searches above harmful information.[205] Many of these tools are currently utilized by private online reputation management services. However, individuals can learn how to use them without having to pay the fees charged by the private services.[206] Some literature is now available to assist individuals in learning strategies that commercial reputation management services have typically utilized.[207]

Another mechanism for protecting some aspects of an individual's online reputation is available under the notice and takedown provisions of the Digital Millennium Copyright Act (DMCA).[208] These provisions allow a copyright holder to send a notice to a website operator requesting removal of material that infringes a copyright. If the operator complies with the notice, it can avoid copyright infringement liability.[209] The effectiveness of this technique in the hands of a private individual will depend on the extent to which the individual actually holds copyright in damaging text and images about her. In many cases, third parties will have generated such materials.[210] Thus, the victim will not have a copyright claim that could support the use of the DMCA.[211]

The ability of an individual to make use of any of the techniques described here will depend on her awareness of the techniques. One of the problems for victims of online abuses has been lack of awareness of how to protect one's own reputation online, outside of resorting to the law or engaging the services of a private reputation management company. While

sense that some people may now associate it with conduct like seeding the Internet with false political information. However, in the absence of a better term, "astroturfing" is utilized in this paper in reference to seeding any type of positive or neutral information about an individual in an attempt to protect her reputation online.

205. Bartow, *supra* note 14, at 427 (describing use of search engine optimization techniques by private reputation management services).

206. *See id.* at 421 ("It is doubtful that any reputation defense service offers clients anything that they cannot do for themselves if they have a basic understanding of applicable laws, of the way that search engines function, and of the vulnerability of search engines to targeted manipulation.").

207. For example the founder and general counsel of ReputationDefender have released a book detailing some strategies for individuals to protect their own online reputations. FERTIK & THOMPSON, *supra* note 1.

208. 17 U.S.C. § 512(c) (2006).

209. § 512(c)(1)(C).

210. For example, a person who takes an embarrassing photograph of the victim will generally hold copyright in the photograph. *See* DANIEL J. SOLOVE, THE FUTURE OF REPUTATION: GOSSIP, RUMOR, AND PRIVACY ON THE INTERNET 184 (2007).

211. *See id.*

private reputation management services unquestionably have a useful place in protecting people's online reputations, they are motivated by profits and they can charge high fees[212] for doing a number of things that private individuals could do on their own if they knew how.[213] Cynically, one might also argue that private reputation management services actually benefit from online abuses and that it is in their own commercial interests for online abuses to continue to some extent.[214]

2. Education

The increased ability of private individuals to protect their reputations online might put more pressure on private reputation management services to develop new products and services, or to price their services more competitively. The question remains how best to empower private individuals to protect their reputations online. Clearly, some level of public education would be useful. Education might be government funded and targeted at schools and other public institutions,[215] such as libraries and universities. It may also be that private non-profit organizations, such as the Electronic Frontier Foundation[216] and the Electronic Privacy Information Center,[217] will play an increasingly important role. Education can focus both on empowering victims to protect their reputations against online attacks, and on training participants in online communities to behave in a socially acceptable manner more generally.

---

212. Bartow, *supra* note 14, at 423–26 (describing fees charged by ReputationDefender for its various services).

213. FERTIK & THOMPSON, *supra* note 1, at 235–47 (describing ways in which private individuals and small businesses can act to protect their own online reputations); Bartow, *supra* note 14, at 421 ("It is doubtful that any reputation defense service offers clients anything that they cannot do for themselves if they have a basic understanding of applicable laws, of the way that search engines function, and of the vulnerability of search engines to targeted manipulation.").

214. Bartow, *supra* note 14, at 419 ("[T]he greater the quantity of sexual harassment toward affluent victims that appears on the Internet, the wealthier reputation defense services can become."). Of course, one could make similar arguments about the home security system industry. This industry unquestionably profits from home burglaries. However, that is not to say that they condone the conduct of burglars.

215. While government regulation of speech generally raises First Amendment concerns, the government is generally able to attach speech-restrictive provisions to funding legislation without running afoul of the First Amendment. United States v. Am. Library Ass'n Inc., 539 U.S. 194 (2003) (upholding legislation that required internet filtering as a condition of libraries accepting government funding).

216. *See* ELECTRONIC FRONTIER FOUNDATION (EFF), http://www.eff.org (last visited Apr. 20, 2010).

217. *See* ELECTRONIC PRIVACY INFORMATION CENTER (EPIC), http://www.epic.org (last visited Apr. 20, 2010).

A number of private organizations already provide information about online harms in addition to providing tools for addressing them. Many of these organizations focus on protecting children from online predators and bullies. For example, NetSmartz provides information to parents, guardians, educators, law enforcement authorities, and children about staying safe on the Internet.[218] NetSmartz also offers free multimedia safety presentations that can be used in classrooms and other communities. Its website also links to the Internet Crimes Against Children website,[219] a government-sponsored educational initiative to protect children online.

Another service aimed at protecting children online is GetNetWise,[220] which provides information, advice, and free online tools for keeping children safe online. It contains an inventory of suggested software tools parents might utilize to protect their children as well as critiques of the available software options. It also provides a suggested contract that parents can enter into with their children containing guidelines to help children stay safe in their online interactions.[221]

C. A CRITIQUE OF EXISTING COMMERCIAL REPUTATION MANAGEMENT SERVICES

While an increasing number of services provide free information and tools for combating online abuses, some of the most well known services are the for-profit reputation management services like ReputationDefender, Reputation Hawk, and YouDiligence. Private reputation management services raise some practical concerns, despite their usefulness. As noted in the previous Section, reputation management services offer a variety of options for protecting individual reputations online. They will monitor an individual's online reputation,[222] typically for a monthly fee.[223] They then

218. *See* NETSMARTZ, http://www.netsmartz.org (last visited May 20, 2010).

219. *See* FVTC INTERNET CRIMES AGAINST CHILDREN TRAINING & TECHNICAL ASSISTANCE PROGRAM, http://www.icactraining.org (last visited May 20, 2010).

220. *See* GETNETWISE ABOUT… KIDS' SAFETY, http://kids.getnetwise.org (last visited May 20, 2010).

221. *See Tools for Families*, GETNETWISE ABOUT… KIDS' SAFETY, http://kids.getnetwise.org/tools/toolscontracts (last visited July 8, 2010).

222. Focusing on popular services like MySpace and Facebook. *See* Bartow, *supra* note 14, at 424 ("ReputationDefender claims it will monitor blogs and sites like MySpace, Facebook, Xanga, Bebo, Flickr, LiveJournal, and many others for any material that might be damaging or distressing to a client . . . .").

223. *See id.* ("The SEARCH part of [ReputationDefender's] service requires payment of a subscription fee, which costs $14.95 per month, with discounts to people who sign up for one or more years at a time."). YouDiligence currently charges between $9.99 and $14.99 per month for its monitoring services. *See* YOUDILIGENCE, http://www.youdiligence.com (last visited June 5, 2011).

provide monthly reports to a client summarizing information about the client available online.[224]

If the service detects information that the client objects to, the service will offer to remove the damaging content from the Internet at a charge relating to each piece of information the client wants to destroy.[225] The client may even ask the service to target information that may be true.[226] However, many reputation management services now focus on the removal of slanderous or damaging information and refrain from removing much information that is true or newsworthy.[227] Most reputation management services regard their techniques for sanitizing a person's online reputation as "proprietary"[228] and do not disclose those techniques publicly.[229] However, their methods likely include: (a) using notice and takedown procedures from the DMCA;[230] (b) contacting blogs and other web hosts and asking them to remove damaging information;[231] (c) astroturfing the Internet with newly manufactured neutral or positive information about their clients;[232] and (d) engaging in search engine optimization techniques to ensure that neutral and positive information about their clients is prioritized in search results.[233]

---

224. *See* Bartow, *supra* note 14, at 423 (citing ReputationDefender's "SEARCH" process).

225. *See id.* at 424 ("The DESTROY aspect of the enterprise costs $29.95 per piece of unwanted information, with no guarantee of positive or sustainable results.").

226. *Id.* (noting that ReputationDefender does not require information to be inaccurate, harassing or defamatory in order to remove it; and that the service is prepared "to sanitize any inconvenient truths"); *id.* at 425 ("ReputationDefender is also willing to mask or bury accounts of mainstream news stories even if they are true.").

227. *FAQ, Can You Help Remove Absolutely ANY Content from the Internet*, REPUTATIONDEFENDER, http://www.reputationdefender.com/faq/ (last visited July 8, 2010) (describing the service's removal procedures and its limitations, such as its refusal to remove media articles, government records, and issues of "legitimate public interest").

228. Bartow, *supra* note 14, at 421 (noting ReputationDefender's reference to its techniques as being "proprietary").

229. *Id.* at 425 ("ReputationDefender refuses to disclose the exact nature of its so-called destruction tools, and presumably its competitors do as well."). More recently, ReputationDefender has disclosed a number of its reputation management techniques. FERTIK & THOMPSON, *supra* note 1.

230. Bartow, *supra* note 14, at 421 (discussing use of the notice and take-down provisions of copyright law by online reputation management services); *see also* discussion *supra* Section III.B.

231. Bartow, *supra* note 14, at 425 ("In addition to utilizing the notice and take-down procedures of copyright law, another of ReputationDefender's vaunted proprietary techniques is apparently to send e-mails to blogs and websites hosting information that its clients want to disappear.").

232. *Id.* at 426–27.

233. *Id.* at 427 ("Another avenue available to reputation defense organizations is Search Engine Optimizing, which has been characterized by at least one legal scholar as fraud. It is

These services provide a number of advantages over legal solutions to online abuses, including the fact that several of them now have many years of experience with reputation management and have established solid working relationships with websites that host harmful communications.[234] The use of private commercial services does not raise the specter of a First Amendment challenge. As noted above, many laws directed at curtailing online speech may raise First Amendment concerns and may be open to constitutional challenge.[235] Reputation management services also avoid many of the practical problems associated with litigation including jurisdictional challenges and difficulties identifying a defendant in the first place. A commercial service does not need to identify or locate a potential defendant in order to engage in astroturfing or search engine optimization. Resort to a reputation management service also avoids drawing public attention to the damaging content.[236] Harmful content can simply be unobtrusively de-prioritized in search engine results.

However, reliance by individuals on these commercial services has a number of disadvantages, despite the obvious benefits. One of the key disadvantages relates to cost and equity issues. Many of the victims of online harassment and other abuses will not be able to afford the fees charged by these services.[237] While engaging a service to monitor one's reputation on the Internet may be relatively affordable,[238] paying fees to repair one's online reputation may be prohibitive for many. Additionally, while these commercial services are available—at least to wealthier people—there may be less pressure on the government to act. If the government thinks the market is

an effort to manipulate search engine results for profit."); Lidsky, *supra* note 8, at 1390 (describing services provided by commercial reputation management companies).

234. FERTIK & THOMPSON, *supra* note 1, at 206 ("Professionals have built thousands of websites and know exactly how to optimize them to rank the highest in Google and other search engines. They often know the right tone to strike and the right balance of links to create. And professionals often have an arsenal of deals with specialized websites that allow rapid improvement in search results.").

235. *See, e.g.*, Diane Leenheer Zimmerman, *Is There a Right To Have Something To Say? One View of the Public Domain*, 73 FORDHAM L. REV. 297, 348–49 (2004).

236. Lidsky, *supra* note 8, at 1390 ("Hiring a reputation management company sometimes provides an attractive alternative to suing for libel because suing often brings more attention to the libelous statements.").

237. Citron, *Cyber Civil Rights*, *supra* note 5, at 106 ("Few free or inexpensive resources are available for defending one's online reputation, and the services of groups like ReputationDefender are expensive and beyond the means of many victims.").

238. The fees for monitoring one's reputation are typically in the ballpark of around $10 to $15 a month. *See supra* note 223; Bartow, *supra* note 14, at 424 (noting that ReputationDefender charges $14.95 per month to monitor a client's online reputation).

handling the problem, government agencies may put less effort into investigating and prosecuting the abuses.[239]

The apparent availability of reputation management services may also negatively impact the level of monitoring undertaken by those who provide online speech forums. These forum providers are generally immunized from tort liability for the speech of others under § 230 of the CDA.[240] This legislation is a powerful disincentive for online service providers to monitor and act against harmful speech. The perceived availability of reputation management services may further disincentivize online forum providers from monitoring their own forums. Service providers might assume that they need not monitor their forums because not only are they generally immune from legal liability for the speech of their contributors, but also if there is a problem, they will receive a notice from a reputation management service. Better yet, the reputation management service may simply take care of the problem through astroturfing or search engine optimization without requiring any action on the part of the online service provider.[241] Recent statistics suggest that many online service providers will quickly remove harmful information on request.[242] However, it is difficult to gauge how proactive any of these services are in removing damaging information absent a formal request to do so.

Another practical limitation of reputation management services is that the actions they take to protect their clients' reputation may backfire dramatically. Most of them will not offer any guarantees of success[243] or refunds for backlash caused by their activities.[244] For example, one ReputationDefender client, Ronnie Segev, suffered a significant backlash as a result of ReputationDefender's efforts to remove embarrassing content

239. *Id.* at 422 (describing how market solutions benefit affluent parties that can afford services such as ReputationDefender but paradoxically decrease governmental incentive to provide low-cost solutions to individuals who are likely to need such services the most).

240. *See* discussion *supra* Section III.B.1.

241. *See* discussion *supra* Section IV.B.

242. MADDEN & SMITH, *supra* note 58, at 4 (noting that a significant majority of people who have sought removal of information about them posted online have been successful).

243. The disclaimer in YouDiligence's terms of service is a good example of how little these services guarantee in practice. *See Terms of Service*, YOUDILIGENCE, http://www.youdiligence.com/yd/TermsOfUse.htm (last visited May 20, 2010) (outlining the service's broad and strongly stated disclaimer of any warranties of any kind to customers).

244. Bartow, *supra* note 14, at 424 (noting that reputation management services do not give "guarantees of positive or sustainable results"); Citron, *Cyber Civil Rights*, *supra* note 5, at 105 (noting how online attackers are likely to be emboldened when a victim attempts to stay online or fight back, as many attackers aim to force victims off the Internet).

about him from a website.[245] After ReputationDefender sent a notice to the website operator requesting removal of the harmful information,[246] a blogger from the website wrote a scathing post entitled "Ronnie Segev and ReputationDefender Can Eat a Dick."[247]

Another limitation of private reputation management services is that they cannot do much in the face of personal attacks directed at a victim, rather than posted publicly online. The tools utilized by reputation management services do not specifically address situations where a person is sending harassing and abusive communications directly to a victim. In the Megan Meier scenario, for example, where harmful communications were directly sent to the victim, there is little that a private reputation management service can do. This may be a situation where legal solutions are more appropriate. Victims of such abuses can, in relevant jurisdictions, rely on cyber-bullying and cyber-harassment laws if police and prosecutors are prepared to act on the complaints.[248]

D. EFFECTIVE REPUTATION MANAGEMENT

1. Enhanced Access to Reputation Management Services

Empowering individuals to fight online abuses themselves requires a number of strategies, many of which rely largely on the availability of funding and public education. For example, pro bono legal services could be encouraged to take on more online abuse cases if they could be staffed and funded to do so. There is also no reason why more pro bono reputation management services could not be developed if government or other funding were available.

The development of more pro bono reputation management services and public education initiatives would be a useful supplement to currently available commercial reputation management services. As noted above,

---

245. Bartow, *supra* note 14, at 425–27 (discussing the Segev incident). The comments posted online were details of a scheme Segev was involved in during his youth to defraud Priceline of an airticket.

246. ReputationDefender sent the following message:

> We are writing to you today because our client, Ronnie Segev, has told us that he would like the content about him on your website to be removed as it is outdated and disturbing to him. Would you be willing to remove or alter the content? It would mean so much to Mr. Segev, and to us. Considerate actions such as these will go a long way to help make the Internet a more civil place.

Id. at 426.

247. *Id.*

248. *See* discussion *supra* Section II.A.

commercial services are expensive and out of the reach of many victims of online abuses.[249] At the same time, some of the tools they utilize are readily available to private individuals who know how to use them.[250] If victims of online abuses had better information about some of these tools, they could more easily protect themselves online without necessarily having to pay for a commercial reputation management service.

If appropriate funding were available, victims might also have the option of using a pro bono reputation management service. Naturally the choice to pay for a commercial service would still be available. If individuals were savvier about protecting their own reputations online and more pro bono options were available, the commercial services may be incentivized to develop even more sophisticated solutions to online abuses. They would after all be competing for increasingly technologically sophisticated clients with more practical options. This could ultimately lead to the development of new innovations for protecting individual reputations.

Access to existing legal remedies for online abuses might also be improved if pro bono legal services were better equipped to take on these cases. Many legal clinics and other pro bono services may not deal with many of these cases because they are unfamiliar with the relevant laws, or they may assess current law as inadequate to cover the victims' harms. A reworking of laws, and increased funding and education to those providing pro bono services to victims of online harassment, might usefully redress the balance here.

2. Cyber-abuse Hotlines

Another extra-legal approach to protecting online reputation is the increased use of internet hotlines that can be established on a voluntary basis by various online service providers.[251] Users of online services can be empowered to report online abuses by telephone, fax, email, or submission of an online form. Hotlines should ideally be as confidential as possible, and those who claim abuse should be given some information about how complaints will be handled and the circumstances under which complaints

249. Citron, *Cyber Civil Rights*, *supra* note 5, at 105 (noting often prohibitive expense of utilizing these services).

250. FERTIK & THOMPSON, *supra* note 1, at 234–47 (advising individuals and small businesses on techniques to self-protect online reputations).

251. Cohen-Almagor, *supra* note 3, at 426–27 (critiquing several existing Internet hotlines).

may be referred to a public authority.[252] Of course, this assumes the existence of an appropriate authority to deal with relevant complaints.

The British Internet Watch Foundation exemplifies the hotline approach in reporting illegal online conduct involving certain types of internet content including: (a) sexual images of children, (b) obscene adult content, (c) material inciting racial hatred, and (d) inappropriate behavior towards a child online.[253] Users can report such content in a variety of ways including submission of an online form.[254] In the United States, the CyberTipline is another example of a hotline for reporting certain damaging conduct much of which involves children, such as child prostitution, child molestation, and sex tourism involving children.[255]

Among the more salient advantages of hotlines in the context of online abuses is the fact that they can open up channels of communication between victims, observers of harmful conduct, and law enforcement authorities.[256] Hotlines also enable ready collection of data about online abuses including data about the nature of prevalent abuses and demographic characteristics of typical abusers and victims.[257] Hotlines can thus enable law enforcement agencies to gain a clearer picture of online abusive conduct and to target enforcement activities appropriately. Reports generated by hotlines, when released to the public, can also serve an important public education function by increasing awareness of damaging online conduct. This can enable individuals as well as pro bono and private services to develop targeted tools to respond to specific abuses.

3. Evolving Online Norms

Social norms interact with other regulatory modalities in cyberspace as in the physical world. Norms both influence and respond to legal and market developments. For example, a law may alter normative behavior by requiring compliance or simply by expressing appropriate behavioral standards.[258] Markets will often respond to online norms. For example, reputation management businesses developed as society became less civil online and a market demand grew for tools to protect individual reputations. The

252. *Id.*

253. *See* INTERNET WATCH FOUNDATION (IWF), http://www.iwf.org.uk/reporting.htm (last visited May 19, 2010).

254. *Id.*

255. *See* NATIONAL CENTER FOR MISSING AND EXPLOITED CHILDREN, http://www.missingkids.com (last visited May 19, 2010).

256. Cohen-Almagor, *supra* note 3, at 427.

257. *Id.*

258. *See* NETANEL, *supra* note 158, at 104–05 (on law's expressive functions).

question today is how to develop norms that foster more civil and accountable online communities.

One approach is to develop online forums that promote community standards of responsibility and accountability. For example, to counter the Juicy Campus debacle,[259] a Princeton student created the Own What You Think website, asking students to pledge not to visit anonymous gossip sites and to be accountable for their own online communications.[260] The site sports the banner headline "Anonymity = Cowardice."[261]

Of course, norms may work in opposing directions and society, or large sectors of society, may simply become desensitized to many online abuses. As one commentator has noted: "Maybe we soon will simply yawn in boredom the next time we see a tweet typed in an inebriated rant, or a Facebook photo of a friend—or perhaps even ourselves—dancing on a table with bloodshot eyes."[262] Even if we become desensitized to these kinds of communications, one would hope that we never become desensitized to dangerous and harmful conduct like cyber-bullying and harassment involving threats of physical harm, or online communications that seriously damage an individual's livelihood or reputation.

4. Industry Self-Regulation

Market self-regulation initiatives may also be an important part of the regulatory matrix. Self-regulation may be adopted voluntarily or may be a result of pressure from customers or from governments. In the cyber-abuse context, the relevant industry is difficult to define. Online abuses occur in a variety of online forums including social networking sites, blogs, and even online multi-player games. Search engines like Google will be implicated here because they play such a significant role in determining which Internet users see what information. Self-regulation initiatives in at least some industries might serve an important educational and normative function for those involved in online communications more generally.

An example of the interplay between government and market self-regulation in the social networking context is the 2008 Joint Statement on Key Principles of Social Networking Sites Safety adopted between MySpace

259. Cohen-Almagor, *supra* note 3, at 419–20 (describing harmful online postings about college students on the juicycampus.com website).

260. *See* OWN WHAT YOU THINK, http://www.ownwhatyouthink.com (last visited May 19, 2010).

261. *Id.*

262. McGarvey, *supra* note 201, at 29.

and state Attorneys General.[263] These principles are aimed at protecting children from inappropriate and harmful online content.[264] They encompass strategies such as developing software tools to protect children from harmful content, designing social networking sites in a way that prevents minors from accessing inappropriate content, educating parents and children about online safety issues, and ensuring that social networking sites cooperate with law enforcement agencies in protecting children online.[265]

Companies might also be compelled to self-regulate if they were subjected to a system of labeling, naming, and shaming websites that provide a platform for cyber-wrongs. For example, several years ago in the United Kingdom, the culture minister and her shadow minister presented the idea that online service providers might be named and shamed into dealing more proactively with violent and sexually explicit conduct on their sites.[266] The hope is that by the government calling websites out on irresponsible behavior, websites would potentially regulate their own content.

This is a difficult result to achieve in practice because it involves cooperation between a central agency and some realistic pressure brought to bear on websites to take action against harmful online conduct. Additionally, because of the global nature of the Internet, definitions of "harmful conduct" may vary from community to community and country to country. Some countries, with stronger free speech protections, may protect speech that others sanction. Of course, certain speech, like realistic threats of harm, should not be protected anywhere. However, beyond that, it is difficult to draw clear lines about what kinds of conduct should lead to naming and shaming.

Some other recent examples of self-regulation involve Google's relatively new Google Search Wiki and Google Profile service.[267] Google's experimental Search Wiki enables Internet users to make comments on search results.[268] Thus, a victim of reputational harm could use the service to contextualize or refute a criticism made about her. However, the Search Wiki comments are not displayed unless an Internet searcher goes out of his way

263. MySpace and State Attorneys General, Joint Statement on Key Principles of Social Networking Sites Safety 1 (Jan. 14, 2010), http://ago.mo.gov/newsreleases/2008/pdf/MySpace-JointStatement0108.pdf.

264. *Id.*

265. *Id.*

266. Patrick Wintour, *Web Providers To Be Named and Shamed over Offensive Conduct*, THE GUARDIAN (Nov. 15, 2008), http://www.guardian.co.uk/technology/2008/nov/15/internet-children.

267. *See generally* FERTIK & THOMPSON, *supra* note 1, at 91 (describing these services).

268. *Id.*

to enable them. Additionally, anyone can comment on any search result, so there is no way for an Internet user to screen for true or false comments. Google now also offers a Google Profile service that enables individuals to write a brief profile about themselves.[269] These profiles may be displayed at the bottom of Google search results for personal names. However, this service is currently limited in its impact because of the placement of the profiles at the bottom of a page of search results where they may be missed by a searcher. Additionally, they have limited use for people with common names.

The Wikipedia online dispute resolution service provides another form of self-regulation that could potentially protect one's online reputation. It is more and more common for individuals to be profiled on Wikipedia, which is a participatory and interactive repository for knowledge on many different subjects.[270] The participatory nature of Wikipedia means that an individual will not necessarily control information about her that may be posted on a Wikipedia page.[271] Wikipedia has its own online dispute resolution procedure to verify the accuracy of information posted, and individuals harmed by false or de-contextualized postings may utilize this service.[272] While this approach is specific to Wikipedia, other online service providers could adopt similar approaches if they want to assist their users in combating reputational harms.

V. CONCLUSION

> "The Internet is a powerful and wonderful tool that has ushered in a new information age. If purposely misused, however, the internet can be terrifying, and even deadly."[273]

The Internet is an unparalleled global communications medium. However, online interactions can be harmful, leading to emotional suffering and physical harm. The current legal system has gone some way towards protecting victims of online harms. However, the law still has a long way to go. Legal remedies will always suffer limitations related to time, cost, and jurisdictional challenges in a borderless online world. Further, the

269. *Id.*

270. *Id.* at 182 ("Wikipedia . . . is a free collaboratively edited encyclopedia. Anyone can edit any article, and anyone can create new articles.").

271. *Id.* ("The vast majority of readers will find no relevant information about them on Wikipedia, but every now and then a malicious editor will slip an inappropriate reference or an unsubstantiated attack into the site.").

272. *Id.* at 182–83 (describing applications of Wikipedia's dispute resolution procedure to reputational injuries).

273. Goodno, *supra* note 41, at 125.

embarrassment and humiliation often associated with a victim bringing a complaint will chill much legal action.

Like many other aspects of internet regulation, effective responses to online abuse will require a multi-modal regulatory framework. Regulatory modalities such as social norms, public education and market forces will need to interact to create more comprehensive responses to online abuses. Reputation management services play an important role in this regulatory matrix, but they are subject to their own limitations. Current approaches to online abuse might be improved if the existing commercial services could be supplemented with more easily affordable pro bono services, and if individuals could be empowered themselves to engage proactively in reputation management strategies. Increased funding for, and use of, hotlines would also be a step forward both in combating specific abuses and in providing more reliable and comprehensive data about online abuses. Attempts at industry self-regulation, potentially in concert with government incentives, would also be a useful development.

A number of the proposals made in this article would require funding, which is always a tall order, particularly in troubled economic times. On a more positive note, most of the suggestions made here are not particularly difficult to implement. They predominantly take advantage of tools already available and apply them in new ways. The extra-legal remedies advocated here also have the advantage that they do not rely on government action other than potentially some funding, so they do not run into significant First Amendment concerns.[274] Additionally, enhancing private mechanisms avoids some of the problems typically inherent in litigating to identify and to assert jurisdiction over often anonymous or pseudonymous defendants. Tackling online abuses is a global problem. Private bodies acting in concert with each other and with domestic governments have a better chance of reaching optimum solutions than governments acting alone.

274. For example, governments are generally permitted to fund programs that impact speech. *See, e.g.*, United States v. Am. Library Ass'n Inc., 539 U.S. 194, 194 (2003) (upholding a funding program that required libraries to filter internet access as a condition of accepting government funding).

EXPLAINING THE DEMISE OF THE DOCTRINE OF EQUIVALENTS

David L. Schwartz[†]

TABLE OF CONTENTS

---

† Assistant Professor of Law, Chicago-Kent College of Law. E-mail: dschwartz@kentlaw.edu. I would like to thank John Allison, Bernadette Atuahene, William Birdthistle, Christopher Buccafusco, T.J. Chiang, Colleen Chien, Chester Chuang, Kevin Collins, Christopher Cotropia, Dennis Crouch, Jason Du Mont, John Duffy, Michael Evans, Jeanne Fromer, John Golden, Eric Goldman, Stuart Graham, Richard Gruner, Sarah Harding, Timothy Holbrook, Paul Janicke, Jay Kesan, Hal Krent, David Pekarek Krohn, Edward Lee, Peter Lee, Mark Lemley, Raizel Liebler, Edward Manzo, Tyler Ochoa, Kristen Osenga, Lee Petherbridge, Matthew Sag, Christopher Schmidt, Carolyn Shapiro, Ted Sichelman, Ned Snow, Stephanie Stern, Katherine Strandburg, Suja Thomas, Emerson Tiller, Peter Yu, Corey Yung, Samantha Zyontz, and the participants at the Northwestern Law—Searle Center Roundtable on Empirical Studies of Patent Litigation, the Intellectual Property Scholars Conference 2009, the Santa Clara Patent Colloquium, the 2009 Midwest Law and Economics Annual Meeting, the Loyola Law School—Los Angeles Symposium on the Federal Circuit as an Institution, Drake Law School—IP Scholars Roundtable 2010, and workshops at the University of Illinois and Chicago-Kent College of Law for their insightful comments and suggestions. I also would like to thank Tom Gaylord, research librarian at Chicago-Kent, and my research assistants Teresa Clark and Elsie Washington for their hard work and dedication. Finally, as always, I would like to thank my wife Naomi for her patience and support.

I. INTRODUCTION

Over 150 years ago, the Supreme Court expanded the potential scope of patents by adopting a doctrine to prevent "substantial copies" of an invention by providing coverage over inventions that are "equivalent" to that patented.[1] The doctrine of equivalents had been consistently applied by courts until its rapid "demise" between the mid-1990s and the mid-2000s.

In recent years, distinguished academics have studied the so-called "demise" of the doctrine of equivalents. Professors John Allison, Mark Lemley, and Lee Petherbridge have each empirically analyzed this doctrine. All of their studies conclude that successful use of the doctrine has substantially diminished over time.[2] With very little detail or support, Allison and Lemley speculated that trial court judges caused the death of the doctrine of equivalents after they were tasked with construing the scope of patent

1. Winans v. Denmead, 56 U.S. 330 (1853).

2. John R. Allison & Mark A. Lemley, *The (Unnoticed) Demise of the Doctrine of Equivalents*, 59 STAN. L. REV. 955, 958 (2007); Lee Petherbridge, *On the Decline of the Doctrine of Equivalents*, 31 CARDOZO L. REV. 1371, 1378–79 (2010) [hereinafter Petherbridge, *Doctrine of Equivalents*]; Lee Petherbridge, *The Claim Construction Effect*, 15 MICH. TELECOMM. & TECH. L. REV. 215, 233 (2008) [hereinafter Petherbridge, *Claim Construction Effect*].

claims in *Markman v. Westview Instruments*.[3] They contended that if trial judges learned the technology and ruled against the patentee on claim construction,[4] they desired to resolve the entire dispute, which required adjudicating the equivalents claim against the patentee as well.[5] Petherbridge offered a different theory. He provided evidence that the decline occurred years after *Markman*, only after a significant Supreme Court decision on the doctrine of equivalents, *Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co.*,[6] which reduced the applicability of the doctrine.[7] While each of these researchers noted that the doctrine of equivalents had decreased in its successful use and provided some grounds for the decrease, none clearly explained why. As such, the cause and precise mechanism behind the so-called "demise" of the doctrine of equivalents have largely been mysterious.

This Article sheds light on the mystery by providing a novel theoretical model and extensive empirical evidence to explain the decline of the doctrine. In large part, the demise occurred as a result of two complementary forces discussed for the first time in this Article: "doctrinal reallocation" and "doctrinal displacement."[8]

Appellate courts have the power to engage in "doctrinal reallocation" by altering adjudicatory control of a doctrine. Control can be regulated in numerous ways. For example, the decision-maker tasked with adjudicating the doctrine in question can be shifted at the trial-court level from the jury to the judge, or vice-versa. Or, the appellate court may increase its control by reviewing lower court decisions de novo instead of under a clearly erroneous standard. These are forms of "doctrinal reallocation." Once control of a doctrine increases, a higher court may alter the prominence of the doctrine in the adjudicatory process. When a judge instead of a jury makes a decision, cases are more easily resolved by summary judgment and readily reviewed on appeal. Lowering the deference in appellate review permits the higher court to more easily correct decisions with which it disagrees.

---

3. Allison & Lemley, *supra* note 2, at 958; *see* Markman v. Westview Instruments, Inc., 52 F.3d 967, 978 (Fed. Cir. 1995) (en banc) (overruling precedent permitting juries to construe patent claims), *aff'd*, 517 U.S. 370 (1996).

4. Claim construction refers to the process of determining the literal scope of a patentee's rights.

5. Allison & Lemley, *supra* note 2, at 958.

6. 535 U.S. 722 (2002).

7. Petherbridge, *Doctrine of Equivalents*, *supra* note 2, at 1391–92.

8. For a broader discussion of these new theories, see David L. Schwartz & Ted M. Sichelman, Doctrinal Displacement (2011) (unpublished manuscript), *available at* http://ssrn.com/abstract=1832705.

Reallocating a doctrine to a judge does more than simply empower a court to profoundly influence the importance of that doctrine—there are further-reaching consequences. A change in the importance of a given doctrine may lead to the decline of other, typically related, doctrines in the same field of law. These "displaced" doctrines may have been extremely important prior to being dislodged.

The theories of doctrinal reallocation and displacement explain the chain reaction resulting in the demise of the doctrine of equivalents. Initially, the Court of Appeals for the Federal Circuit reshaped patent litigation in *Markman*. There, it overruled previous precedent and held that claim construction was an issue of law that should be exclusively examined by a judge rather than a jury.[9] Shortly thereafter, the Federal Circuit in *Cybor Corp. v. FAS Technologies* ruled that claim construction should be reviewed by the appellate court using the expansive de novo standard.[10] These decisions triggered significant changes in patent litigation, not only in connection with claim construction, but also with respect to the doctrine of equivalents. Claim construction, which was significant but not critical before these decisions, rapidly became the centerpiece of patent litigation. Nearly contemporaneously, the doctrine of equivalents declined in importance. In effect, these doctrines switched places in terms of significance as a judicial tool. This switch occurred, in part, because both doctrines are essentially substitute ways for the court to evaluate the proper reach of an invention. When one means of evaluating scope—claim construction—became relatively easier for the court to apply, courts began to rely upon it more. As a result of this shift, patent litigation today is far different than litigation in the early 1990s.

The purpose of this Article is two-fold. Testing the theories of doctrinal reallocation and doctrinal displacement, the Article first presents evidence that doctrinal reallocation occurred in patent litigation in the wake of *Markman*. In the aftermath of that reallocation, the Federal Circuit increased the importance of claim construction in patent litigation. This observation, based in part on empirical data, highlights the ability of the court to shape patent law after reallocating the responsibility for claim construction to the judge.

9. Allison & Lemley, *supra* note 2, at 978; Markman v. Westview Instruments, Inc., 52 F.3d 967, 978 (Fed. Cir. 1995) (en banc) (overruling precedent permitting juries to construe patent claims), *aff'd*, 517 U.S. 370 (1996).

10. Cybor Corp. v. FAS Techs., Inc., 138 F.3d 1448 (Fed. Cir. 1998) (en banc).

Second, the Article examines the demise of the doctrine of equivalents in the lens of doctrinal displacement.[11] Specifically, this Article provides empirical evidence showing that the rise in the importance of claim construction foreshadowed a sharp decline in the importance of the doctrine of equivalents. Reducing the significance of the doctrine of equivalents was in line with the Federal Circuit's goal of curbing the unpredictability of patent jury trials.

This Article has four additional Parts. Part II explains the theory of doctrinal reallocation and displacement in the context of adjustments to claim construction and the doctrine of equivalents by the Federal Circuit. Part III describes the empirical study design and methodology. Part IV propounds hypotheses about the expected effect of *Markman* and *Cybor* on patent litigation and the expected demise of the doctrine of equivalents. It also delivers empirical results relating to the displacement of the doctrine of equivalents in patent law. Part V concludes with some brief remarks about the significance of the findings.

II. THEORY

Doctrinal reallocation[12] and doctrinal displacement[13] may occur in almost any area of law. Scholars have recognized that a shift in decision-making authority from jury to judge (and vice-versa) can alter substantive doctrine,[14] but they have yet to provide a formal theory to explain this jurisprudential phenomenon. This Article provides such a theory and tests it with data in the

11. Although the doctrine of equivalents and claim construction both affect the ultimate reach of a given patent claim, they are separate doctrines. *See* Tate Access Floors, Inc. v. Interface Architectural Res., Inc., 279 F.3d 1357, 1367 (Fed. Cir. 2002) ("The doctrine of equivalents expands the reach of claims beyond their literal language."). However, others have noted that the doctrine of equivalents and claim construction overlap. *See, e.g.*, Kevin Emerson Collins, *The Reach of Literal Claim Scope into After-Arising Technology: On Thing Construction and the Meaning of Meaning*, 41 CONN. L. REV. 493 (2008) (arguing that courts at times allow literal claim scope to grow to encompass after arising technologies, a task traditionally performed by the doctrine of equivalents).

12. Doctrinal reallocation refers to reallocating the responsible decision-maker for a doctrine. It is a loose analogy to asset reallocation. In asset reallocation, a portfolio is adjusted among different asset classes to reduce risk.

13. Doctrinal displacement refers to displacing one doctrine with another. *See infra* Section II.B.

14. For a review of the history of legal realists and their views on procedure, see NEIL DUXBURY, PATTERNS OF AMERICAN JURISPRUDENCE 55–169 (1995); see also Frank B. Cross, *Legal Process, Legal Realism and the Strategic Political Effects of Procedural Rules* 2 (Univ. of Tex. Sch. of Law, Law & Econ. Working Paper No. 065, 2005), *available at* http://ssrn.com/abstract=837665 ("Realists argue that the apparently neutral procedural requirements are created or applied precisely for their ideological implications.").

context of the Federal Circuit, which reviews nearly all appeals involving issues of patent law.[15] This Part briefly explains the theories of doctrinal reallocation and doctrinal displacement using two doctrines from patent law: claim construction and the doctrine of equivalents.

A. DOCTRINAL REALLOCATION AT THE FEDERAL CIRCUIT

There are many reasons why a court may wish to increase the importance of a doctrine. Typically the change is made because the court wants to implement an institutional preference. These institutional preferences often are designed to reduce the uncertainty of litigation outcomes—in other words, courts aim to improve the predictability and stability of the adjudicatory process. Separately, courts are worried about institutional legitimacy and fairness. Alternatively, courts may have concerns about excessive caseloads, and they raise the significance of a doctrine to permit swifter resolution of lawsuits.

Before discussing the theory in detail, several terms must be defined. First, this Article uses "importance" of a doctrine to mean how central a doctrine is to an area of law, including, for example, how often it is raised and how often it is dispositive of the entire dispute.[16] Second, "control" refers to the ability of judges to determine the importance of a doctrine not just for a given case, but for an area of law as a whole. The importance of the doctrine may change with or without changing the substance of the underlying law.

Any court can alter the importance of a doctrine by various procedural mechanisms. Through these procedural mechanisms, examples of which are described *infra*, district and appellate judges can change the quantum of their control over the doctrine. One final note about jurisprudential control is appropriate. Increased control over a doctrine does not always lead to the doctrine becoming more important. Rather, control provides courts with the *ability* to make the doctrine more or less important. The direction of importance, either increased or decreased, is typically dictated by the court's overall institutional goal. If the goal is to lighten judicial workload, for example, the court may increase the importance of a statute of limitations. Alternatively, if the goal is to heighten predictability, the court may decrease the importance of an unstructured jury doctrine.

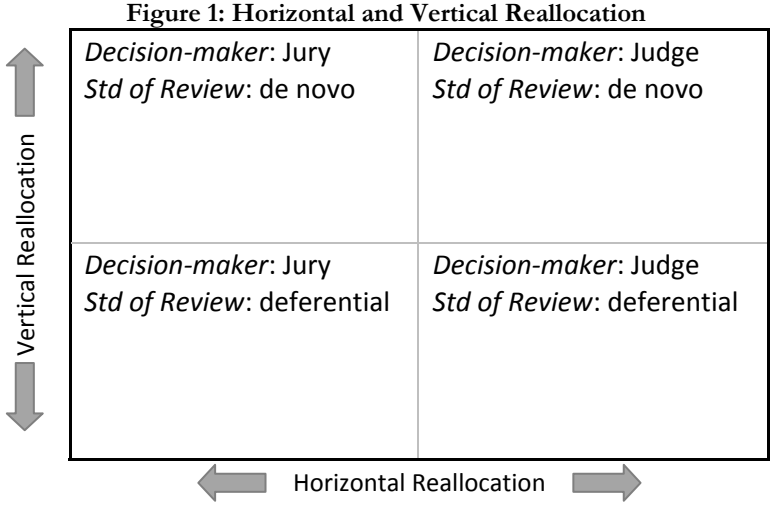
---

15. 28 U.S.C. § 1295(a) (2006); Holmes Grp., Inc. v. Vornado Air Circulation Sys., Inc., 535 U.S. 826, 829 (2002).

16. This Article analyzes various aspects of judicial decisions to evaluate importance, necessarily making the assumption that these decisions are a good measure of the centrality of an issue to the universe of disputes including unlitigated disputes.

Judges can increase or decrease their control over a doctrine in several fashions. One method is to "reallocate" the decision-making authority over a particular doctrine among various institutional actors. "Horizontal" reallocation refers to shifts of authority within the trial court—for example, between the judge and jury. "Vertical" reallocation refers to shifts of authority between upper and lower courts—for instance, by changing the standard of review on appeal. Horizontal reallocation and vertical reallocation are depicted in the two-by-two matrix shown in Figure 1, *infra*.[17]

Figure 1: Horizontal and Vertical Reallocation



Within the trial court, decision-making responsibility is divided between the jury and the judge. In Figure 1, judges have more control when the district court judges are the decision-makers (shown on the right) than when the juries are the decision-makers (shown on the left).[18] Horizontal reallocation alters the control over the shifted issue. Judicial determination of an issue provides more control to judges. Judges are repeat players in litigation and hear the same issue more than a jury selected and seated for a single case. Of course, the Seventh Amendment bounds horizontal

17. To fit reallocation into a simple two-by-two matrix, all deferential standards of review are denoted as the same. In reality, there is a difference between the types of deferential review such as substantial evidence and clear error. *E.g.*, Frank B. Cross, *Decisionmaking in the U.S. Circuit Courts of Appeals*, 91 CALIF. L. REV. 1457, 1502–03 (2003).

18. Changing the decision-maker on an issue between judge and jury is not a substantive doctrinal shift per se. *See* Emerson H. Tiller & Frank B. Cross, *What Is Legal Doctrine?*, 100 NW. U. L. REV. 517, 517 (2006) (noting that doctrine comprises "[j]udicial opinions [that] create the rules or standards"). The underlying substantive *doctrine* on the shifted issue remains unchanged. In other words, the rules or standards are unchanged; it is the evaluator of these rules or standards who changes.

reallocation, and more specifically, the extent to which issues may be recalled from the jury.[19]

Vertical reallocation involves control within the hierarchy of the judicial system, namely between lower courts and upper courts. Not all aspects of a trial court judgment are reviewed on appeal with the same scrutiny. Courts use the standard of review to differentiate among the various appealed matters.[20] The most permissive standard of review from the perspective of a higher court—de novo—permits the upper court to review the matter without deference.[21] De novo (shown on the top in Figure 1) provides the most control for the higher court because it can freely revise findings from the court below.[22] Other standards of review such as clearly erroneous or abuse of discretion (shown on the bottom in Figure 1) provide more control to the lower court because the upper court cannot modify the lower decisions unless there is clear error.

1. The Claim Construction Doctrine

In the mid-1990s, the Federal Circuit desired to make patent law more predictable.[23] Consequently, this Article argues, the Federal Circuit elevated the importance of one aspect of patent litigation: claim construction. Claim construction refers to the process of determining the literal scope of a

19. The Seventh Amendment states: "In Suits at common law, where the value in controversy shall exceed twenty dollars, the right of trial by jury shall be preserved, and no fact tried by a jury shall be otherwise re-examined in any Court of the United States, than according to the rules of the common law." U.S. CONST. amend. VII.

20. Amanda Peters, *The Meaning, Measure, and Misuse of Standards of Review*, 13 LEWIS & CLARK L. REV. 233, 240 (2009) ("If appellate courts examine all of the decisions made below without any deference to rulings, then the trial court's proceedings are meaningless. However, a deferential standard of review not only works to preserve the integrity of the trial court, it also serves to protect the appellate court's valuable time and resources.").

21. *Id.* at 246 ("Courts using de novo review examine the trial court's application of the law without affording the lower court discretion.").

22. Admittedly, very little if anything falls within the top-left box.

23. *See* Paul R. Michel, *The Challenge Ahead: Increasing Predictability in Federal Circuit Jurisprudence for the New Century*, 43 AM. U. L. REV. 1231, 1235 (1994) ("I therefore argue that our court works best when it so defines generic legal rights that, in most individual situations, the parties to a potential lawsuit could, if willing, reason together and agree on the likely outcome of prospective litigation. Specifically, the parties' lawyers could reliably predict how our court would ultimately rule on the matter in dispute. Surely, moving in the opposite direction—toward more uncertainty of rights, more unpredictability of adjudicatory outcomes, and therefore more lawsuits—is an undesirable and ultimately an unsustainable result."). Obviously, there are limits to how predictable patent law can be, especially given that it must apply to currently undiscovered technologies.

patentee's rights.[24] The word "construction" in claim construction refers to interpreting the meaning of the words used in a patent claim.[25] Claim construction occurs in various contexts, and perhaps most prominently in litigation.[26]

The law of claim construction is embodied in a series of canons of construction, which are similar to the canons of statutory construction.[27] The Federal Circuit believed that claim construction was key to making patent law

24. Abbott Labs. v. Sandoz, Inc., 544 F.3d 1341, 1358–60 (Fed. Cir. 2008) (" 'Claim construction' is for the purpose of explaining and defining terms in the claims, and usually requires use of words other than the words that are being defined."). The word "claim" in claim construction refers to the claims of the patent. All patents contain at least one and typically multiple claims. John R. Allison & Mark A. Lemley, *The Growing Complexity of the United States Patent System*, 82 B.U. L. REV. 77, 103 (2002) (reporting that a random sample of patents issued between 1996 and 1998 had an average of 14.87 claims per patent). The claims are each a single sentence written in a technical manner. *See* U.S. PATENT & TRADEMARK OFFICE, U.S. DEP'T OF COMMERCE, MANUAL OF PATENT EXAMINING PROCEDURE (MPEP) § 706.03(d) (8th ed. Rev. 8, July 2010). The U.S. Patent Office (Patent Office) has detailed formatting and structural rules that apply to claims. *Id.* §§ 608.01(i)–.01(o). The exact language in the claims is carefully considered by patent attorneys and the Patent Office. Patent attorneys spend substantial time selecting the language to use in patent claims. Jason M. Okun, *To Thine Own Claim Be True: The Federal Circuit Disaster in* Exxon Chemical Patents, Inc. v. Lubrizol Corp., 21 CARDOZO L. REV. 1335, 1341 (2000) ("It is in the process of drafting the claim that the most crucial and close communication occurs between the claim drafter and the inventor."). The Patent Office substantively examines a patent application to confirm that it meets the requirements for a patent (including utility, patentable subject matter, novelty, non-obviousness, written description, enablement, and best mode). During examination, the Patent Office considers the precise claim language chosen by the patent applicant. The claims define the outer limit of the patentee's rights. 35 U.S.C. § 271 (2006); Jeanne C. Fromer, *Claiming Intellectual Property*, 76 U. CHI. L. REV. 719, 731 (2009) ("[C]laiming communicates the set to the public to encourage efficient investment in the invention, by requiring licensing or abstinence from the set's embodiment and by permitting free use of embodiments not in the set.").

25. Scripps Clinic & Research Found. v. Genentech, Inc., 927 F.2d 1565, 1580 (Fed. Cir. 1991) ("[T]he construction of claims is simply a way of elaborating the normally terse claim language: in order to understand and explain, but not to change, the scope of the claims.").

26. In litigation, it is common for the parties to disagree on the meaning of a particular word or phrase used in a patent claim. Dan L. Burk & Mark A. Lemley, *Fence Posts or Sign Posts? Rethinking Patent Claim Construction*, 157 U. PA. L. REV. 1743, 1751 (2009) ("[T]here is essentially always a dispute over the meaning of the patent claims.").

27. KIMBERLY A. MOORE, PAUL R. MICHEL & TIMOTHY R. HOLBROOK, PATENT LITIGATION AND STRATEGY 287–311 (3d ed. 2008) (discussing canons of claim construction); *see also* Timothy R. Holbrook, *Substantive Versus Process-Based Formalism in Claim Construction*, 9 LEWIS & CLARK L. REV. 123, 144–45 (2005); Arti K. Rai, *Engaging Facts and Policy: A Multi-Institutional Approach to Patent System Reform*, 103 COLUM. L. REV. 1035, 1047 n.43 (2003) ("Like canons of statutory construction, canons of claim construction assist the court in interpreting language consistently.").

more predictable.[28] If companies know ex ante whether their activities infringe on the rights of another, they can plan accordingly.[29] They can rationally decide whether or not to engage in an activity after evaluating the risks.[30] They can invest in "design around" solutions that add to the storehouse of available technologies. A lack of predictability results in companies not knowing with reasonable certainty whether their activities infringe upon the rights of another,[31] which limits their ability to avoid infringement in the first place.[32] This uncertainty arguably leads to a loss of efficiency, because some companies may avoid making new products altogether to eliminate the risk of liability, or pay damages unnecessarily when they otherwise could have designed to avoid infringement.[33] To alleviate this problem, in the mid-1990s the Federal Circuit focused on making patent law more efficient and predictable; to do so, this Article argues, the Federal Circuit decided to make claim construction more important.[34]

It is interesting that the Federal Circuit focused on claim construction as a means of introducing greater certainty. This may have been because claim construction serves a gatekeeper function in infringement suits: before infringement may be determined, the claim must first be construed.[35] Furthermore, many invalidity defenses require claim construction as a first step. Examples include whether the invention is novel,[36] obvious,[37] or patentable subject matter.[38] As a result, claim construction affects a variety of other issues, especially those decided much later in the case. In other words,

28. *See* Cybor Corp. v. FAS Techs., Inc., 138 F.3d 1448, 1473 (Fed. Cir. 1998) (Rader, J., dissenting in part) ("By removing lay juries from complex technological decisions, these decisions promised to improve the predictability and uniformity of patent law.").

29. *See* JAMES BESSEN & MICHAEL J. MEURER, PATENT FAILURE 46–48 (2008).

30. *See id.* at 6–8.

31. *Id.*

32. *Id.* at 70–72.

33. *Id.*

34. *See* John R. Thomas, *Formalism at the Federal Circuit*, 52 AM. U. L. REV. 771, 774–75 (2003) (arguing that the Federal Circuit has used formalism in an effort to make patent law more certain and predictable).

35. Merck & Co. v. Teva Pharm. USA, Inc., 347 F.3d 1367, 1369 (Fed. Cir. 2003) ("In determination of patent infringement, as the first step the claims are construed; then, the construed claims are compared to the alleged infringing device.").

36. Helifix Ltd. v. Blok-Lok, Ltd., 208 F.3d 1339, 1346 (Fed. Cir. 2000) ("The first step of an anticipation analysis is claim construction.").

37. *E.g.*, Sys. Div., Inc. v. Teknek LLC, 59 F. App'x 333, 338 (Fed. Cir. 2003) (citing Smiths Indus. Med. Sys., Inc. v. Vital Signs, Inc., 183 F.3d 1347, 1353 (Fed. Cir. 1999)) ("The first step in an obviousness analysis is to construe the language of the claims.").

38. *See, e.g.*, *In re* Nuijten, 500 F.3d 1361, 1352 (Fed. Cir. 2007) (construing patent claims before considering whether the claims were patentable subject matter).

a change in the salience of claim construction may impact issues decided after and reliant upon claim construction.[39]

2. Reallocation of the Claim Construction Doctrine

In the early 1990s, claim construction was largely performed by juries.[40] Using jury instructions, the judge would instruct the jury on the tools to determine the proper claim construction.[41] The judge would include information on the canons of claim construction, but the jury would be responsible for applying the canons. The judge still maintained some modicum of control through post-trial motions[42]: if the judge believed that the jury's verdict was unsupported or in substantial error, the judge could always grant judgment as a matter of law.[43]

Figure 2 illustrates the status of claim construction in the early 1990s, prior to *Markman* and *Cybor*.

39. The Federal Circuit initially focused on claim construction rather than the doctrine of equivalents to increase predictability. Claim construction is largely based upon the "four corners" of the patent (and the associated prosecution history). In contrast, the doctrine of equivalents is based on a variety of factors, including the patent, prosecution history, and the operation and structure of the accused device. The court must have believed that it could more easily use claim construction to arrive at definite and foreseeable results.

40. Edmund J. Sease, Markman *Misses the Mark, Miserably*, 2004 J.L. TECH. & POL'Y 99, 101 ("The jury was allowed to hear all the evidence and then decide what the term 'absorbent' meant factually in the context of the invention. This was the state of the Federal Circuit in 1984. . . . Since *Markman* in 1996, juries are no longer allowed to determine the meaning of a patent claim.").

41. *See, e.g.*, U.S. Surgical Corp. v. Ethicon, Inc., 103 F.3d 1554, 1566 (Fed. Cir. 1997) (recounting detailed jury instructions provided to construe a means-plus-function claim limitation).

42. In fact, in the famous *Markman* case, the trial judge entered judgment as a matter of law for the accused infringer after the jury had found for the patentee. Markman v. Westview Instruments, Inc., 52 F.3d 967, 973 (Fed. Cir. 1995) (en banc), *aff'd*, 517 U.S. 370 (1996).

43. FED. R. CIV. P. 50.

Figure 2: Claim Construction in the Early 1990s

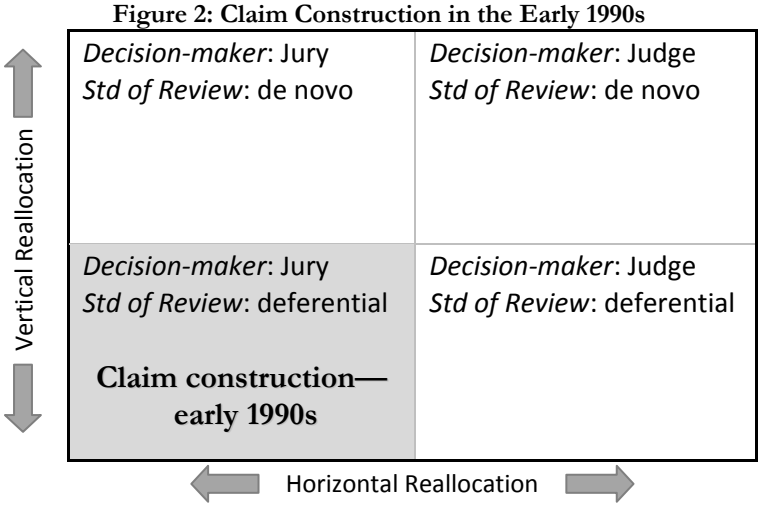


Figure 2 places claim construction in the early 1990s as a doctrine evaluated by the jury and reviewed on appeal with some deference. However, the figure slightly oversimplifies the state of patent litigation, for even in the early 1990s, district court judges interpreted claims in some instances.[44] For example, district court judges granted summary judgment in some patent cases.[45] To decide summary judgment, the court construed the claims and compared the properly construed claims to the accused product or method, granting all inferences to the non-moving party.[46] Judges also construed patent claims when deciding motions for preliminary injunctions, a form of equitable relief reserved only for the court.[47] Moreover, judges conducted bench trials of patent cases without a jury demand, and in doing so construed patent claims.[48] But, in general, in the early 1990s, claim construction was the province of the jury.[49]

44. It should be noted that before the late 1980s, there were fewer jury demands and most patent cases were bench trials. *See* Kimberly A. Moore, *Jury Demands: Who's Asking?*, 17 BERKELEY TECH. L.J. 847, 850–51, 851 fig.1 (2002).

45. *See, e.g.*, Int'l Visual Corp. v. Crown Metal Mfg., 991 F.2d 768 (Fed. Cir. 1993).

46. *See, e.g.*, UTStarcom, Inc. v. Starent Networks, Corp., No. 07-CV-2582, 2009 WL 3122554, at *7 (N.D. Ill. Sept. 22, 2009) ("At the summary judgment stage, the accused device is compared to the construed claims to determine whether there is a genuine issue of material fact.").

47. *See, e.g.*, Devon Indus., Inc. v. Am. Med. Mfg., 19 F.3d 39 (Fed. Cir. 1994); Al-Site Corp. v. Cable Car Sunglasses, 911 F. Supp. 410 (N.D. Cal. 1994).

48. *See, e.g.*, Conopco, Inc. v. May Dep't Stores, 46 F.3d 1556 (Fed. Cir. 1994).

49. Jeffrey Peabody, *Under Construction: Towards a More Deferential Standard of Review in Claim Construction Cases*, 17 FED. CIR. B.J. 505, 506 (2008) ("Prior to 1995, claim construction issues were typically decided by the jury.").

At a recent conference, Federal Circuit Judge Jay Plager revealed some information about the Federal Circuit's thinking about claim construction just before the mid-1990s.[50] According to Judge Plager, the Federal Circuit was concerned about the lack of transparency and the "black box" nature of the jury process, particularly with regards to claim construction. To reverse this problem, the Federal Circuit made major changes in the process of claim construction between 1995 and 1998. None of the changes affected the substantive claim construction doctrine. Instead, the Federal Circuit reallocated the responsibility for claim construction that had rested in juries entirely to judges. The reason for this reallocation can be gleaned from the judicial opinions themselves. The Federal Circuit held that "it is only fair . . . that competitors be able to ascertain to a reasonable degree the scope of the patentee's right to exclude."[51] The Federal Circuit further stated that "competitors should be able to rest assured . . . that a *judge*, trained in the law, will . . . apply the established rules of construction, and in that way arrive at the true and consistent scope of the patent owner's rights to be given legal effect."[52]

Horizontal and vertical reallocation in patent law began in April 1995 when the Federal Circuit issued its en banc decision in *Markman v. Westview Instruments* (*Markman I*).[53] *Markman I* held that claim construction was exclusively reserved for the judge.[54] The majority acknowledged that there were two lines of cases in the Federal Circuit claim construction precedent, one holding that "claim construction may have underlying factual inquiries that must be submitted to a jury,"[55] and the second holding that claim construction "is strictly a question of law for the court."[56] The Federal Circuit found that the first line was incorrect and should be abandoned because it had no firm basis in Federal Circuit precedent.[57] The Federal

50. Judge S. Jay Plager, Comments at the Loyola of Los Angeles Law Review Symposium: The Federal Circuit as an Institution (Oct. 30, 2009).

51. Markman v. Westview Instruments, Inc., 52 F.3d 967, 978 (Fed. Cir. 1995) (en banc), *aff'd*, 517 U.S. 370 (1996).

52. *Id.* at 979 (emphasis added). To be fair, the Federal Circuit is not a single monolithic court. For example, there were three judges who declined to join in the *Markman* majority, Judges Mayer and Rader (concurring) and Judge Newman (dissenting).

53. *Id.*

54. *Id.* at 979 ("We therefore settle inconsistencies in our precedent and hold that in a case tried to a jury, the court has the power and obligation to construe as a matter of law the meaning of language used in the patent claim.").

55. *Id.* at 976.

56. *Id.* at 976–77.

57. *Id.* at 977–78.

Circuit also found that claim construction must be reviewed de novo when raised in an appeal.[58]

The next year, the Supreme Court unanimously affirmed the Federal Circuit's *Markman I* holding, but on somewhat different reasoning (*Markman II*).[59] The Supreme Court supported the Federal Circuit, stating that claim construction is "exclusively within the province of the court."[60] After deciding that no Supreme Court precedent controlled the issue, the Court decided to "consider both the relative interpretive skills of judges and juries and the statutory policies that ought to be furthered by the allocation."[61] According to the Supreme Court, judges are more likely to properly construe a written instrument.[62]

While the Supreme Court in *Markman II* upheld that judges must construe claims (horizontal reallocation), it was silent on the standard of review of claim construction rulings (vertical reallocation). This silence by the Supreme Court led to some short term uncertainty with respect to the standard of review of claim construction. *Markman I* held that it was to be reviewed de novo.[63] *Markman II* was silent on this point.[64] Nonetheless, a majority of Federal Circuit claim construction opinions after *Markman II* found that claim construction was to be reviewed de novo.[65] However, a minority of cases concluded that there was a factual component to claim construction, and those facts were reviewed with deference.[66]

In 1998, several years after *Markman II*, the Federal Circuit en banc decided *Cybor Corp. v. FAS Technologies*, which resolved the standard of review issue.[67] In *Cybor*, the Federal Circuit held that claim construction is purely a matter of law and should be reviewed de novo on appeal. The Federal Circuit stated that most panels after *Markman II* had followed the de novo

58. *Id.* at 979.

59. Markman v. Westview Instruments, Inc. (*Markman I*), 517 U.S. 370 (1996).

60. *Id.* at 372.

61. *Id.* at 384.

62. *Id.* at 388–89.

63. *Markman I*, 52 F.3d at 975.

64. *Markman II*, 517 U.S. 370.

65. *See, e.g.*, Alpex Computer Corp. v. Nintendo Co., 102 F.3d 1214 (Fed. Cir. 1996); Insituform Techs., Inc. v. Cat Contracting, Inc., 99 F.3d 1098 (Fed. Cir. 1996); Gen. Am. Transp. v. Cryo-Trans, Inc., 93 F.3d 766 (Fed. Cir. 1996).

66. *See, e.g.*, Eastman Kodak Co. v. Goodyear Tire & Rubber Co., 114 F.3d 1547, 1555–56 (Fed. Cir. 1997); Serrano v. Telular Corp., 111 F.3d 1578, 1586 (Fed. Cir. 1997); Wiener v. NEC Elecs., Inc., 102 F.3d 534, 539 (Fed. Cir. 1996); Metaulics Sys. Co. v. Cooper, 100 F.3d 938, 939 (Fed. Cir. 1996).

67. Cybor Corp. v. FAS Techs., Inc., 138 F.3d 1448 (Fed. Cir. 1998) (en banc).

standard.[68] The Federal Circuit rejected the clearly erroneous standard.[69] Just as shifting the decision from jury to judge horizontally reallocated control to the judges, shifting the standard of review from deference to de novo vertically reallocated control from the trial courts to the appellate courts.

Patent litigation after *Markman* and *Cybor* is represented in Figure 3.

Figure 3: Claim Construction After *Markman* and *Cybor*

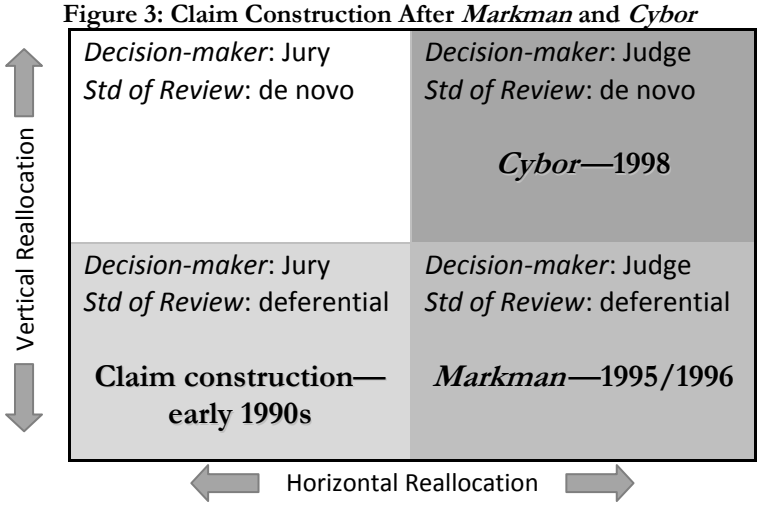


Figure 3 makes apparent the quick change in claim construction control. Within a three-year window, the Federal Circuit moved claim construction from the bottom left box in the matrix—weakest control by the court—to the top right box—greatest control by the court. *Markman I* clearly brought horizontal reallocation of claim construction,[70] moving it from the left box to the right box. A move in this direction gave trial court judges more control of the claim construction doctrine.[71] *Cybor* resulted in vertical reallocation, as is evidenced by moving upward on the matrix as shown in Figure 3.[72] By

68. *Id.* at 1454.

69. *Id.*

70. The Federal Circuit's 1995 en banc ruling reallocated the decision-maker for claim construction. The Supreme Court's decision merely affirmed. Consequently, the 1995 date is used as the date of reallocation in this Article.

71. It also should be noted that Figure 2 is an oversimplification. As noted by the Federal Circuit in *Markman I*, there was a split in authority before 1995. Some cases before *Markman I* had held that claim construction was for the judge. So while all of the post-*Markman* cases fall on the right side of Figure 3, a majority but not all of the pre-*Markman* cases fall on the left side.

72. *Markman I* also brought vertical reallocation to claim construction. However, *Markman II* created some uncertainty on the effectiveness of that reallocation. *See supra* note 66 and accompanying text. *Cybor* confirmed it.

using both horizontal and vertical reallocation, the Federal Circuit seized maximum control over the claim construction doctrine.[73]

B. DOCTRINAL DISPLACEMENT IN PATENT LAW

Doctrinal reallocation generally results in elevating or diminishing the importance of a given doctrine. In turn, after one doctrine increases in importance, one or more other doctrines generally decrease in importance—in other words, they are "displaced" by the doctrine that became more prominent.[74]

1. Causes of Doctrinal Displacement

There are three major causes of doctrinal displacement. First, there is a practical explanation: litigation constraints. Early in the life of a lawsuit, many claims and defenses are raised. After discovery progresses, the parties have an opportunity to explore the merits and strengths of their respective cases. Later, at critical moments in the lawsuit—for example, summary judgment, trial, and appeal—the parties typically focus their claims and defenses. Litigants understand that their chances of success increase if they focus their arguments on a few winnable points.[75] Raising multiple, weaker arguments dilutes the strength of the promising ones.[76] Word and page limits further

73. Figure 3 is a useful illustration of the changes occurring in the mid-1990s. However, the two-by-two matrix has a few limitations. For the purposes of illustration, each time period has been placed in a single box. As noted above, there were some Federal Circuit opinions in the early 1990s that reviewed claim construction de novo. *See, e.g.*, Oscar Mayer Foods Corp. v. ConAgra Inc., 45 F.3d 443 (Fed. Cir. 1994) (unpublished table decision). And after *Markman II* in 1996, most but not all of the Federal Circuit panels considered claim construction using the de novo standard. A very few cases were tried to juries before *Markman I* and decided on appeal after *Markman II*. For example, in *B. Braun Medical, Inc. v. Abbott Laboratories*, the entire case was submitted to a jury in 1994. 124 F.3d 1419 (Fed. Cir. 1997). The case was not decided by the Federal Circuit until 1997. *Id.* Almost all jury claim constructions had been settled or resolved by *Markman II*. Thus, perhaps *Markman* and *Cybor* should be considered together. Using this understanding, most of the cases after *Markman II* should be in the upper right box like those emphasized in Figure 3.

74. The corollary is that as one doctrine becomes less important, one or more other doctrines are usually enhanced.

75. Stephen Easton, *Losing Your Appeal*, 42 FED. LAW. 24, 31 (1995) (directing appellants to "choose the arguments that give you your best chances for success. Force yourself to pare the list to two or three strong grounds for reversal (or affirmance). Concentrate on them in both the briefs and the oral argument").

76. *Id.* at 31 ("An attorney who swamps a judge with every possible argument runs the risk of causing the judge to miss the best arguments. If you throw a diamond into a mud pile, it starts to look pretty muddy.").

push the parties to focus their claims and defenses.[77] Parties to an appeal are capped on the number of words permitted in the documents they submit to the court, and many district courts impose page limits on summary judgment or other important briefs. Litigation constraints are driven by the strategic decisions by the lawyers of which issues to develop and press.[78]

When evaluating which issues to raise, litigants weigh at least two strategic considerations. First, litigants consider the likelihood of success on the merits of a given issue. Second, litigants consider whether the judge or jury will find the issue important in the context of the overall dispute. When one doctrine becomes more important, and consequently is raised more frequently, other doctrines will be raised less frequently. This displacement of other doctrines is normally diffuse. In each case, depending on the particular facts and circumstances, a different doctrine may be displaced. Alternatively, no doctrine may be displaced, and instead litigants may expand the force with which they argue a more limited number of points. Over a series of cases, no single doctrine is displaced directly. In these instances, multiple doctrines are displaced to a lesser degree, and the displacement proceeds largely unnoticed. However, the court can focus the displacement on one or several doctrines through additional case law. En banc decisions of an appellate court (or decisions of the Supreme Court) are particularly useful to displace a single doctrine.

The litigation constraints rationale may be particularly important with claim construction. In patent infringement lawsuits, there are frequently numerous potential claim construction disputes. As claim construction became more likely to be a critical and winnable issue, litigants often devoted several of their limited number of arguments to claim construction disputes. Devoting more to claim construction left far less room for other arguments and doctrines, thereby enhancing the displacement of other doctrines.

A second reason for doctrinal displacement is judicial constraints. District court judges have limited time and resources and, when appropriate, rely on summary judgment to expeditiously resolve lawsuits. When summary judgment is available on several possible grounds, judges frequently select the "cheapest" basis. The cheapest option expends the least judicial time or effort. When one doctrine is increased in importance, the time and effort for the judge to consider the various summary judgment avenues change. In

77. *Id.* ("In almost every appellate case, the length of the briefs, the time for oral argument, and, most importantly, the attention spans of judges with overloaded dockets are all severely limited.").

78. As litigation tactics, they are subject to "selection effects" in that displacement may affect which cases settle and do not result in a written opinion.

patent law, after judges construe claims, they can, for instance, evaluate summary judgment on either non-infringement or invalidity. Non-infringement is typically easier for the judge because there is often a single accused product to evaluate, in contrast with a more complicated analysis of multiple pieces of prior art under, for example, the obviousness doctrine. Furthermore, the judge may already be familiar with the accused product from the claim construction process.[79] Thus, judicial claim construction lowers the adjudicative cost of non-infringement relative to other defenses. For these reasons, non-infringement is commonly the preferred route to dispose of patent cases. Non-infringement and invalidity are not substitute doctrines. However, they are substitute methods of resolving cases. Moreover, even outside of the summary judgment context, judges may strategically choose which doctrine to use to dispose of the case. In situations in which the standard of review differs for the two doctrines, district court judges may rely upon the doctrine afforded more deference on appeal to reduce the risk of reversal.[80]

Other academic literature supports the view that courts behave strategically in response to systemic changes to patent litigation. Professors Matthew Henry and John Turner studied the impact of the creation of the Federal Circuit on patent litigation.[81] They conducted a time-series analysis of decisions from 1953 until 2002.[82] According to Henry and Turner, the Federal Circuit, which was created in 1982, was substantially less likely to find a patent invalid than its predecessor courts.[83] If a district court judge found a patent invalid, the Federal Circuit was more likely to reverse than the regional circuit courts had been. Henry and Turner assert that, because the tendency

79. There is some case law stating that the accused product is legally irrelevant to claim construction. *See, e.g.*, SRI Int'l v. Matsushita Elec. Corp., 775 F.2d 1107, 1118 (Fed. Cir. 1985) (en banc) (stating that a claim is not to be construed in light of the accused device). Other cases expressly permit viewing the accused product to provide further context for the claim construction analysis. *See, e.g.*, Wilson Sporting Goods Co. v. Hillerich & Bradsby Co., 442 F.3d 1322, 1326 (Fed. Cir. 2006).

80. Max M. Schanzenbach & Emerson H. Tiller, *Strategic Judging Under the United States Sentencing Guidelines: Positive Political Theory and Evidence*, 23 J.L. ECON. & ORG. 24 (2007) (finding evidence that trial judges choose the bases for their sentencing decisions to protect their decision from higher court review); Joseph L. Smith & Emerson H. Tiller, *The Strategy of Judging: Evidence from Administrative Law*, 31 J. LEGAL STUD. 61 (2002) (finding support that strategic considerations influence trial courts when reviewing decision from the Environmental Protection Agency).

81. Matthew D. Henry & John L. Turner, *The Court of Appeals for the Federal Circuit's Impact on Patent Litigation*, 35 J. LEGAL STUD. 85, 85 (2006).

82. *Id.* at 88–89.

83. *Id.* at 90.

to reverse invalidity findings was well-known, district courts more frequently relied on non-infringement to resolve cases after 1982.[84]

There is a third explanation for doctrinal displacement: the judge's role as a gatekeeper. In some instances, such as in claim construction after *Markman*, the judge must decide one doctrine before reaching others. These downstream doctrines can include doctrines relating to liability and damages. If the judge decides the first doctrine in a manner that resolves the dispute, the second doctrine need not be reached. For instance, assume that the decision-maker for the duty of reasonable care requirement in tort law was shifted from the jury to the judge. Thereafter, judges if they so desired could dispose of negligence actions without a jury trial on the basis that no duty was owed to the defendant. In this example, the doctrines of negligence and damages could be displaced due to the judge's gatekeeping role.

The gatekeeper theory permits extreme displacement of a doctrine. The displaced doctrine does not merely fall incrementally in the hierarchy of doctrines. A slight decrease would be consistent with the litigation constraints theory alone; rather, the gatekeeper theory adds that the doctrine drops substantially. A reallocation can drop another previously important doctrine below numerous other unaffected doctrines. And substitute doctrines that address the same equity concerns are strong candidates for displacement. When one doctrine becomes more prominent, the court can use it extensively. Substitute doctrines are not needed as much, and subsequently diminish in stature.

Doctrinal displacement may also be enhanced with the rise in summary judgment of another doctrine. If summary judgment is granted for the defendant on a particular defense, no jury trial may be necessary in the case. Doctrines which would have been evaluated by the jury become moot. Thus, increasing the grants of summary judgment has the effect of displacing doctrines typically considered by the jury downstream of the summary judgment decision.

The litigation constraints, judicial constraints, and gatekeeper explanations for doctrinal displacement are interrelated. When litigants know that the court acts as a gatekeeper before a doctrine is reached, they are more likely to downplay it. Instead, litigants focus their primary efforts on the gatekeeping doctrine. Similarly, when litigants understand that courts prefer to grant summary judgment on a particular basis, they are likely to file motions on that basis more often. Consequently, litigation realities amplify

84. *Id.* at 103.

the effect of the gatekeeper. As such, all three theories support the same result—the displacement of one doctrine as another gains importance.[85]

2. The Displaced Doctrine: The Doctrine of Equivalents

The doctrine of equivalents permits a finding of infringement even if the accused device or method does not literally fall within the scope of the construed patent claims.[86] Instead, a device or method may infringe under the doctrine of equivalents if it performs "substantially the same function in substantially the same way to obtain the same result" as the patented invention.[87] Thus, the doctrine of equivalents permits an expansion of patent rights beyond the literal scope of the patent claims. One purpose of the doctrine of equivalents is to protect patentees from those who seek "to evade liability for infringement by making only insubstantial changes to a patented invention."[88] The Supreme Court explained that without the doctrine of equivalents, a patent would be "a hollow and useless thing" and "unscrupulous copyist[s]" would be "encourage[d]."[89] At the onset of litigation, most patentees allege infringement both literally and under the doctrine of equivalents, with the latter being a fallback position.[90]

The Federal Circuit endorsed the doctrine of equivalents in 1995, nearly simultaneously with the claim construction reallocation resulting from *Markman*. In *Hilton Davis Chemical Co. v. Warner-Jenkinson Co.*, the Federal Circuit, sitting en banc, affirmed a jury verdict of infringement under the doctrine of equivalents.[91] In a 7–5 ruling, the majority held that an exception to the doctrine of equivalents known as prosecution history estoppel did not apply.[92] The Federal Circuit also declined to reallocate the doctrine of equivalents to the judge, instead holding that the doctrine of equivalents was

85. An important question is whether displacement is intentional or an unintended consequence. This Article cannot answer that question with any certainty. However, the fact that no court has acknowledged displacement leads me to believe it is unintentional.

86. Warner-Jenkinson Co. v. Hilton Davis Chem. Co., 520 U.S. 17, 39–40 (1997).

87. Graver Tank Mfg. Co. v. Linde Air Prods., Co., 339 U.S. 605, 608 (1950) (quoting Sanitary Refrigerator Co. v. Winters, 280 U.S. 30, 42 (1992)). An accepted alternative test for the doctrine of equivalents is whether there are insubstantial differences between the accused device or method and the claimed invention. *Warner-Jenkinson*, 520 U.S. at 24. Another relevant consideration is the known interchangeability of the elements.

88. Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 535 U.S. 722, 727 (2002).

89. *Graver Tank*, 339 U.S. at 607.

90. Allison et al., *supra* note 2, at 977. ("Rather, a patentee is almost always arguing the doctrine of equivalents as an alternative to a theory of literal infringement.").

91. Hilton Davis Chem. Co. v. Warner-Jenkinson Co., 62 F.3d 1512, 1528–29 (Fed. Cir. 1995) (en banc), *rev'd*, 520 U.S. 17 (1997).

92. *Id.* at 1514.

a jury question.[93] In the years that followed, first the Supreme Court and then the Federal Circuit began limiting the reach of the doctrine of equivalents.

In the late 1990s and early 2000s, the Federal Circuit and Supreme Court issued several significant decisions involving the doctrine of equivalents.[94] These cases, in turn, placed legal limits—administered by judges—on when the doctrine of equivalents is applicable. First, in 1997, the Supreme Court decided *Warner-Jenkinson v. Hilton Davis*, which reversed the Federal Circuit's en banc decision.[95] In *Warner-Jenkinson*, the Supreme Court held that the All Element Rules must be employed,[96] meaning that each claim element must be present in the accused device or method either literally or equivalently.[97] Prior to this Supreme Court opinion, the Federal Circuit had not consistently and stringently applied the All Elements Rule, and sometimes it permitted patentees to loosely argue that the accused product was equivalent to the invention as a whole.[98]

Thereafter, in 2000, the Federal Circuit en banc voted in *Festo v. Shoketsu Kinzoku Kogyo Kabushiki Co.* to bar the application of the doctrine of equivalents for certain claim elements.[99] The Federal Circuit found that the doctrine of equivalents was not available if a claim element had been amended during the process of examination by the Patent Office.[100]

---

93. *Id.* at 1522 ("In answer to the second question posed by this court *en banc*, infringement under the doctrine of equivalents is an issue of fact to be submitted to the jury in a jury trial with proper instructions, and to be decided by the judge in a bench trial.").

94. For a good discussion of the development of the doctrine of equivalents during this time frame, see Timothy R. Holbrook, *Possession in Patent Law*, 59 SMU L. REV. 123, 164–69 (2005); Michael Meurer & Craig A. Nard, *Invention, Refinement and Patent Claim Scope: A New Perspective on the Doctrine of Equivalents*, 93 GEO. L.J. 1947 (2005); Petherbridge, *Doctrine of Equivalents*, *supra* note 2, at 1385–1393.

95. Warner-Jenkinson Co. v. Hilton Davis Chem. Co., 520 U.S. 17 (1997).

96. *Id.* at 40 ("The determination of equivalence should be applied as an objective inquiry on an element-by-element basis.").

97. *Id.* at 21 ("Nearly 50 years ago, this Court in [*Graver Tank*] . . . set out the modern contours of what is known in patent law as the 'doctrine of equivalents.' Under this doctrine, a product or process that does not literally infringe upon the express terms of a patent claim may nonetheless be found to infringe if there is 'equivalence' between the elements of the accused product or process and the claimed elements of the patented invention.").

98. *Cf.* Joseph R. Re & Lynda J. Zadra-Symes, *Infringement Under the Doctrine of Equivalents: The Federal Circuit's First Ten Years*, 785 ALI-ABA PAT. L. & LITIG. 77, 93 (1992) ("The Federal Circuit has applied the element-by-element approach it adopted in banc in *Pennwalt* [in 1987] almost consistently since that decision.").

99. Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 234 F.3d 558 (Fed. Cir. 2000) (en banc), *vacated*, 535 U.S. 722 (2002).

100. *Id.* at 563 ("Therefore, an amendment that narrows the scope of a claim for any reason related to the statutory requirements for a patent will give rise to prosecution history estoppel with respect to the amended claim element.").

Specifically, if the patentee had amended a particular element for "reasons relating to patentability"—such as to convince the Patent Office that the claim was new—then no equivalence was available for such element.[101] After granting certiorari, the Supreme Court in 2002 rejected the Federal Circuit's categorical test.[102] Instead, the Supreme Court held that a rebuttable presumption applies to any claim amended for reasons relating to patentability.[103] However, the Supreme Court's flexible and rebuttable presumption approach has been very difficult for patentees to overcome.[104] Thus, while the Supreme Court rejected the Federal Circuit's rigid test, the effect of these rulings has been a substantial reduction in instances in which the doctrine of equivalents is applicable.[105]

In sum, the doctrine of equivalents, a fairness doctrine juries apply, is arguably inconsistent with a patent system premised on predictability and on clear prior notice of the scope of rights.[106] In the years since *Markman*, there have been several Federal Circuit and Supreme Court cases touching on the doctrine of equivalents. These cases introduced substantive restrictions on

101. *Id.* at 569 ("When a claim amendment creates prosecution history estoppel with regard to a claim element, there is no range of equivalents available for the amended claim element. Application of the doctrine of equivalents to the claim element is completely barred (a 'complete bar').").

102. Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 535 U.S. 722, 723–25 (2002).

103. *Id.* at 741 ("When the patentee has chosen to narrow a claim, courts may presume the amended text was composed with awareness of this rule and that the territory surrendered is not an equivalent of the territory claimed. In those instances, however, the patentee still might rebut the presumption that estoppel bars a claim of equivalence. The patentee must show that at the time of the amendment one skilled in the art could not reasonably be expected to have drafted a claim that would have literally encompassed the alleged equivalent.").

104. Christopher A. Harkins, *Choosing Between the Advice of Counsel Defense to Willful Patent Infringement or the Effective Assistance of Trial Counsel: A Bridge or the Troubled Waters?*, 5 NW. J. TECH. & INTELL. PROP. 210, 218 (2007) ("[W]hen a claim amendment is an amendment related to patentability, there arises a presumption of estoppel against the doctrine of equivalents, which presumption may only be overcome in a few 'narrow ways.' ").

105. Daniel H. Shulman & Donald W. Rupert, *"Vitiating" the Doctrine of Equivalents: A New Patent Law Doctrine*, 12 FED. CIR. B.J. 457, 459 (2003) ("These other decisions, which have become more and more frequent in the last few years, limit the DOE by effectively creating a *per se* rule as to what constitutes an equivalent."). In 2001, the Federal Circuit continued narrowing the reach of the doctrine of equivalents. The Federal Circuit en banc eliminated the doctrine of equivalents for a different type of claim element—no equivalents are available for subject matter disclosed in a patent specification but not literally claimed. Johnson & Johnston Assocs. v. R.E. Serv. Co., 285 F.3d 1046 (2002) (en banc).

106. Timothy R. Holbrook, *Equivalency and Patent Law's Possession Paradox*, 23 HARV. J.L. & TECH. 1, 5 (2010) ("The express purpose of [the doctrine of equivalents] is to ensure fair and adequate protection to the patentee and to solidify the patent incentive."); *cf.* Meurer & Nard, *supra* note 94, at 1953–54.

the doctrine of equivalents.[107] However, the displacement of the doctrine of equivalents, which led to its decreasing importance, occurred after *Markman I*, well before any direct assaults on the doctrine in these cases. For the most part, these cases reduced the occasions on which a patentee may successfully raise infringement under the doctrine of equivalents.[108] But the displacement had already occurred. While they surely further diminished the stature of the doctrine, they occurred subsequent to the doctrinal displacement.

Other scholars have empirically studied the success of patentees who attempt to rely upon the doctrine of equivalents, and their data supports the theory that the doctrine of equivalents has diminished in several respects. Petherbridge found that by 2007, patentees only rarely succeeded under the doctrine of equivalents.[109] Rather than tying the decrease in success to the *Markman* decision, Petherbridge traces it to the *Festo* decision in 2000, which provided strong limits on the availability of the doctrine of equivalents.[110] Petherbridge also posited that the decline in the doctrine of equivalents is directly related to the rise in power of claim construction.[111] As the Federal Circuit increased its rate of modifying lower court claim constructions, it decreased the rate of success of a patentee on the doctrine of equivalents.[112] However, Petherbridge did not provide an explanation of why the increased importance of claim construction decreased the success rate on the doctrine of equivalents.[113] Petherbridge, instead, argues that an intra-circuit dispute on

107. John R. Thomas, *Claim Re-construction: The Doctrine of Equivalents in the Post-*Markman *Era*, 9 LEWIS & CLARK L. REV. 153, 159 (2005).

108. *See* Fromer, *supra* note 24, at 735–38; Nicole S. Robbins, *The Curtailment of the Doctrine of Equivalents: Courts Emphasize the Public Notice Function of Patent Claims*, 35 SUFFOLK L. REV. 323, 339–43 (2001).

109. Petherbridge, *Doctrine of Equivalents*, *supra* note 2, at 1386–92. From reviewing the figures in Petherbridge's articles, it appears that the decline did not commence until at least *Markman I*.

110. *Id.* at 1390–92 (reporting that logistical regression predicts that changes in procedural circumstances in the doctrine of equivalents largely explain the decline).

111. Petherbridge, *Claim Construction Effect*, *supra* note 2, at 236.

112. *Id.* ("[R]egression analysis provides evidence that the Federal Circuit's rejection of lower court claim construction determinations most strongly predicts a decrease in predictability, while other variables that could have explained it, like changes in the rules surrounding the doctrine of equivalents have either no impact or predict predictability rather than unpredictability.").

113. In fact, Petherbridge argues that the procedural changes relating to claim construction did not cause the decline in the doctrine of equivalents. *See* Petherbridge, *Claim Construction Effect*, *supra* note 2, at 244 ("Procedural changes (e.g., increases in relative rates of incoming summary judgments potentially wrought by the *Markman*/*Cybor* framework) do not provide a strong explanation for the decline in doctrinal stability.").

claim construction methodology appeared around 2000, which affected the doctrine of equivalents.[114]

Allison and Lemley draw somewhat different conclusions from their empirical study of outcomes of doctrine of equivalents cases.[115] They studied district court and Federal Circuit doctrine of equivalents decisions in three periods surrounding the *Festo* decision.[116] Allison and Lemley assert that the multiple substantive changes in the doctrine of equivalents law had "surprisingly little effect on the actual outcome of doctrine of equivalents cases."[117] More significant to the present Article, Allison and Lemley state that by the late 1990s, patentee assertions under the doctrine of equivalents almost never prevailed at trial or on appeal.[118]

Allison and Lemley argue that *Markman* killed the doctrine of equivalents.[119] They speculate that after *Markman*, district court judges were inclined to err on the side of granting summary judgment of non-infringement under the doctrine of equivalents.[120] There was rarely a dispute about the structure or function of the accused product. So after the judge construed the claims, summary judgment of literal infringement or non-infringement was often appropriate.[121] But granting summary judgment of no literal infringement would not resolve the entire lawsuit. To end the lawsuit, the court had to consider the patentee's charge of infringement under the doctrine of equivalents.[122] Allison and Lemley argue that judges "will be doubly inclined to hold for the accused infringer" on the doctrine of equivalents as the only way to dispose of the case.[123] Allison and Lemley's explanation for the decline of the doctrine of equivalents is similar to the gatekeeper theory for displacement, and their empirical results are consistent with doctrinal displacement. However, their explanation is incomplete as they focus only on the trial court, not the appellate court, as a gatekeeper. They also do not fully articulate the litigation and judicial constraints aspects of displacement.

114. *Id.* at 245 ("The evidence to this point suggests that [the change in the realm of claim construction] more strongly involves changes in claim construction jurisprudence than it does changes in the rules of the doctrine of equivalents.").

115. Allison & Lemley, *supra* note 2.

116. *Id.* at 963–66.

117. *Id.* at 957.

118. *See id.* at 970–71.

119. *See id.* at 977–78.

120. *See id.* at 977.

121. *Id.*

122. *Id.*

123. *Id.*

So why was the doctrine of equivalents displaced, and not some other doctrine? One explanation is that claim construction and the doctrine of equivalents serve similar functions—both are directed to the scope of protection for a patentee. Claim construction provides the literal reach of the patent. The doctrine of equivalents permits the patentee a further reach, as long as the differences between the literal claim scope and the accused product are insubstantial. When construing claims, the judge often knows the structure of the accused products. Using this knowledge, the judge may provide a broader construction to ambiguous claim language so as to avoid confronting the doctrine of equivalents. In other words, courts may have found these doctrines to be substitutes for each other. And as previously noted, substitute doctrines are strong candidates for displacement.[124]

A related explanation is that claim construction has arguably expanded to encompass the doctrine of equivalents. The doctrine of equivalents permits the patentee to cover items that perform substantially the same function although they have different structure. For instance, a stent with an oval cross-section might be found equivalent to a claimed one with a "circular" cross-section. After *Markman I*, claim construction is performed solely by judges. Juries did not need to provide a written record of their claim construction. Judges, when forced to do this, regularly defined the claim terms—structural terms—using functional definitions.[125] Post *Markman I*, the judicial craft of claim construction has subsumed the doctrine of equivalents. Consequently, the need for the doctrine of equivalents was effectively eliminated.

As outlined above, Allison and Lemley provide a different reason for the displacement of the doctrine of equivalents. After the judge steeped herself in the technology and construed the claims, she was less inclined to submit the case to a jury.[126] However, to fully resolve the case, the doctrine of equivalents needed to be decided.[127] Thus, as a practical matter, judges quickly decided the doctrine of equivalents under the guise of summary judgment to keep the case from the jury.[128] If this is true, one would expect less success on the doctrine of equivalents after *Markman* hearings became important.[129] While *Markman* mandated that district court judges construe the

124. *See supra* Section II.B.1.

125. An exception to this is organic chemistry, a field in which structure can be defined by structure alone.

126. *See* Allison & Lemley, *supra* note 2, at 977.

127. *Id.*

128. *See id.*

129. *Id.* at 977–98.

patent claims, it left open when and how judges do so. Some courts adopted local patent rules that specified the timing of claim construction.[130] Many judges elected to hold separate hearings, often called *Markman* hearings.[131] These hearings solely focused on the meaning of the claims, and they are typically divorced from consideration of the issues of infringement, validity, or enforceability.[132] *Markman* hearings could last up to several weeks and sometimes included live witness testimony. Other judges elected to construe the patent claims simultaneously with deciding dispositive motions. These judges often did not hold a *Markman* hearing but, instead, decided the issue of claim construction based on the written record developed during summary judgment briefing. The Federal Circuit has taken no position on the timing and procedure used by district courts to construe claims, and it has approved of both major approaches.[133] The sole Federal Circuit mandate was that claim construction must be performed by the court, not the jury.[134] Judges who held separate hearings may have been more likely to learn the technology and have a greater desire to dispose of the case in its entirety after claim construction. Because the substantive changes to the doctrine of equivalents in *Festo* are so close in time to the rise of separate hearings, it would be difficult to directly test whether the hearings are correlated with the displacement.

III. STUDY DESIGN AND METHODOLOGY

The findings of this Article are based upon data derived from three databases: (1) a claim construction appellate decision database consisting of information from all published and unpublished claim construction decisions from 1991 until 2008; (2) an appellate issue database consisting of information for all electronically available Federal Circuit decisions from the

130. *See, e.g.*, N.D. CAL., P.R. 4-5, 4-6 (2010); E.D. TEX., P.R. 4-5, 4-6 (2010); N.D. GA., P.R. 6.5, 6.6 (2009); W.D. PA., P.R. 4 (2009); S.D. TEX., P.R. 4.5, 4.6 (2008); E.D. N.C., P.R. 304.5, 304.6 (2010).

131. William F. Lee & Anita K. Krug, *Still Adjusting to* Markman*: A Prescription for the Timing of Claim Construction Hearings*, 13 HARV. J.L. & TECH. 55, 59 (1999).

132. *See* Sease, *supra* note 40, at 99.

133. The Federal Circuit has even authorized the court to conduct rolling claim construction, revising an initial claim construction after the record was more fully developed. *See, e.g.*, SanDisk Corp. v. Memorex Prods., Inc., 415 F.3d 1278, 1291 (Fed. Cir. 2005) ("After discovery the court expects the parties to refine the disputed issues and learn more about the claim terms and technology, at which point a more accurate claim construction can be attempted.").

134. Lee & Krug, *supra* note 131, at 56–57 (noting that *Markman* did not proscribe any particular timing to claim construction, and further noting that district courts have construed patent claims as early as the onset of litigation and as late as prior to jury instruction).

years 1991, 1994, 1997, and 2000; and (3) a word count appellate database consisting of published appeals following grants of summary judgment from 1987 until 2004. The appellate issue and word count appellate databases were created for this Article. This Part describes the databases and the process for constructing them. It then discusses limitations of the study including limitations of empirical legal studies of appellate court decisions more generally.

A. THE DATABASES

1. The Claim Construction Appellate Decision Database

The claim construction appellate decision database includes all claim construction appellate decisions from district court litigation[135] from January 1, 1991 until December 31, 2008.[136] Overall, the database contains 1,288 Federal Circuit decisions, including 157 decisions before *Markman I*.[137] Appeals are included regardless of the procedural posture—whether resolved via preliminary injunction, summary judgment, trial, or otherwise. The database only includes appeals from district courts. Accordingly, it does not include appeals from the United States Patent & Trademark Office, the Court of Federal Claims, or the International Trade Commission.

The claim construction appellate decision database includes all merits resolutions by the Federal Circuit of claim construction appeals from district court litigation. The Federal Circuit can resolve appeals involving claim construction through several mechanisms including a precedential written opinion, a non-precedential written opinion, and a summary affirmance. The database includes precedential and non-precedential opinions, as well as appeals resolved without a written opinion. A detailed explanation of how

135. The dataset only includes utility patents. Appellate review of claim construction of design patents (and plant patents) is relatively infrequent. Because claim construction of design patents is substantively different from claim construction of utility patents, design (and plant) patents were excluded from the present study.

136. The claim construction appellate decision database has been used in several previous studies. *See* David L. Schwartz, *Practice Makes Perfect? An Empirical Study of Claim Construction Reversal Rates in Patent Cases*, 107 MICH. L. REV. 223, 237–41 (2008) [hereinafter Schwartz, *Practice Makes Perfect*]; David L. Schwartz, *Courting Specialization: An Empirical Study of Claim Construction Comparing Patent Litigation Before Federal District Courts and the International Trade Commission*, 50 WM. & MARY L. REV. 1699 (2009); David L. Schwartz, *Pre*-Markman *Reversal Rates*, 43 LOY. L.A. L. REV. 1073 (2010) [hereinafter Schwartz, *Pre*-Markman *Reversal Rates*].

137. The reliability and validity of the original database is high. *See* Schwartz, *Practice Makes Perfect*, *supra* note 136, at 272–73.

the original dataset was derived is available elsewhere[138] and consequently, not repeated here.

2. The Appellate Issue Database

The appellate issue database identifies the issues discussed in Federal Circuit decisions from the years 1991, 1994, 1997, and 2000. The years were selected to include equally spaced samples, before and after *Markman I* and *II*.

The database records specific issues explored in all electronically-available Federal Circuit opinions, both precedential and non-precedential, during those years. The issues are generally identified by headings in the opinions such as anticipation, obviousness, inequitable conduct, and infringement. Most cases involved multiple issues. Claim construction may be analyzed as a precursor to any of these issues. Even if the Federal Circuit's analysis did not use a heading, if the opinion discussed the issue, it was included in the database. For claim construction and the doctrine of equivalents, an opinion was not recorded unless there was a specific discussion in the opinion analyzing the relevant law or facts.[139] A bare bones recitation in an opinion, such as one which notes that infringement was not shown literally or under the doctrine of equivalents or that the court must construe the claims, was not recorded.[140]

To develop the appellate issue database, a Lexis query was performed to locate potentially relevant cases.[141] A human coder recorded the issues

138. For a thorough discussion of the selection, coding, and reliability of the dataset, see Schwartz, *Practice Makes Perfect*, *supra* note 136, at 269–74. For a discussion of particular issues with locating earlier (pre-1995) cases, see Schwartz, *Pre*-Markman *Reversal Rates*, *supra* note 136, at 1073, 1091–92.

139. Technically, claim construction is a doctrine, not an issue. *See* Tiller & Cross, *supra* note 18 (discussing the distinction between doctrine and issue.) As for the doctrine of equivalents, the Supreme Court itself uses the phrase "*doctrine* of equivalents." In coding the appellate issue database, claim construction and the doctrine of equivalents were included, in addition to the general issues of §§ 101, 102, 103, and 112; literal infringement; and inequitable conduct.

140. The Federal Circuit in some cases noted that an issue was briefed but would not be decided. *See, e.g.*, Purdue Pharma, L.P. v. Faulding, Inc., 230 F.3d 1320, 1329–30 (Fed. Cir. 2000) ("Because we have upheld the district court's determination that the asserted claims of the '360 patent are invalid, it is unnecessary to address Faulding's cross-appeal from the district court's finding of infringement."). These undecided issues were excluded from the database.

141. In the CAFC database, the following query was executed: "court and date(geq (1/1/1991) and leq (12/31/1991)) and not name(trademark or department or secretary or "international trade" or "merit systems" or veteran or "federal claims" or "in re") and not("patent appeals and interferences" or "united states claims court")." Because summary affirmances did not discuss any issues, they were not responsive to the Lexis search.

discussed by the Federal Circuit in every case. Opinions from merits decisions of district courts were included, regardless of whether the district court resolved the case via preliminary injunction, summary judgment, bench trial, or jury trial.[142] The appellate issue database includes 297 total opinions, including 54 opinions from 1991; 55 from 1994; 90 from 1997; and 98 from 2000.

3. The Word Count Appellate Database

The word count appellate database includes specific word count information about certain appellate decisions from 1987 until 2004.[143] The database notes the number of words in each opinion devoted to claim construction and the number of words devoted to the doctrine of equivalents. Words devoted to the doctrine of equivalents includes discussion about any test or argument relating to the doctrine of equivalents including prosecution history estoppel limitation. The database also records the total word count discussing all issues on appeal, typically organized in the opinion under the heading "Discussion." Most opinions in the database have separate headings under which the patent and claims were introduced. Consequently, the claim construction word count typically does not include a recitation of the claim as a whole.

To attempt to keep as much as possible constant across the 1987 until 2004 time period, only appeals reviewing a grant of summary judgment are included. No appeals from jury or bench trials are included.[144] Thus, the word count appellate database excludes decisions in which the jury construed the claims or the jury decided the doctrine of equivalents. If all opinions were included, including those reviewing jury verdicts, we may expect a change in word count due to the doctrinal reallocation of claim construction alone. This is because there would be a larger appellate record after the judge construed the claims. In other words, after *Markman*, the record in a jury case would include a judicial claim construction. In contrast, the record from jury trials before *Markman* frequently did not, as the judge did not expressly

142. Only appeals from utility patent litigation at the district court were included. Non-merit appeals, such as from motions to dismiss for lack of personal jurisdiction, were omitted. Cases in which liability and attorney fees generated separate appeals were only included in the dataset for the liability appeal.

143. The beginning year of 1987 was selected because jury demands in patent cases were less frequent before the late 1980s. *See* Moore, *supra* note 44, at 851 fig.1.

144. In some cases, summary judgment is granted in part, such as for literal infringement, and denied in part. In these cases, a trial is conducted on the doctrine of equivalents. This sort of case was excluded. Only cases in which claim construction and the doctrine of equivalents was resolved on summary judgment were included.

construe the claims. Because there is more to review, we may expect more words in the appellate opinions. There is not a similar problem in summary judgment appeals. In summary judgment cases before and after *Markman*, the judge construed the claims. By using summary judgment cases only, this concern is substantially alleviated. However, the study assumes that the district courts would grant summary judgment in the same cases before and after *Markman*.

The database is limited to precedential opinions because non-precedential opinions typically are not as well organized. They often lack the organization present in precedential opinions, such as a separate "Discussion" section. This presents potential coding difficulties. Furthermore, the non-precedential opinions often are very cursory, especially relative to precedential opinions. Because the opinions are short, the use or omission of introductory sentences to either the claim construction or doctrine of equivalents discussion could materially alter the results. For this reason, a word count of non-precedential opinions was not deemed to be as useful.

To develop the word count appellate database, an overbroad query was performed on Lexis to locate potentially relevant cases.[145] Thereafter, a human coder read every case[146] to confirm that all of the following were true: (1) the appeal was from a decision of a federal district court; (2) the federal district court resolved the lawsuit on summary judgment; and (3) the appeal addressed either claim construction or the doctrine of equivalents, or both. The coder noted the number of words devoted to each issue, the number of words in the "Discussion" section which typically included a discussion of all issues on appeal, as well as the number of words in the entire opinion. The "Discussion" section did not include background information about the patent, technology, or procedural posture of the case. The recitation of the claim itself was not counted as part of the claim construction analysis. Words in an alternative opinion, such as concurring or dissenting opinions, were not counted. The word count appellate database includes word count information from 183 opinions.

145. In the CAFC database, the following query was executed: "claim w/10 (construl or interp!) and (doctrine w/5 equivalents) and "summary judgment" and date(geq (1/1/1987) and leq (12/31/2004))."

146. Several significant en banc decisions were omitted to avoid skewing the sample. These included *Markman*, *Warner-Jenkinson*, *Cybor*, *Festo*, and *Johnson & Johnston*. The appeal in each of these cases only involved one of the doctrines, and the word count of the discussion in each was substantial.

B. LIMITATIONS OF THE DATABASES AND EMPIRICAL STUDIES OF APPELLATE DECISIONS

All projects involving empirical studies of legal decisions have limitations and the present study is no exception. First, patent litigation is extremely complex. Typically, there are numerous issues raised by the parties. These issues are often fact-specific for each case. For example, patent litigation between branded and generic drug manufacturers differs from patent litigation over a business method patent held by a non-practicing entity. Not only is the underlying technology different in these scenarios, but the parties' strategic goals vary as well. Consequently, it is difficult to make generalizations about patent litigation from the study of individual cases.

Second, the present Article uses data gathered through content analysis of judicial opinions, which has well-known limitations.[147] These include unobserved reasoning, strategic behavior, and selection bias.[148] Judge Harry Edwards argues that empirical methods are not useful to understanding judicial decision-making.[149] He argues that statistics cannot distinguish among extralegal factors that affect judicial decision-making.[150] These unobserved factors include the state of the case record on appeal and the judicial deliberations that preceded the opinion.[151] He also argues that most empirical legal studies of case law lack firm support because they exclude summary affirmances. In the present study, the claim construction appellate decision database includes summary affirmances; however, the word count appellate database and appellate issue database do not. By definition, there are no words to count in a summary affirmance. For the appellate issue database and the related analysis, the study assumes that the issues raised in summary affirmances and opinions are the same, an assumption which may not be correct.

147. Mark A. Hall & Ronald F. Wright, *Systematic Content Analysis of Judicial Opinions*, 96 CALIF. L. REV. 63, 105–06 (2008) (discussing the limitations of content analysis as part of their call for greater use of content analysis).

148. For a discussion of these limitations, see R. Polk Wagner & Lee Petherbridge, *Is the Federal Circuit Succeeding? An Empirical Assessment of Judicial Performance*, 152 U. PA. L. REV. 1105, 1128–29 (2004).

149. Harry T. Edwards & Michael A. Livermore, *Pitfalls of Empirical Studies That Attempt To Understand the Factors Affecting Appellate Decisionmaking*, 58 DUKE L.J. 1895 (2009).

150. *Id.* at 1899 ("Legal scholars remain interested in trying to use empirical methods—most notably the statistical analysis of case outcomes—to understand the effect of extralegal factors on appellate decisionmaking. In our view, the principal problem with such empirical legal analyses is that they cannot distinguish between legal and extralegal factors without considering and accurately accounting for the most important determinants of appellate decisionmaking: (1) the case records on appeal, (2) the applicable law, (3) controlling precedent, and (4) judicial deliberations.").

151. *Id.*

Separately, patent law changed in many ways in the last twenty years. The Supreme Court and Federal Circuit issued numerous substantive decisions that altered the law, only some of which are described by this Article. Changes to one doctrine may cause substantive effects on the law in other doctrines. Furthermore, Congress also amended the Patent Act several times during the time period of this study. These changes include adjusting how to calculate the patent term[152] and requiring publication of most patent applications before issuance.[153]

Each of these changes may affect patent litigant strategies and substantive patent law doctrines. Because patent litigation as a whole is so complex, it is incredibly complicated to develop and test empirical models. This complexity is especially prevalent when multiple doctrines in patent law are interrelated and studied simultaneously. Changes in precedent can alter lawyers' behavior in drafting patents. Furthermore, changes in precedent can also influence party behavior in litigation. Thus, the patent litigation system is dynamic and, over time, the types of lawsuits brought will change.

Another limitation is that the changes studied involving claim construction and the doctrine of equivalents are endogenous to the Federal Circuit. In other words, while this Article termed *Markman* and *Cybor* as the cause of doctrinal reallocation and displacement, the court itself made these changes. As the Federal Circuit made the *Markman I* and *Cybor* decisions, as well as the subsequent opinions studied in this Article, the events are not truly independent. No empirical methodology can correct for this. Furthermore, this Article does not differentiate among the various judges on the Federal Circuit; rather, the Federal Circuit is treated as a single static court. While the data are largely consistent with the propounded hypotheses, this Article makes no claims regarding causation. To the extent it makes any assertions, it is limited to mere correlation. The explanations for the correlations deserve further empirical and theoretical scrutiny.

Another limitation stems from general changes in litigation over the studied time period. Even outside of patent lawsuits, litigation in general has increased since the early 1990s. For example, there were approximately 265,000 lawsuits filed in federal court in fiscal year 1992.[154] In contrast, in fiscal year 2008, there were approximately 350,000 lawsuits, an increase of

152. Uruguay Round Agreements Act, Pub. L. No. 103-465, sec. 532(a), § 154, 108 Stat. 4809, 4983–85 (1994).

153. Act of Nov. 29, 1999, Pub. L. No. 106-113, § 1000(a)(9), 113 Stat. 1501, 1536 (1999) (enacting into law the American Inventors Protection Act of 1999, S. 1948, 106th Cong., tit. IV, sec. 4502, § 122 (1999) (codified at 35 U.S.C. § 122 (2006))).

154. *U.S. District Court—Judicial Caseload Profile*, UNITED STATES COURTS (Apr. 1998), http://www.uscourts.gov/cgi-bin/cms.pl (accessed by choosing "ALL DISTRICT COURTS" from the dropdown menu and pressing the "Generate" button).

about twenty-five percent.[155] Total federal case appeals also slightly increased in a generally linear fashion between the early 1990s and 2008.[156] The rate of summary judgment in all litigation also may have changed over time.[157]

In addition to the changing nature of patent and civil litigation over time, any study of appellate decisions has certain inherent limitations. These limitations include most notably a potential selection bias.[158] Because previous articles described in detail the potential selection effect, it will be only briefly discussed here.[159] First, appellate decisions are not a random sample of all patent disputes or all patent infringement complaints. Obviously, in real-world patent litigation, in each case the merits, the parties, and the parties' resources differ.[160] Each of these factors affects which disputes become lawsuits, which lawsuits proceed through final, appealable judgment, and which decisions are appealed. The closer cases, such as those fifty-fifty cases wherein either party could prevail (including cases with closer claim construction arguments), may be appealed at a higher frequency.[161] However, the present study does not rely upon case outcomes, which evaluate the performance of the district court and are susceptible to distortion based on selection effects.[162] Instead, because this Article examines

155. *U.S. District Court—Judicial Caseload Profile*, UNITED STATES COURTS (2008), http://www.uscourts.gov/cgi-bin/cmsd2008.pl (accessed by choosing "ALL DISTRICT COURTS" from the dropdown menu and pressing the "Generate" button).

156. *Federal Court Management Statistics*, UNITED STATES COURTS (2010), http://www.uscourts.gov/fcmstat/index.html.

157. Paul W. Mollica, *Federal Summary Judgment at High Tide*, 84 MARQ. L. REV. 141, 141 (2000) (noting the declining percentage of civil cases proceeding to trial in federal courts over time, and tying that to "the emergence of summary judgment as the new fulcrum of federal civil dispute resolution").

158. *See* Schwartz, *Pre*-Markman *Reversal Rates*, *supra* note 136, at 1101–06.

159. *E.g.*, Richard S. Gruner, *How High Is Too High?: Reflections on the Sources and Meaning of Claim Construction Reversal Rates at the Federal Circuit*, 43 LOY. L.A. L. REV. 981, 1003–24 (2010); Schwartz, *Pre*-Markman *Reversal Rates*, *supra* note 136, at 1101–06; Ted Sichelman, *The Myths of (Un)Certainty at the Federal Circuit*, 43 LOY. L.A. L. REV. 1161, 1179–83 (2010).

160. *See* John R. Allison & Mark A. Lemley, *Empirical Evidence on the Validity of Litigated Patents*, 26 AIPLA Q.J. 185, 202–05, 250–51 (1998).

161. Kimberly A. Moore, *Are District Court Judges Equipped To Resolve Patent Cases?*, 15 HARV. J.L. & TECH. 1, 9–10 (2001); George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL. STUD. 1, 4, 16 (1984). Other empirical studies have reported plaintiff win rates in patent jury trials at nearly seventy percent, contrary to what one would expect using the limiting case of the Priest/Klein economic theory. Kimberly A. Moore, *Judges, Juries, and Patent Cases—An Empirical Peek Inside the Black Box*, 99 MICH. L. REV. 365, 385–86 (2000); *see also* Alan C. Marco, *The Selection Effects (and Lack Thereof) in Patent Litigation: Evidence from Trials*, 4 TOPICS ECON. ANALYSIS & POL'Y, iss. 1, art. 21 (2004), http://www.bepress.com/bejeap/topics/vol4/iss1/art21/ (reporting that there does not appear to be a selection bias tending to produce a fifty percent patent infringement win rate, but that there does appear to be a selection bias toward fifty percent in the validity win rate).

162. Wagner & Petherbridge, *supra* note 148, at 1127–29.

appellate choices such as whether to publish an opinion or designate it as precedential, and other metrics of solely appellate court decision-making, the potential selection effects problem is slightly muted.

Studying the time period surrounding a major change such as *Markman* also presents difficulties. For example, the cases that survived until an appellate decision may have changed across this period. When juries performed claim construction, more accused infringers may have settled instead of risking a jury ruling.[163] After *Markman*, accused infringers in similar cases could more freely litigate claim construction before a judge. As more courts utilized a completely separate claim construction hearing, the number of litigants willing to proceed through claim construction may have increased. For these reasons, the types of cases that resulted in a Federal Circuit decision may be different before and after *Markman*.

The word count appellate database has a separate concern because it has a smaller sized dataset. It currently analyzes word count information from less than two hundred Federal Circuit opinions over a fifteen-year period. On average, there were about ten opinions per year,[164] with more opinions in recent years and fewer in earlier years. The small number of observations affects the statistical tests performed. Furthermore, word counts in appellate decisions only illustrate behavior at the Federal Circuit level. The database does not directly report doctrines or word counts raised in the trial court. Moreover, because it is limited to cases in which the court granted summary judgment, it consists almost entirely of cases wherein the accused infringer prevailed at the district court. For each of the foregoing reasons, all results and discussion of the data are subject to the limitations discussed in this Section.

IV. DATA ON THE DEMISE OF THE DOCTRINE OF EQUIVALENTS

Using the theories provided in Part II, this Part provides hypotheses on the ramifications of these phenomena including hypotheses about the doctrine of equivalents. Part IV also sets forth the results of empirical testing of the hypotheses. Section IV.A provides the doctrinal reallocation hypotheses and results. Section IV.B examines the doctrinal displacement hypotheses and results.

163. Schwartz, *Pre*-Markman *Reversal Rates*, *supra* note 136, at 1099.

164. There are more than ten claim construction opinions per year. However, the word count appellate database was created using only cases in which both claim construction and the doctrine of equivalents were discussed.

A. REALLOCATION HYPOTHESES AND RESULTS

This Section examines hypotheses and data that suggest that the horizontal and vertical reallocations in *Markman* and *Cybor* made claim construction more central and important in patent litigation. According to the theory proposed in Part II, *supra*, the horizontal and vertical reallocations illustrated in Figure 3 provided more control of claim construction to the appellate court. The control gave the Federal Circuit the ability to emphasize or deemphasize claim construction in the overall context of patent litigation. When confronted with the opportunity, the Federal Circuit elected to make claim construction more rather than less important. At first glance, this statement seems obvious—claim construction is perhaps the central doctrine in patent litigation today. Practitioners and professors who came of age after *Markman* may believe that claim construction was always central to patent litigation. However, before *Markman*, it is debatable whether claim construction was as important.[165] This hypothesis tests the conventional wisdom that claim construction was always important.[166]

The hypotheses derive from the time period around *Markman* and *Cybor*. The Federal Circuit exerted control over lower courts[167] and over claim construction as a result of *Markman*.[168] This Article argues that the Federal Circuit enhanced the importance of claim construction to increase the predictability of patent litigation. This explanation is consistent with the remarks of Judge Plager,[169] and with the statements in the *Markman* opinion itself.[170] Generally, a court using its control to increase the importance of a

165. *See* Craig Allen Nard, *Process Considerations in the Age of* Markman *and Mantras*, 2001 U. ILL. L. REV. 355, 360 ("It was not until the 1995 en banc Federal Circuit decision in *Markman* . . . that [the notice function of patents] reached the forefront of patent law jurisprudence.").

166. *See, e.g.*, Edward M. O'Toole, *How To Prepare for and Conduct* Markman *Hearings 2006*, *in* HOW TO PREPARE & CONDUCT *MARKMAN* HEARINGS 2006, at 175, 207 (PLI Patents, Copyrights, Trademarks, & Literary Prop., Course Handbook Series No. 873, 2006) ("After all, claim construction has always been an important aspect of resolution of patent disputes, and most patent cases continue, as they always have, to turn on claim construction—impacting the issue of infringement or validity, or both.").

167. Legal doctrine is taught by higher courts to lower courts. *See* Tonja Jacobi & Emerson H. Tiller, *Legal Doctrine and Political Control*, 23 J.L. ECON. & ORG. 326 (2007).

168. John F. Duffy, *On Improving the Legal Process of Claim Interpretation: Administrative Alternatives*, 2 WASH. U. J.L. & POL'Y 109, 123 (2000) ("Together *Markman* and *Cybor* have . . . centralized judicial power to interpret claims in the Federal Circuit.").

169. *See* Plager, *supra* note 50.

170. Markman v. Westview Instruments, Inc., 52 F.3d 967, 989 (Fed. Cir. 1995) (en banc) (Mayer, J., concurring) (noting that the majority opinion attempts "to free patent litigation from the 'unpredictability' of jury verdicts"), *aff'd*, 517 U.S. 370 (1996).

doctrine will perform certain observable tasks. Empirically testing whether these tasks occurred in patent law permits evaluation of the theory.

The Article sets forth *infra* three hypotheses relating to doctrinal reallocation. The first hypothesis is that, after *Markman*, the Federal Circuit issued a greater percentage of written claim construction opinions. The second is that, after *Markman*, the Federal Circuit issued a greater percentage of precedential claim construction opinions. The third is that, after *Markman*, a greater percentage of claim construction appeals arose from summary judgment. Each of these hypotheses, both separately and together, is consistent with the view that the Federal Circuit made claim construction more important after *Markman*.

1. *Reallocation Hypothesis #1: After* Markman I*, the Federal Circuit issued a greater percentage of written opinions*

The first hypothesis contends that the Federal Circuit issued a greater percentage of written opinions on claim construction after *Markman*. More written opinions signal that the Federal Circuit believes that claim construction is important. Obviously, the appellate court can only issue opinions on a particular doctrine that the parties raise on appeal. This caps the maximum number of opinions an appellate court can generate.

However, the federal courts of appeal need not produce a written opinion in every case. They have the option of affirming without providing a written opinion, a procedure known as summary affirmance.[171] When using a summary affirmance, the appellate court disposes of the case without explaining its reasoning.[172] Alternatively, the court may issue a written opinion that sets forth the complete basis for its opinion.[173] The court chooses which cases to decide by written opinions and which to decide by summary affirmance.[174] Issuing more written opinions cues litigants of the increased importance of the doctrine. A greater percentage of claim construction appeals were (and will continue to be) resolved by written opinions after *Markman*.

171. *See* FED. CIR. R. 36.

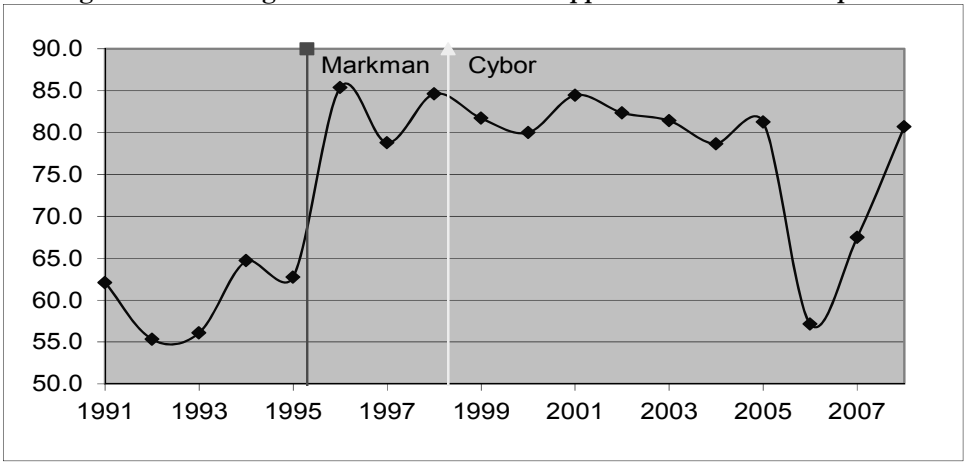
172. Pamela S. Karlan, Comment, *Electing Judges, Judging Elections, and the Lessons of* Caperton, 123 HARV. L. REV. 80, 83 (2009); *see, e.g.*, Sutton v. Nokia Corp., No. 2010-1218, 2010 WL 5230901 (Fed. Cir. 2010) (affirming decision of lower tribunal without providing any explanation).

173. FRANK M. COFFIN, ON APPEAL: COURTS, LAWYERING, & JUDGING 165 (1st ed. 1994).

174. Taylor v. McKeithen, 407 U.S. 191, 194 n.4 (1972) ("We, of course, agree that the courts of appeals should have wide latitude in their decisions of whether or how to write opinions. That is especially true with respect to summary affirmances.").

Moving now to the empirical results, Figure 4, *infra*, shows the annual percentage of claim construction appeals that garnered a written opinion.

Figure 4: Percentage of Claim Construction Appeals with a Written Opinion



The ordinate illustrates the percentage of decisions that received a written opinion plotted against the year of the Federal Circuit disposition. All of the opinions in a given year are collapsed into a single data point.[175] For clarity, Figure 4 notes the dates of the *Markman I* and *Cybor* decisions.

Almost simultaneously with *Markman I* in 1995, the Federal Circuit decreased the rate of summary affirmances and began issuing more written opinions on claim construction. To provide a more comprehensive analysis, a multiple-regression model of the data was developed. According to the model, even when other potentially explanatory variables are controlled for, the odds of a written opinion after *Markman I* are more than twice as high as

175. There have been 1288 opinions over eighteen years, and there are approximately seventy-two opinions per year. Regression assumes that each variable is independent and identically distributed, but this assumption may not hold in a precedential system in which a prior decision influences subsequent decisions. *See* James Greiner, *Judicial Decisions as Data Points*, SOCIAL SCIENCE STATISTICS BLOG (March 20, 2007, 16:40 EST), http://www.iq.harvard.edu/blog/sss/archives/2007/03/ (arguing that certain empirical assumptions should be cautiously considered in a precedential system). After a court decides a significant precedential case which clarifies or alters the substantive doctrine, one might expect a subsequent decrease in written and precedential opinions. Here, the trend illustrated in Figures 4 and 5 shows increased precedential and written opinions after *Markman*, which is in the opposite direction of this prediction. In fact, the role of precedent may if anything be downwardly tampering the effects. However, the effects of precedent are very complicated, and an alternative hypothesis is that new case law invites uncertainty and more precedential written opinions. Further research is needed into the general question of the relationship between legal precedent and the assumptions in empirical models.

before *Markman I*, and the difference is statistically significant.[176] As illustrated in Figure 4, the Federal Circuit resolved only about sixty percent of all claim construction appeals from 1991 until 1995 with a written opinion. Thus, it was quite likely that a claim construction appeal resulted in no case law (either precedential or non-precedential), and no guidance to litigants.[177] After 1995 (the year *Markman I* was decided), however, the rate of written opinions quickly increased to between eighty and eighty-five percent.[178]

It should be noted that there is a downward spike in Figure 4 that begins in September 2006 and ends in very early 2007. The reason for that spike presently cannot be completely explained. The time period of the spike begins in September, which is roughly contemporaneous with the turnover of law clerks. Perhaps for that year only, some judges amassed a backlog of cases with their old clerks and used summary affirmances to pare their dockets. An alternative explanation relates to district court decisions appealed shortly before the en banc *Phillips v. AWH* decision but decided by the Federal Circuit around the time of the spike.[179] Perhaps these were appeals in which the district court correctly construed the claims but used the wrong rationale, such as placing a heavy reliance on dictionaries. The Federal Circuit judges may have agreed that the result was correct and utilized summary

176. The detailed regression results can be found in Tables 4a and 4b in the Appendix. The control variables for the main regression of Tables 4a are the geographic location of the district court (i.e., which Circuit the district court resided in); whether the district court was in one of the ten busiest patent districts during the given year; the technology of the underlying patent (chemical, mechanical, or electrical), the posture of the district court judgment (preliminary injunction, summary judgment, jury trial, or bench trial), and winner at the district court (patentee or accused infringer). The odds ratio is 2.86 with a p-value of 0.000. A p-value of 0.05 or less signifies that the null hypothesis—in this case that there is no difference in the use of summary affirmance/Rule 36 before and after *Markman*—can be rejected with a 95% confidence level. Here, the p-value is 0.000, which means that the null-hypothesis can be rejected. The odds ratio means it was 186% more likely at the mean for a summary affirmance/Rule 36 claim construction decision before *Markman I*, after controlling for the aforementioned variables. A separate regression shown in Table 4b includes the total number of patent opinions and Rule 36 cases on any issue, not just claim construction. The results of the separate regression show that even controlling for Federal Circuit patent opinions and Rule 36 decisions outside of claim construction, it was 96% more likely at the mean for a Rule 36 claim construction decision before *Markman I* than after (p-value=0.000). For a discussion of regression analysis, see *supra* note 175.

177. The author takes no position on the optimal or minimum amount of case law to develop a doctrine. Rather, it is only noted that after *Markman I*, the court employed Rule 36 less frequently than before.

178. It is also worth noting that there were no substantial increases in the number of active Federal Circuit judges during this time period.

179. Phillips v. AWH Corp., 415 F.3d 1303 (Fed. Cir. 2005) (en banc) (resolving an intra-circuit split on whether dictionary definitions should be the default claim construction of a disputed term).

affirmances in these instances. Finally, it is possible that it is an unintentional clustering of results in the data.[180] In any event, this data was included within the regression and the results remain statistically significant.

After *Markman*, the Federal Circuit vastly increased the number of written opinions describing claim construction methodology and analysis.[181] Not only did it receive more appeals involving claim construction after *Markman*,[182] but it drafted written opinions for a greater percentage of those appeals.[183] The larger volume of opinions signals to litigants the increased importance of the doctrine. When more cases address a particular issue, litigants understand that the court is interested in the issue. Because summary affirmances do not include any written opinions, the public (other than the particular litigants involved in the case) cannot easily know what issues were raised in those cases. The focus on claim construction in opinions encouraged litigants to raise this issue on appeal. This further increased the significance of the doctrine of claim construction.

2. *Reallocation Hypothesis #2: After* Markman I*, the Federal Circuit issued a greater percentage of precedential opinions*

The second hypothesis is that the Federal Circuit issued a greater percentage of precedential opinions on claim construction after *Markman*. In addition to resolving cases without any opinion (by summary affirmance), the courts can also issue different types of written opinions. For every written opinion, the courts of appeals may designate the opinion as either precedential or non-precedential.[184] Precedential opinions have various functions—announcing new law, applying settled law to new facts, and

180. A Federal Circuit judge on the bench during this time period told the author that he believes it is random.

181. Some may argue that a court issues a written opinion instead of a summary affirmance when the dispute is complicated. However, there is no reason to believe that the complexity of disputes changed in 1995. So this does not explain the change in the frequency of written opinions in claim construction beginning in 1995.

182. There are several potential reasons that parties brought more appeals involving claim construction issues. One reason, which is consistent with doctrinal reallocation, is that the parties recognized the court's elevation of the doctrine in importance. If parties know the court believes an issue is important, it is not surprising that it is frequently raised.

183. It is possible that the increase in Rule 36 is mere happenstance, or alternatively, due to an increase in the quality of the briefs submitted by the parties. However, because the timing of the increase so closely corresponds to *Markman*, these other explanations appear unlikely.

184. *See* Kimberly A. Moore, Markman *Eight Years Later: Is Claim Construction More Predictable?*, 9 LEWIS & CLARK L. REV. 231, 234–35 (2005).

recording important discussion or criticisms of settled rules.[185] The court spends more time drafting a precedential opinion because it binds future appellate panels as precedent.[186] Non-precedential opinions, also known as unpublished opinions, are citable by litigants but do not serve as precedent in the district or appellate court.[187] The main rationale for unpublished opinions is that they conserve judicial resources.[188] They are typically shorter with less discussion of the facts. All federal courts of appeal utilize non-precedential opinions to some extent.[189] Courts can choose which opinions to make precedential or non-precedential.

Increasing the proportion of precedential opinions may increase the importance of the doctrine. Precedential opinions signal that the appellate court considers the doctrine significant. To increase the importance of claim construction, the Federal Circuit increased the percentage of precedential written opinions after *Markman*.

There are other possible reasons a court may increase the number of written opinions or the designation of precedential opinions. For example, it could be that a new, fledgling doctrine needs to be fleshed out more in case law. It is doubtful that this rationale applies to claim construction, even though a new actor—the judge—was given responsibility for the task. The Federal Circuit articulated the canons of claim construction in numerous

185. William L. Reynolds & William M. Richman, *Limited Publication in the Fourth and Sixth Circuits*, 1979 DUKE L.J. 807, 808; *see also* William M. Richman & William L. Reynolds, *Appellate Justice Bureaucracy and Scholarship*, 21 U. MICH. J.L. REFORM 623, 632–33 (1988); William L. Reynolds & William M. Richman, *The Non-precedential Precedent—Limited Publication and No-Citation Rules in the United States Courts of Appeals*, 78 COLUM. L. REV. 1167, 1182–83 (1978).

186. *See* Lauren K. Robel, *Caseload and Judging: Judicial Adaptations to Caseload*, 1990 BYU L. REV. 3, 50.

187. FED. R. APP. P. 32.1 was amended in 2007 to require all circuit courts to permit citation of unpublished opinions. Before 2007, some circuits permitted citation of unpublished opinions without limitation and some discouraged their citation. For a good list of the differences among circuits pre-2007, see Robert T. Reagan, *Citing Unpublished Federal Appellate Decisions Issued Before 2007*, FEDERAL JUDICIAL CENTER (Mar. 9, 2007), http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Unpub_Opinions.pdf.

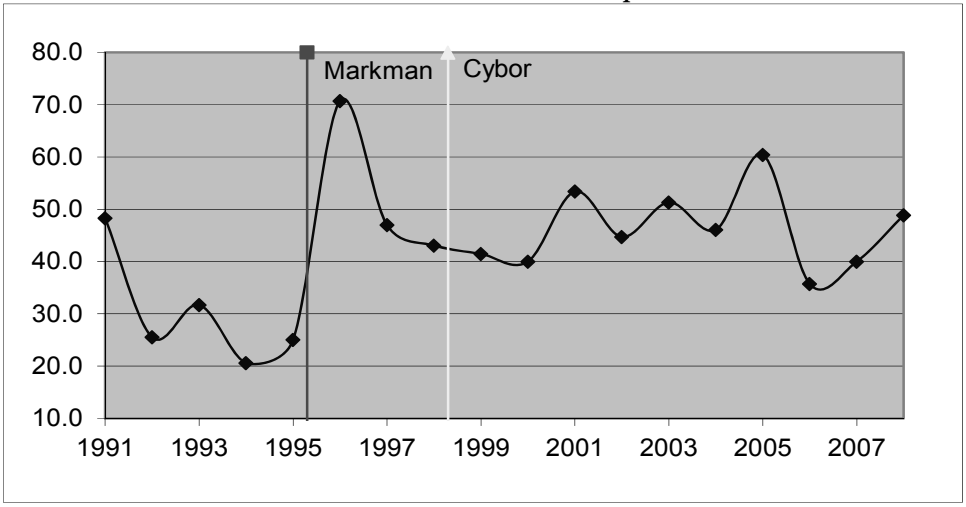
188. *See, e.g.*, William L. Reynolds & William M. Richman, *Studying Deck Chairs on the Titanic*, 81 CORNELL L. REV. 1290, 1293 (1996) (stating that increased judicial caseload required limited publication of cases).

189. Penelope Pether, *Constitutional Solipsism: Toward a Thick Doctrine of Article III Duty; or Why the Federal Circuits' Nonprecedential Status Rules Are (Profoundly) Unconstitutional*, 17 WM. & MARY BILL RTS. J. 955, 960 (2009) ("[T]he percentage of federal appellate decisions that are unpublished presently runs at almost eighty-five percent . . . .").

cases before *Markman*.[190] Furthermore, if the doctrine merely needed to be fleshed out, the increase in precedential opinions should be temporary, ending when the doctrine was sufficiently developed. Consequently, more precedential opinions over a long period of time may signal the increased salience of the doctrine.

Turning now to the empirical data, Figure 5, *infra*, shows the percentage of claim construction appeals resolved by a precedential opinion from 1991 until 2008.

Figure 5: Percentage of Claim Construction Appeals Resolved with a Precedential Opinion



After controlling for potentially explanatory variables in a multiple regression model, the results are statistically significant; it was more than twice as likely for the Federal Circuit to issue a precedential opinion after *Markman I* than before.[191] Beginning in 1996, there was a spike in the percentage of claim construction appeals that were resolved via precedential

190. *See generally* 5A DONALD S. CHISUM, CHISUM ON PATENTS § 18.03[2][a] (rev. 2007) (identifying twelve examples of canons of claim construction and providing citations of case authority for the canons dating back until the 1930s).

191. The regression details can be found in Table 5a in the Appendix. The control variables are the geographic location of the district court (i.e., which Circuit the district court resided in); whether the district court was in one of the ten busiest patent districts during the given year; the technology of the underlying patent (chemical, mechanical, or electrical), the procedural posture of the district court judgment (preliminary injunction, summary judgment, jury trial, or bench trial), and winner at the district court (patentee or accused infringer). The odds ratio is 2.35 with a p-value of 0.000, meaning that it is 135% more likely at the mean. The second regression controlling for overall Federal Circuit patent docket could not be performed. It was not feasible to gather data on precedential opinions on all areas of law.

opinions. Non-precedential opinions and summary affirmances are two different methods of deciding cases without a precedential opinion.

In doctrines such as claim construction that are based upon guidelines rather than rules, precedent is especially important in teaching how to properly decide cases.[192] Precedent binds future court panels. In theory, as the volume of precedents increases, courts should be more likely to find a prior opinion that matches or nearly matches the facts at hand.[193] After the 1996 spike, the level of precedential opinions remained elevated compared to the pre-*Markman* levels. More specifically, in the years before *Markman* (1991–1994), the Federal Circuit decided 30.5% of claim construction appeals with precedential opinions. But, in the years afterwards (1997–2003), the Federal Circuit decided 46.5% with precedential opinions. Some may argue that increased precedential opinions were necessary to develop the claim construction doctrine after *Markman* clarified that it was a matter of law. However, the Federal Circuit articulated the various canons of claim construction in numerous cases before 1995.[194] Furthermore, this would not explain why there is still, fifteen years after *Markman I*, an elevated level of precedential opinions. Today, the Federal Circuit has explained each of the canons in numerous post-*Markman* opinions.

Thus, the Federal Circuit increased the number of written and precedential claim construction opinions after *Markman*. Since the Federal Circuit arguably desired litigants and district court judges to focus on claim construction, it appears to be a reasonable and prudent decision to increase the body of case law analyzing that issue. Now, almost fifteen years after *Markman*, there are a plethora of precedential claim construction opinions.

3. Reallocation Hypothesis #3: After Cybor*, a greater proportion of appeals were from grants of summary judgment*

The third hypothesis asserts that a greater proportion of appeals that reach the Federal Circuit were from grants of motions for summary

192. Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557, 577–78 (1992) (discussing the role of precedent when standards are utilized).

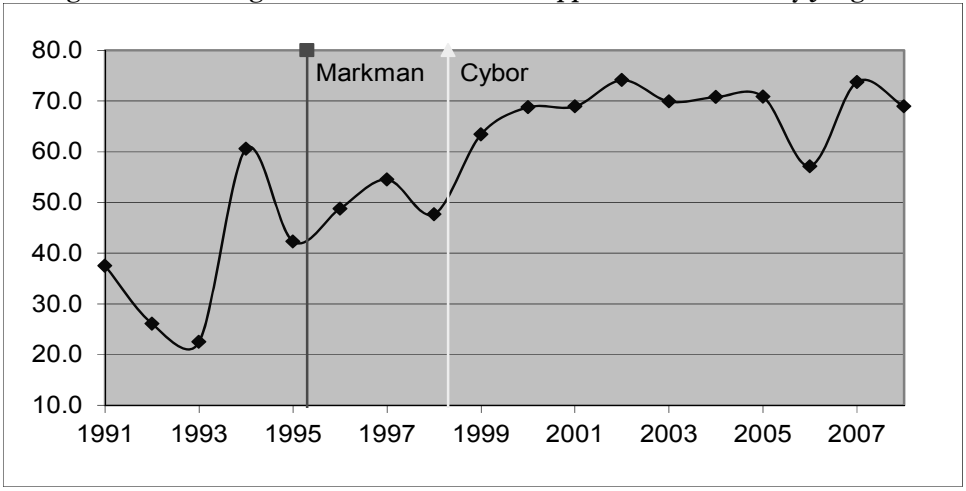
193. *See* David Luban, *Settlements and the Erosion of the Public Realm*, 83 GEO. L.J. 2619, 2622–23 (1995) (arguing that the rules and precedents resulting from litigation have "obvious importance for guiding future behavior and imposing order and certainty").

194. *See generally* 5A CHISUM, *supra* note 190, § 18.03[2][a] (identifying twelve examples of canons of claim construction and providing citations of case authority for the canons dating back until the 1930s). To be fair, there were some short-lived intra-circuit disputes about how to perform claim construction. Holbrook, *supra* note 27, at 146–48 (noting the court's struggle between the *Vitronics* and *Texas Digital* methodologies). However, the level of precedential opinions has been relatively constant, not tied to particular disagreements.

judgment. As the claim construction doctrine became more important, it was dispositive in more cases. District court judges could more easily grant summary judgment because of the horizontal reallocation. In patent cases, there is often less dispute over the structure of the accused product than there is concerning the construction of the patent. Because claim construction is a matter of law, it is resolved without using the summary judgment standard, namely, all inferences to the non-moving party. Once the primary battle on claim construction is resolved, the remaining issues on literal infringement are more straightforward.[195] The vertical reallocation also encouraged resolution by summary judgment. Because the district court judges understood that their decisions would be reviewed on appeal using a de novo standard, they desired to resolve the cases quicker. Finally, as the doctrine became more important and more central to patent law, district courts could entirely dispose of more cases after construing the claims.

The data reflects that the percentage of claim construction appeals decided by summary judgment increased over time. Figure 6, *infra*, shows the percentage of claim construction appeals that arose from district court summary judgment orders.

Figure 6: Percentage of Claim Construction Appeals from Summary Judgment



As claim construction became more important, a greater percentage of appeals reviewed summary judgment decisions as opposed to the results of bench trials, jury trials, or rulings on preliminary injunctions. This data is

195. Literal infringement and claim construction may be seen as doctrinally linked after *Markman I*.

consistent with data from other researchers arguing that *Markman* encouraged summary judgment in patent cases.[196]

The multiple regression model indicates that it was over one hundred percent more likely for the appeal to be from a grant of summary judgment after *Cybor* than before.[197] As can be seen from Figure 6, less than forty percent of claim construction appeals were from summary judgment during the period from 1991 until 1993. The first increase in appeals from summary judgment occurred before *Markman I* in 1994. Looking at this data alone, it does not appear that *Markman I* by itself immediately caused the increase in summary judgments. Instead, a temporary increase appears just before *Markman I*, and a sustained increase appears around 1998, the time of the *Cybor* decision.

However, there is a natural time lag in the litigation process. The appeal process itself takes approximately one year, and the trial court proceedings longer than that.[198] Considering this delay, the possibility that the increase in appeals from summary judgment is due to *Markman* cannot be excluded. After 1998, the percentage of claim construction appeals from summary judgment sharply increased to almost seventy percent. It thereafter remained substantially constant. In the last ten years, approximately seventy percent of appeals of claim constructions arose in the summary judgment context. Others reported similar increases in summary judgment in patent litigation, even beyond claim construction. For example, Petherbridge, in his findings, noted a trend toward an increased percentage of appeals from a finding of summary judgment of non-infringement.[199] Data on whether district courts issued fewer summary judgments in earlier years are not readily available.

---

196. Allison & Lemley, *supra* note 2, at 958 (noting that *Markman* drives summary judgments); Burk & Lemley, *supra* note 26, at 1795 (asserting that *Markman* increased summary judgment); Lee & Krug, *supra* note 131, at 59 (observing that the *Markman* decisions could encourage summary proceedings); Petherbridge, *Claim Construction Effect*, *supra* note 2, at 243.

197. The results of the regression can be found in Table 6a in the Appendix. The control variables are the same as those for precedential opinions described *supra* note 191. The odds ratio for this regression is 3.38 with a p-value of 0.000.

198. Jay P. Kesan & Gwendolyn G. Ball, *How Are Patent Cases Resolved? An Empirical Examination of the Adjudication and Settlement of Patent Disputes*, 84 WASH. U. L. REV. 237, 282–83 (2005) (noting that the average trial court patent case, including and weighted down by those which settled early, pended well over a year).

199. Petherbridge, *Doctrine of Equivalents*, *supra* note 2, at 1394; *see also* Mary A. Woodford, Presentation to Ropes & Gray LLP: Preliminary Analysis of IPLC Data: Patent Infringement Cases 13 (June 2009) (on file with author) (reporting fifty-six percent of patent cases filed between 2000 and 2008 and which reached judgment were decided by summary judgment).

The rise in summary judgment since *Markman* is not surprising. First, as Allison and Lemley argued, once the district court judges spent time evaluating patent claims, it was only natural for them to attempt to resolve the case. In fact, many judges only construed the claims in the context of dispositive motions. Further, summary judgment may be appropriate in a large number of these cases. Once the judges construed the claims, there will not be a genuine issue of material fact in cases where the parties do not dispute the structure or function of the accused device or method. Second, as discussed in Section II.B.2, the doctrinal displacement of the doctrine of equivalents followed the reallocation in claim construction. The doctrine of equivalents was historically a quintessential jury issue. By reducing the importance of this doctrine, judges could grant more summary judgment motions. Third, the increased importance of claim construction and the reduced importance of the doctrine of equivalents may motivate both courts and litigants to resolve cases via summary judgment. The Federal Circuit's high claim construction reversal rate is well known.[200] The Federal Circuit also does not review claim constructions through an interlocutory appeal.[201] Summary judgment permits quick review by the Federal Circuit, a goal often shared by both litigants and the district court.[202]

The doctrinal reallocation in patent law had other effects. Overall, the reallocation focused resources—of the Federal Circuit, of district courts, and of litigants—more on a single issue in the case. Many believe that claim construction ought to be central to patent litigation. By focusing resources on this one issue, the end product is better-organized Federal Circuit opinions. Before and immediately after *Markman*, the Federal Circuit issued often-confusing claim construction opinions.[203] The Federal Circuit would blend claim construction and infringement discussions. Now the Federal Circuit's claim construction opinions are better written and more

200. Amgen Inc. v. Hoechst Marion Roussel, Inc., 469 F.3d 1039, 1040 (Fed. Cir. 2006) (Michel, C.J., dissenting) (noting the problem in claim construction of "a steadily high reversal rate").

201. *See, e.g.*, V. Ajay Singh, *Interlocutory Appeals in Patent Cases Under 35 U.S.C. § 1292(c)(2): Are They Still Justified and Are They Implemented Correctly?*, 55 DUKE L.J. 179, 196 (2005) ("The Federal Circuit has thus far refused to hear permissive appeals related to claim construction.").

202. Kathleen M. O'Malley et al., *A Panel Discussion: Claim Construction from the Perspective of the District Judge*, 54 CASE W. RES. L. REV. 671, 681 (2004).

203. An old, illustrative case is Morton Int'l, Inc. v. Cardinal Chem. Co., 959 F.2d 948 (Fed. Cir. 1992). In *Morton*, the opinion has an unlabeled background section. It follows with a discussion containing a section labeled I, but no other subsections. Section I blends claim construction and literal infringement and also addresses attorneys' fees under 35 U.S.C. § 285.

organized.[204] The clearer organization after *Markman* aids the reader, whether it is district court judges or potential litigants, in following the courts' reasoning.

B. DOCTRINAL DISPLACEMENT HYPOTHESES AND RESULTS

This Section examines hypotheses and data that suggest that the doctrine of equivalents has been displaced by claim construction. Turning back to the theory, doctrinal displacement suggests that claim construction should have displaced another doctrine. The Federal Circuit horizontally reallocated the doctrine of claim construction to the judge, vertically reallocated the standard of review to de novo, and raised the profile and importance of the claim construction doctrine. Raising the importance of claim construction meant that more litigants would elect to focus on it. This caused a displacement of other doctrines in patent law. The first hypothesis is that as claim construction became more important, the doctrine of equivalents became less important.

1. Displacement Hypothesis #1: After Markman I*, the frequency with which the Federal Circuit analyzed the doctrine of equivalents decreased and claim construction increased*

According to displacement theory,[205] three influences—litigation constraints, judicial constraints, and the gatekeeping nature of claim construction—together caused a displacement of the doctrine of equivalents. Courts used claim construction to resolve all claim scope issues. District court judges began to rely more on summary judgment of non-infringement. Litigants subsequently must have learned that the doctrine of equivalents was unlikely to prevail; consequently, arguments relating to the doctrine of equivalents were dropped or downplayed in many briefs. Because claim construction consumed more words, less space was left for other issues.[206] This hypothesis will be evaluated by analyzing the issues addressed in Federal Circuit written opinions in patent infringement appeals over selected years.

---

204. An exemplary recent case is *Baldwin Graphic Systems, Inc. v. Siebert, Inc.*, 512 F.3d 1338 (Fed. Cir. 2008). There, the opinion consists of three parts: (1) a background description of the technology and proceedings in the district court; (2) a detailed discussion of the claim construction dispute and resolution; and (3) a brief conclusion that a grant of summary judgment must be vacated because of an erroneous claim construction. *Id.*

205. *See supra* Section II.B.

206. While traditionally a jury issue, the doctrine of equivalents can be resolved on summary judgment if one of the legal limitations to the doctrine of equivalents applies, or if there is no disputed issue of material fact. Consequently, when evaluating appeals of judgments for non-infringement, the Federal Circuit must consider the doctrine of equivalents, if raised by the patentee.

After *Markman*, the frequency with which the Federal Circuit discussed the doctrine of equivalents should have dropped because litigants did not press it on appeal.

Previous scholars have noted the decline of the doctrine of equivalents; yet, there is debate on the cause and timing of its demise. As noted *supra*, Allison and Lemley argue that *Markman* itself ended the doctrine of equivalents.[207] They assert that after a judge construes the claims and concludes that the accused product is not within the literal scope of the claims, the judge likely desires to resolve the case on summary judgment. To completely resolve the case requires that the judge also conclude that the product is not equivalent. Petherbridge argues that *Festo* appeared to be a tipping point for the doctrine of equivalents, showing that after *Festo*, a patentee's success rate on appeal on the doctrine of equivalents significantly dropped. Thus, Allison and Lemley disagree with Petherbridge as to the triggering event of the decline of the doctrine of equivalents.

The present Article cannot conclusively resolve the debate. Both scholars may be partially correct. However, along with a new theoretical framework to understand the decline of the doctrine of equivalents, it presents some additional evidence on this question. The new data supports the view that at least part of the decline occurred immediately after the increased importance of claim construction. The doctrine of equivalents may have continued its decline after *Festo*. As discussed in Section IV.A, *supra*, the doctrinal reallocation of claim construction resulted in claim construction becoming more important in patent litigation. Shifting the importance of a single doctrine has larger implications in real-world litigation, and the claim construction shift preceded a decline in the significance of the doctrine of equivalents.

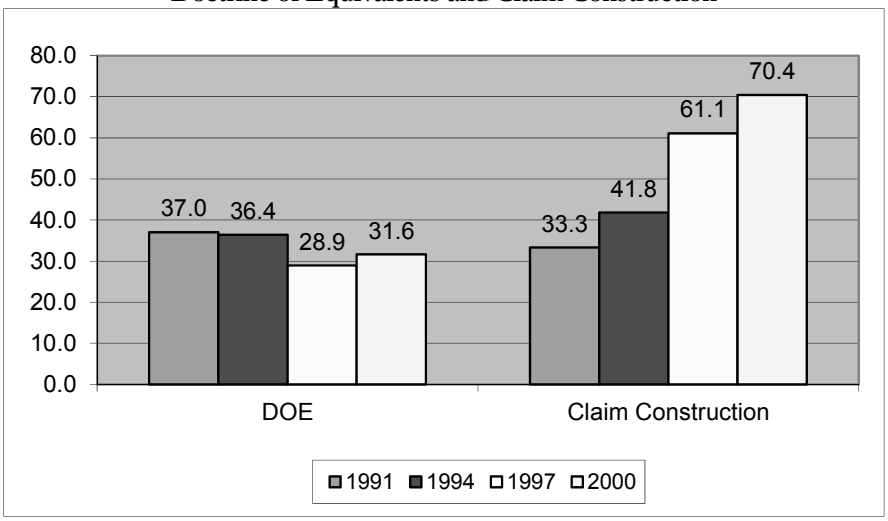
The doctrine of equivalents appears to have been in a more prominent position before *Markman*. When reading the opinions issued from 1991 until 1995, the author found that the Federal Circuit addressed the doctrine of equivalents more frequently during that time period than the present time. In particular, the Federal Circuit often discussed the doctrine of equivalents in robust detail, much the way the Federal Circuit discusses claim construction today. Sometimes the Federal Circuit discussed the doctrine of equivalents in the same breath as claim construction. Additionally, claim construction was less important in patent litigation pre-*Markman*. Unlike patent litigation today, patentees then did not focus on the claim language to prove their charges of infringement. Instead, they presented arguments to the jury about

207. *Supra* Part I.

the "invention" and attempted to divorce the "invention" from the specific claim language.

Figure 7, *infra*, displays the prevalence of claim construction and the doctrine of equivalents before the Federal Circuit in 1991, 1994, 1997, and 2000.[208]

Figure 7: Percentage of Federal Circuit Written Opinions Analyzing the Doctrine of Equivalents and Claim Construction



The most striking aspect of Figure 7 is the increase in claim construction at the Federal Circuit. Before *Markman*, claim construction appeared in less than fifty percent of appellate decisions, as shown in the left two bars. After *Markman*, it substantially increased, reaching seventy percent of decisions by 2000.[209] During the same time period, the doctrine of equivalents declined, albeit less dramatically. While the results surrounding the doctrine of equivalents are not statistically significant, the general trend is not in the wrong direction. These results reflect a limited sample of cases.

One reason why a larger drop in the doctrine of equivalents is not evident relates to the increase in appeals from grants of summary judgment of non-infringement. In order to grant summary judgment of non-infringement, the district court must determine that there is no issue of

208. The percentage is based upon the number of opinions raising the issue/doctrine relative to the total number of merits opinions for the year.

209. Claim construction is a gatekeeper doctrine that may be present in invalidity and infringement discussions, while the doctrine of equivalents may be present only in infringement discussions. However, claim construction always had this status as a gatekeeper; *Markman* did not start it. Consequently, the gatekeeper status cannot explain the difference between the rates before and after *Markman*.

material fact to either literal infringement or the doctrine of equivalents. To affirm, the Federal Circuit should mention, at least briefly, both bases of the district court's ruling. There was a large jump in appeals from orders granting summary judgment of non-infringement from 1991 until 2000. Considering all merits appeals in the appellate issue database, appeals from summary judgments of non-infringement comprised 13.0% of 1991 opinions, 20.0% of 1994 opinions, 28.9% of 1997 opinions, and 46.9% of 2000 opinions. Consequently, changes in summary judgment practice in patent litigation, perhaps driven by *Cybor*, may have played a role in inflating the number of doctrine of equivalents arguments raised in later years. This is consistent with the judicial constraints explanation for displacement.

The coding mechanism used in the appellate issue database may also partially explain why the drop in the doctrine of equivalents appears modest. Each issue raised in the appellate decision was weighted equivalently. For example, if the Federal Circuit discussed doctrine of equivalents for a paragraph and claim construction for five pages, the database coded each doctrine the same.

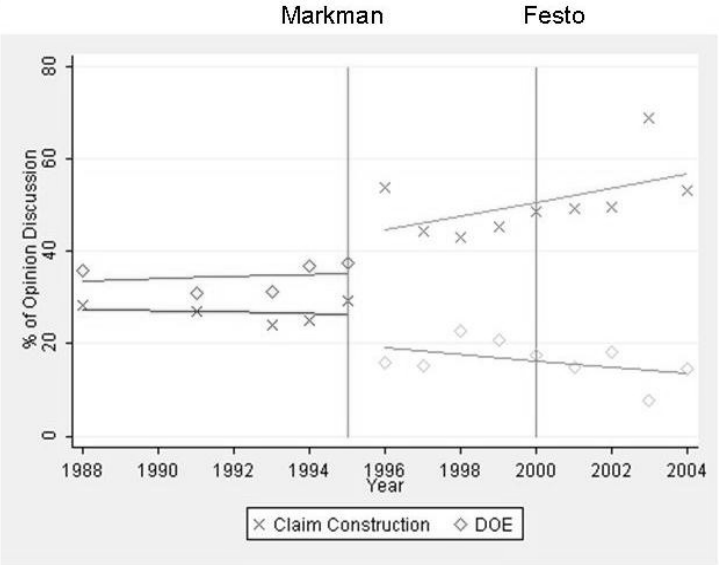
2. *Displacement Hypothesis #2: After* Markman I*, the Federal Circuit discussed the doctrine of equivalents in fewer words, and claim construction with more words*

The second displacement hypothesis was tested using the word count appellate database. Word count data permits analysis of displacement in greater detail and overcomes the aforementioned limitation of the appellate issue database. Even when the Federal Circuit analyzed the doctrine of equivalents by the Federal Circuit after *Markman*, it should occupy less space in the opinions. Decreased word count can support the hypothesis, assuming that word count is a proxy for importance or at least a proxy for how much analysis the court deemed sufficient for explanation and resolution of an issue.[210]

210. Other studies have used word count as a rough proxy for importance. *See, e.g.*, Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978–2005*, 156 U. PA. L. REV. 549, 587 (2008) (analyzing word count data because "in explaining (or defending) their analysis of a legal issue, judges are generally more likely to dedicate a greater share of their explanations to considerations that they deem to be more important"); Jennifer L. Groscup et al., *The Effects of Daubert on the Admissibility of Expert Testimony in State and Federal Criminal Cases*, 8 PSYCHOL. PUB. POL'Y & L. 339 (2002); Hall & Wright, *supra* note 147, at 117 ("For instance, some studies count the number of words or paragraphs devoted to discussing particular factors as an indication of the factors' relative importance."); Carl W. Roberts, *A Conceptual Framework for Qualitative Text Analysis*, 34 QUALITY & QUANTITY 259, 263 (2000) ("Analyses of word-counts yield inferences about the predominance of themes in texts.");

Figure 8, *infra*, shows the word count of summary judgment opinions involving claim construction and the doctrine of equivalents over time. As described in more detail in Section III.A.3, *supra*, the percentages were calculated by dividing the number of words in the opinion addressing claim construction or the doctrine of equivalents by the number of words in the "Discussion" section of the opinion addressing all issues on appeal.[211]

Figure 8: Percentage of Words Devoted to Claim Construction and Doctrine of Equivalents



The vertical lines at 1995 and 2000 in Figure 8 represent *Markman I* and *Festo*. It is important to remember that the word count appellate database only includes resolutions from summary judgment so juries did not evaluate either issue in the district court in any of the observations.

Figure 8 shows that the doctrine of equivalents occupied about a third of the discussion of Federal Circuit opinions before *Markman I* while claim construction entailed about a quarter of the opinions. During this time period, the doctrine of equivalents averaged a slightly greater percentage of the words than claim construction. After *Markman I*, the percentage of words devoted to the doctrine of equivalents dropped off, and varied from

Paul J. Wahlbeck, *The Development of a Legal Rule: The Federal Common Law of Public Nuisance*, 32 L. & SOC'Y REV. 613 (1998).

211. The yearly percentage was calculated by averaging the percentages for each decision within the year. If the yearly percentage was calculated by averaging the total number of relevant words in the opinions, then a few long opinions in a given year would skew the results. In other words, an opinion with a large word count would have a disproportionate influence on the overall results.

approximately twelve percent to approximately twenty-five percent. In contrast, during the same time period, claim construction increased in terms of the percentage of words in an opinion. From 1996 until 2000, claim construction comprised between forty and fifty-five percent of the opinions each year. Before *Markman I*, it encompassed between twenty and thirty percent of opinions. These results are statistically significant.[212]

To keep this result in the proper context, the total word count in the discussion section of opinions varied slightly over time. It increased approximately ten percent from the pre-*Markman I* period until *Festo*, and approximately another ten percent from *Festo* until the end of 2004.[213] So while the opinions increased in size, the increase was not substantial and most likely does not account for the change in word count of claim construction and the doctrine of equivalents around these events. And more importantly, the word count devoted to the doctrine of equivalents decreased while the count for claim construction increased.

This data supports the view that doctrinal displacement occurred after doctrinal reallocation and the doctrine of equivalents became less significant after *Markman*.[214] The drop in Figure 8 after *Markman* is significant. The trend continued after *Festo*, with the doctrine of equivalents becoming more marginalized, and claim construction more important.

However, patent litigation is complex and word count information cannot fully capture the significance of the doctrines. There may be multiple confounding factors a word count analysis cannot decode. For example, the data does not control for changes in complexity and difficulty in analyzing claim construction or the doctrine of equivalents, if any, over time. Nonetheless, the data is consistent with Allison and Lemley's narrative that

212. The results of a t-test suggest a statistically significant difference between the proportion of the word count devoted to claim construction before and after the *Markman* opinion had been issued (t=-3.9403, p=0.0001). The results of a t-test also suggest a statistically significant difference between the proportion of the word count devoted to the doctrine of equivalents before and after the *Markman* opinion (t=2.9810, p=0.0033). Because the distribution of the proportions of doctrine of equivalents word counts was not normal (in contrast to the proportions for claim construction, which were), two alternative statistical tests were performed. Both of these tests, a Wilcox-Mann-Whitney test (z=2.679, p=0.0074) and a general linear model (z=-2.89, p=0.004), provide the same result.

213. The average discussion section before *Markman I* was 3154 words, between *Markman I* and *Festo*, it was 3483 words, and after *Festo* until the end of 2004 it was 3876 words. That works out to a 10.4% increase after *Markman I* and an 11.3% increase after *Festo*.

214. Scholars have used word court as a proxy for measuring the importance of an issue in legal opinions. *See, e.g.*, sources cited *supra* note 210.

the doctrine of equivalents became less important as claim construction became more so.

Some may assert that when a doctrine becomes more important, it is natural for a court to temporarily increase the word count devoted to that doctrine.[215] The court in these circumstances needs to explain the enhanced doctrine to litigants and lower courts. This account does not completely explain the results of the present study. Figure 8 shows that the increase in word count for claim construction was not an aberration lasting only a few years. In fact, over time, claim construction occupied more and more decision space.[216] Similarly, some may argue that the increasing complexity of technologies and patents may explain the results.[217] However, the increasing complexity should affect both claim construction and the doctrine of equivalents.

This data, while only one way of analyzing the events, supports the view that the doctrine of equivalents dropped in importance as claim construction increased. The same trend can be analyzed by scrutinizing the opinions themselves. For example, in the illustrative 1994 case *Wolverine World Wide, Inc. v. Nike, Inc.*,[218] the Federal Circuit affirmed a district court's grant of summary judgment of non-infringement.[219] The court first considered the district court's claim construction of the term "forefoot-enveloping."[220] The court affirmed the claim construction with a three-paragraph discussion over the space of a page and a half.[221] The court then disposed of the patentee's literal infringement appeal in two paragraphs.[222] Finally, the Federal Circuit rejected the patentee's doctrine of equivalents appeal.[223] The doctrine of equivalents analysis, although only two paragraphs in length, was more thorough than the court's analysis on other issues. The detailed analysis examined two portions of the specification of the patent-in-suit and

215. The same can be said if the doctrine merely changes or becomes uncertain.

216. Furthermore, the variance in average word counts of the "Discussion" sections over time does not appear to cause the results. There is some variance in the yearly word count averages. The average word count of the "Discussion" sections of the opinions over this ten-year period was approximately 3300 words. The average for eight of the ten years fell within a relatively narrow band of the overall average, within twenty percent of 3300. The data for two years fell outside this band, 1994 being lower and 1998 being higher.

217. Allison & Lemley, *supra* note 24, at 79 (noting increased complexity in patents from the 1970s when compared to those from the 1990s).

218. 38 F.3d 1192 (Fed. Cir. 1994).

219. *Id.* at 1194.

220. *Id.* at 1196–98.

221. *Id.*

222. *Id.* at 1198–99.

223. *Id.* at 1199–2000.

compared how the patented invention operated with the accused products.[224] After that analysis, the court affirmed the district court's holding.[225]

In more recent cases, the doctrine of equivalents plays a lesser role, especially when compared to claim construction. For example, in *Welker Bearing v. PHD*, the Federal Circuit reviewed an opinion granting summary judgment of non-infringement.[226] After setting forth details about the patented technology and the district court proceedings, the opinion devoted nearly five pages to the issue of claim construction.[227] The term in dispute was "mechanism for moving said finger."[228] The court devoted nearly two pages discussing whether the claim was in means-plus-function format, eventually concluding that the claim included language in means-plus-function format.[229]

Thereafter, the *Welker Bearing* opinion delves into the details of the claim construction for three solid pages of analysis.[230] The Federal Circuit considered the patent specification, explicitly reciting and analyzing information provided in six locations in the patent specification.[231] It discussed the claim construction doctrines: claim differentiation, ordinary meaning and clear disavowal of claim scope.[232] Finally, the Federal Circuit affirmed the district court's claim construction.[233] As for literal infringement, the opinion contains two paragraphs disposing of the issue.[234] In those two paragraphs, the Federal Circuit affirmed the district court's conclusion of no literal infringement.[235]

Finally, the Federal Circuit reached the doctrine of equivalents.[236] In affirming the district court's grant of summary judgment, the Federal Circuit opinion included only two paragraphs relating to the doctrine of equivalents.[237] As described above, the Federal Circuit's current opinions focus much less on the doctrine of equivalents than pre-*Markman* opinions. The lack of Federal Circuit focus supports the declining importance of the

224. *Id.*
225. *Id.* at 2000.
226. Welker Bearing Co. v. PHD, Inc., 550 F.3d 1090 (Fed. Cir. 2008).
227. *Id.* at 1095–99.
228. *Id.* at 1095.
229. *Id.* at 1095–97.
230. *Id.* at 1097–99.
231. *Id.* at 1098–99.
232. *Id.* at 1099.
233. *Id.*
234. *Id.*
235. *Id.*
236. *Id.*
237. *Id.* at 1099–1100.

doctrine of equivalents. As discussed *infra*, this decline—whether the direct result of the increased prominence of the claim construction doctrine, substantive changes to the law of the doctrine of equivalents, or some combination of the two—is consistent with doctrinal reallocation.

This decline of the doctrine of equivalents is partially attributable to the case law's development of limitations on the doctrine of equivalents. The Supreme Court and the Federal Circuit decided several doctrine of equivalents cases.[238] These legal limits provided substantive changes in the doctrine. A substantive change can directly increase or decrease the importance of the doctrine. If a decrease in significance occurs, litigants may raise the doctrine less frequently.

Alternatively, these case law developments may be thought of as another doctrinal reallocation—moving part of the decision-making on the doctrine of equivalents from the jury to the judge. Shifting to judicial decision-making provides the court control to decide the importance of a doctrine. In contrast to claim construction, the Federal Circuit used its control to diminish the doctrine of equivalents. Petherbridge showed that the Federal Circuit reversed district court holdings of infringement under the doctrine of equivalents and affirmed district court rejections of such infringement.[239] In other words, the Federal Circuit used its institutional power to weaken the doctrine of equivalents after its doctrinal reallocation.

The conclusions reached by this study are consistent with the conclusions of Allison, Lemley, and Petherbridge. Federal Circuit opinions reduced emphasis on the doctrine of equivalents after *Markman*.[240] Patentees have little success on the doctrine of equivalents after *Markman*.[241] The doctrine of equivalents has lost power as claim construction increased in prominence and importance within patent law.[242] However, further study is needed on the exact timing of the decline of the doctrine of equivalents. It is still unclear how much of the decline followed *Warner-Jenkinson* and *Festo*, and how much already occurred before these decisions.

---

238. Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 535 U.S. 722 (2002); Warner-Jenkinson Co. v. Hilton Davis Chem. Co., 520 U.S. 17 (1997); Johnson & Johnston Assocs. v. R.E. Serv. Co., 285 F.3d 1046 (Fed. Cir. 2002) (en banc); Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 234 F.3d 558 (Fed. Cir. 2000) (en banc), *vacated*, 535 U.S. 722 (2002).

239. Petherbridge, *Doctrine of Equivalents*, *supra* note 2, at 1386–87, 1399.

240. *Id.* at 1394; Petherbridge, *Claim Construction Effect*, *supra* note 2, at 233.

241. Allison & Lemley, *supra* note 2, at 966–67; Petherbridge, *Doctrine of Equivalents*, *supra* note 2, at 1387.

242. Allison & Lemley, *supra* note 2, at 966–67; Petherbridge, *Claim Construction Effect*, *supra* note 2, at 233.

Claim construction opinions are of limited precedential value beyond interpreting the particular patent at issue. But others in the marketplace, such as competitors, are often interested in the construction of the terms of any litigated patent. Thus, appellate claim construction opinions are often valuable beyond the immediate parties to the litigation. In contrast, opinions on the doctrine of equivalents are not. The doctrine of equivalents analysis will be specific to each individual accused product. In most cases, the doctrine of equivalents analysis is not applicable to third parties. Consequently, it makes sense to devote more resources to claim construction information and opinions. That information is valuable to more entities than information on the doctrine of equivalents.

V. CONCLUSION

This Article provides a novel theoretical model and extensive empirical evidence to explain the decline of the doctrine of equivalents. In recent years, John Allison, Mark Lemley, and Lee Petherbridge studied the doctrine of equivalents. While these scholars noted and provided some evidence that the successful use of the doctrine of equivalents decreased, none clearly explained why. As such, the cause and precise mechanism behind the so-called "demise" of the doctrine of equivalents have largely remained a mystery.

This Article explains that the demise occurred because of two complementary forces discussed for the first time in this Article: doctrinal reallocation and doctrinal displacement. Under doctrinal reallocation, a substantive doctrine may become more important after a shift in adjudicative control over that doctrine. Doctrinal displacement posits that an increase in the importance of a doctrine may in turn decrease the importance of another, typically related, doctrine. This Article's empirical results support the position that the demise of the doctrine of equivalents was a result of these twin forces.

The study of doctrinal reallocation and doctrinal displacement in the law and its after effects is merely beginning. Until this study, it has never been formally discussed or empirically examined. The present study uses the theories of doctrinal reallocation and doctrinal displacement to explain the demise of the doctrine of equivalents. Further study is warranted to see whether the phenomena can explain other changes in the law, in areas within and beyond patent law. And it raises the further important question: is doctrinal displacement intentional? Does the court know ex ante that doctrinal reallocation likely leads to doctrinal displacement? If the court does, it has never expressly acknowledged it.

The general theories of doctrinal reallocation and displacement may allow hypotheses on how proposed procedural changes will affect existing doctrines. For example, some have argued to remove the doctrine of obviousness in patent law from the control of the jury.[243] What would be the likely consequences of such a change? Which doctrine would be displaced? Separately, what will happen if the pending patent reform bills[244] are passed by Congress? Displacement theory can help find the answers.

243. Petition for Writ of Certiorari, Medela AG v. Kinetic Concepts, Inc., 130 S. Ct. 624 (2009) (No. 09-198), 2009 U.S. S. Ct. Briefs LEXIS 806.

244. Patent Reform Act, S. 515, 111th Cong. (2009); Patent Reform Act, H.R. 1260, 111th Cong. (2009).

APPENDIX: DETAILED REGRESSION TABLES

Table 4a: Regression for Hypothesis #1: Written Opinions
(corresponding to Figure 4, page 1193)

Variable	Logistic Regression Odds Ratio (Std. Error)[245]	
Case decided before *Markman I*	2.861***	(.589)
District court in 2nd Circuit[246]	1.098	(.426)
District court in 3rd Circuit	1.560	(.605)
District court in 4th Circuit	1.618	(.720)
District court in 5th Circuit	0.940	(.368)
District court in 6th Circuit	1.468	(.582)
District court in 7th Circuit	1.198	(.441)
District court in 8th Circuit	1.502	(.615)
District court in 9th Circuit	1.274	(.437)
District court in 10th Circuit	1.300	(.612)
District court in 11th Circuit	0.731	(.320)
District court in one of 10 busiest patent courts	1.133	(.177)
Chemical Patent (PTO class)[247]	2.704***	(.569)
Electrical Patent (PTO class)	1.389	(.234)
Appeal from grant of summary judgment[248]	1.427	(.334)
Appeal from bench trial	1.366	(.302)
Appeal from jury trial	0.475**	(.135)
Patentee won at district court	1.309	(.249)
Pseudo R^2	.0569	
# Obs	1247	

245. *** Significant at the .001 level, ** Significant at the .01 level, * Significant at the .05 level, Standard errors in parentheses.

246. Base circuit is 1st Circuit.

247. Base technology is mechanical.

248. Base appeal is from preliminary injunction ruling.

Table 4b: Additional Logistic Regression for Hypothesis #1: Written Opinions
(corresponding to Figure 4, page 1193)

Variable	Logistic Regression Odds Ratio (Std. Error)[249]
Case decided before *Markman I*	1.959*** (.193)
Case involved Claim Construction	0.510*** (.043)
Pseudo R^2	.0245
# Obs	4234

Table 5a: Regression for Hypothesis #2: Precedential Opinions
(corresponding to Figure 5, page 1197)

Variable	Logistic Regression Odds Ratio (Std. Error)[250]
Case decided before *Markman I*	2.345*** (.490)
District court in 2nd Circuit[251]	1.595 (.541)
District court in 3rd Circuit	1.236 (.400)
District court in 4th Circuit	1.989 (.748)
District court in 5th Circuit	1.209 (.417)
District court in 6th Circuit	1.550 (.532)
District court in 7th Circuit	1.309 (.422)
District court in 8th Circuit	1.304 (.462)
District court in 9th Circuit	1.282 (.385)
District court in 10th Circuit	1.674 (.683)
District court in 11th Circuit	1.137 (.457)
District court in one of 10 busiest patent courts	1.089 (.148)
Chemical Patent (PTO class)[252]	2.784*** (.437)
Electrical Patent (PTO class)	1.729*** (.251)
Appeal from grant of summary judgment[253]	1.582* (.308)
Appeal from bench trial	1.358 (.243)
Appeal from jury trial	0.473** (.136)
Patentee won at district court	1.395* (.219)
Pseudo R^2	.0587
# Obs	1247

249. *** Significant at the .001 level, ** Significant at the .01 level, * Significant at the .05 level, Standard errors in parentheses.

250. *** Significant at the .001 level, ** Significant at the .01 level, * Significant at the .05 level, Standard errors in parentheses.

251. Base circuit is 1st Circuit.

252. Base technology is mechanical.

253. Base appeal is from preliminary injunction ruling.

Table 6a: Regression for Hypothesis #3: Summary Judgment
(corresponding to Figure 6, page 1199)

Variable	Logistic Regression Odds Ratio (Std. Error)[254]	
Case decided before *Cybor*	3.383***	(.537)
District court in 2nd Circuit[255]	1.136	(.428)
District court in 3rd Circuit	1.175	(.424)
District court in 4th Circuit	1.318	(.548)
District court in 5th Circuit	1.387	(.535)
District court in 6th Circuit	2.981**	(1.188)
District court in 7th Circuit	1.400	(.499)
District court in 8th Circuit	1.472	(.587)
District court in 9th Circuit	2.573**	(.866)
District court in 10th Circuit	1.230	(.561)
District court in 11th Circuit	1.203	(.543)
District court in one of 10 busiest patent courts	1.140	(.182)
Chemical Patent (PTO class)[256]	0.797	(.142)
Electrical Patent (PTO class)	1.438*	(.254)
Patentee won at district court	0.084***	(.013)
Pseudo R^2	.2437	
# Obs	1262	

254. *** Significant at the .001 level, ** Significant at the .01 level, * Significant at the .05 level, Standard errors in parentheses.

255. Base circuit is 1st Circuit.

256. Base technology is mechanical.

GETTING INTO THE "SPIRIT" OF INNOVATIVE THINGS: LOOKING TO COMPLEMENTARY AND SUBSTITUTE PROPERTIES TO SHAPE PATENT PROTECTION FOR IMPROVEMENTS

Kevin Emerson Collins[†]

TABLE OF CONTENTS

† Professor of Law, Washington University School of Law in St. Louis. I thank Scott Baker, T.J. Chiang, Mark Janis, Mark Lemley, Oskar Liivak, Mike Madison, Robert Merges, and Josh Sarnoff for their comments. I also thank the participants of the 2010 IPSC conference hosted by Berkeley Law (University of California).

I. INTRODUCTION

Technological progress is a cumulative endeavor, and how patent protection should be structured to most effectively promote technological progress in light of its cumulative nature is the subject of many distinct and robust debates.[1] One thread in these debates addresses the conditions under which the patent protection given to earlier inventors should be expansive enough to encompass later-developed improvements.[2] Patent applicants are

1. *See, e.g.*, *infra* notes 116 & 213 (discussing several patent doctrines that are controversial because they implicate the cumulative nature of technological progress).

2. The number of articles relevant to this topic is enormous, as nearly every discussion of claim scope implicates patent protection for improvements, at least tangentially. For articles that are focused on patent scope and improvement and that address the mechanics of patent law, see Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEX. L. REV. 989 (1997); Robert Merges, *Intellectual Property Rights and Bargaining Breakdown: The Case of Blocking Patents*, 62 TENN. L. REV. 75 (1994); Robert P.

clearly entitled to protection for the set of nonobvious things that they disclose in their patent applications, and thus make available to the public, at the time of filing.[3] In improvement scenarios, inventors acting at a later point in time make innovative changes to the things disclosed by earlier patentees, generating new things that the earlier patents did not make available to the public. The debate over patent protection for later-developed improvements addressed in this Article centers on the conditions under which initial inventors' patent rights should reach beyond the set of things that inventors actually make available to the public at the time they file their patents and into the improved things that are only made available to the public by subsequent inventors.[4]

This Article stirs the pot on patent protection improvements with a two-step argument. First, it demonstrates that there is a blind spot in the conventional theory on optimal patent protection for after-arising improvements.[5] This theory has been developed on the basis of only a subset of improvement cases (that this Article calls *classic*-improvement cases), and it has ignored another subset (that this Article calls *overlooked*-improvement cases) that lies in plain sight. Although many classic-improvement cases are

Merges & Richard R. Nelson, *On the Complex Economics of Patent Scope*, 90 COLUM. L. REV. 839, 909–11 (1990). For articles addressing patent scope and improvement from a purely economic perspective, see James Bessen & Eric Maskin, *Sequential Innovation, Patents, and Imitation*, 40 RAND J. ECON. 611 (2009); Howard F. Chang, *Patent Scope, Antitrust Policy, and Cumulative Innovation*, 26 RAND J. ECON. 34 (1995); Jerry R. Green & Suzanne Scotchmer, *On the Division of Profit in Sequential Innovation*, 26 RAND J. ECON. 20 (1995); Ted O'Donoghue et al., *Patent Breadth, Patent Life, and the Pace of Technological Progress*, 7 J. ECON. & MGMT. STRATEGY 2 (1998); Suzanne Scotchmer, *Standing on the Shoulders of Giants*, 5 J. ECON. PERSP. 29 (1991).

3. The set of things that a patent "discloses" and thus "makes available to the public" is used here as shorthand for the set of things that is both enabled for and possessed by the person having ordinary skill in the art. This set is smaller than the set of things that can fall within the scope of a claim that is valid under the enablement and written description requirements. *See infra* notes 130–33 and accompanying text.

4. A distinct, yet interconnected, debate about patent protection for improvements that this Article does not address centers on the nonobviousness requirement. *See* 35 U.S.C. § 103 (2006). Scholars involved in this debate often presume that earlier patentees can control later improvements and examine the conditions under which the later-acting improvers should be able to patent their improvements and obtain blocking patents. *See, e.g.*, Robert M. Hunt, *Patentability, Industry Structure, and Innovation*, 52 J. INDUS. ECON. 401 (2004); Ted O'Donoghue, *A Patentability Requirement for Sequential Innovation*, 29 RAND J. ECON. 654 (1998); Suzanne Scotchmer, *Protecting Early Innovators: Should Second-Generation Products Be Patentable?*, 27 RAND J. ECON. 322 (1996).

5. On a doctrinal level, the Article also identifies a flaw in a common understanding of the mechanism through which earlier patentees' rights can encompass later-developed improvements. *See infra* Part VIII (presenting and undermining the strong fixation theory of peripheral claims).

legitimately contested, the overlooked-improvement cases are all "easy" cases in which the literal scope of earlier-issued patents routinely and uncontroversially grows over time to encompass later-developed improvements.[6] Critically, the conventional theory that outlines when earlier patents should encompass later improvements cannot explain why overlooked-improvement cases are "easy" cases. A puzzle therefore emerges. Contemporary theory suggests that a set of common, everyday improvement cases should have one outcome (or at least that their outcomes should be highly contested), when, in practice, the cases routinely and uncontroversially yield the opposite outcome. Theory that fails to explain how law does, and should, operate is ripe for revision. The second step of this Article's argument therefore amends the conventional theory on patent protection for improvements, correcting its blind spot and reducing the explanatory gap between theory and practice. There is a to-date-hidden factor that reliably distinguishes the overlooked- and classic-improvement cases in which courts are employing the different rules. Furthermore, there is a convincing story about why this factor is a normative trump card that allows earlier-issued patents to encompass later-developed improvements as a matter of routine only in the overlooked-improvement cases.

Identifying the new factor requires some conceptual heavy lifting because seeing it entails a shift in one of the most basic conceptual frameworks, or paradigms, structuring contemporary understandings of patent protection.[7] Today, the inventions that give rise to patent rights are only identified as the sets of innovative *things* that an inventor discloses, and these innovative things are taken to be the primitives of what an inventor invents. This conceptual framework is a coarse-grained framework that blinds contemporary patent scholarship to the factor that differentiates overlooked- and classic-improvement cases. What is needed to see this factor is a finer-grained framework that gets into the "spirit" of innovative things—i.e., that recognizes a "spirit" of inventions that is somehow embodied in particular

6. The cases are "easy" because the outcome is routine and uncontroversial. The term "easy" remains in quotation marks throughout the Article because, prior to this Article, there was no coherent explanation of the outcome in logical, doctrinal terms. *Cf.* Mark Kelman, *Interpretive Construction in the Substantive Criminal Law*, 33 STAN. L. REV. 591, 662–69 (1981) (discussing the important role of the interpretive construction of facts by courts in criminal cases that are "easy" only in the sense of their outcomes being uncontroversial in the relevant legal community).

7. The mention of paradigms invokes Kuhn. *See infra* notes 337–40 and accompanying text (arguing that the overlooked cases have been overlooked because the focus on properties as the locus of invention that is required to identify them does not fit with the dominant conceptual paradigms of contemporary patent theory).

features of the things that are disclosed and claimed by the earlier patentee.[8] To structure this finer-grained analysis, this Article appropriates the metaphysical concept of the innovative *properties of things*. A property is an intuitive and familiar entity. It is simply "what is variously called a feature, quality, attribute or . . . a way that something is,"[9] and things, in turn, can be identified by the sum total of the properties that they possess.[10] Paying attention to innovative properties in improvement cases requires sustained effort, as it runs against the grain of contemporary patent discourse. However, the payoff is worth the effort. This Article proves the pragmatic value of identifying the innovative properties of things, rather than innovative things in their entireties, as the locus of invention when assessing the optimal reach of earlier-filed patents into later-developed improvements. A focus on innovative properties corrects the blind spot in the contemporary theory on patent protection for improvements, reducing the gap between the theory and reality of contemporary patent practice by revising the theory to create a better fit with reality. It is necessary to identify, explain, and justify an uncontroversial and desirable feature of real-world patent rights that already exists today, namely the different treatment afforded to classic and overlooked improvements.[11] In brief, the descriptive insight enabled by a focus on properties is that the properties invented by successive generations of inventors in improvement scenarios can relate to each other differently in different types of improvements. Classic improvements involve later improvers who *refine* the properties invented by the earlier patentees: the more general of the earlier inventor's innovative properties persist in the

8. The proposal to pay attention to the "spirit" of an invention has a clear conceptual kinship with the point of novelty approach to patent law that is often denigrated by both scholars and courts. *See infra* notes 64–68 (discussing the point of novelty approach). However, this Article develops an argument about the role that the point of novelty should, and already does, play in patent law in a different direction than previous scholarship has taken it. *See infra* notes 69, 318 (comparing this Article's focus on properties as the locus of invention in improvement cases with recent scholarship that addresses a point of novelty approach or central claiming).

9. Chris Daly, *Properties*, *in* ROUTLEDGE ENCYCLOPEDIA OF PHILOSOPHY (E. Craig ed., 1998) (discussing disagreements among philosophers about the nature and existence of properties). This definition of property conveys a metaphysical concept, and it is entirely distinct from the legal concept of property—whatever that legal concept is. *Cf.* Thomas C. Grey, *The Disintegration of Property*, *in* LIBERTY, PROPERTY AND THE LAW (Richard A. Epstein ed., 2000) (discussing the difficulty of defining property as a legal concept).

10. *See infra* note 72.

11. Another effect of a focus on innovative properties is the refinement of judicial and scholarly analysis of optimal protection for classic improvements. *See infra* notes 242–48 and accompanying text (introducing the concept of the "least-general naked property" of a classic improvement).

improvement, but the more specific of his innovative properties are supplanted by the improver's innovative properties. In contrast, overlooked improvements involve later improvers who invent new properties that simply *compound with* the properties invented by the earlier patentees in the improved thing: the earlier and later innovators' innovative properties both are fully present in the improvement.

A focus on properties, rather than things in their entireties, as the locus of invention in improvement cases also explains why courts should treat classic- and overlooked-improvement cases differently, as they already do. (Again, paying attention to innovative properties in improvement scenarios does not mandate a radical shift in the on-the-ground scope of contemporary patent protection. It alters patent theory so that it is better able to account for a desirable feature of the contemporary patent regime.) One of the core principles of patent law is that inventors should be rewarded in proportion to the value of their inventions. When inventions are defined in terms of innovative properties, this proportionality principle mandates differential treatment for classic and overlooked improvements. In a classic improvement, the persistence of some, but not all, of the earlier innovator's properties in the improvement gives the earlier inventor a weaker normative claim to rights that encompass the improvement. In contrast, in an overlooked improvement, the persistence of all of the earlier innovator's properties in the improvement gives the inventor a very strong normative claim to rights that encompass the improvement. Thus, when properties are viewed as the locus of invention, the contested nature of the overlooked improvement cases and the "easy" nature of the classic-improvement cases both make good economic sense.

Perhaps the most compelling payoff of identifying properties as the locus of invention is that, for the first time, the economic concepts of complements and substitutes can be brought to bear on the crafting of optimal claim scope in improvement cases. Today, patent scholarship on improvement uses the concepts of complements and substitutes in a variety of ways. However, these concepts have no relevance to the crafting of optimal claim scope in general or the differential reach of earlier-filed claims into classic and overlooked improvements in particular. This Article demonstrates that the concepts of complements and substitutes can be turned into useful tools for crafting optimal patent scope if, and only if, innovative properties are identified as the goods that are either complements or substitutes. Overlooked improvements result from successively-invented, complementary properties that are instantiated in the self-same thing, and the normative claim of earlier inventors to rights that encompass later-developed improvements is at its strongest when the successively-invented properties

are pure complements. Therefore, the scope of earlier-filed claims should routinely be construed so as to encompass later-developed, overlooked improvements. In contrast, classic improvements result from successively invented properties of the self-same thing that are complement-substitute mixtures, and the normative claim of earlier inventors grows weaker as the substitute properties come to predominate the mixture. Therefore, the reach of earlier-filed claims into classic improvements should remain a contested issue.[12] In sum, by understanding properties as the locus of invention in improvement cases, the legal doctrine that courts use to craft claim scope can be revolutionized to incorporate the economic concepts of complements and substitutes.

On a more theoretical level, a fine-grained focus on the innovative properties of things is a fruitful addition to patent theory because it provides a link in the contemporary conceptualization of patent rights that heretofore has been missing. Contemporary patent theory is replete with discussions of innovative ideas, but the role of innovative ideas in the patent regime is ambiguous. To reflect the fact that patent claims are not limited to the set of things that a patentee makes available to the public and that they should grow over time to encompass improvements, patents are often characterized as granting property rights in ideas.[13] Yet, at the same time, ideas per se are clearly beyond the reach of patent protection.[14] This paradox can be resolved by understanding that patent claims describe and propertize sets of things, not ideas themselves, but that the scope of the set of things that can be claimed is defined roughly as the set of things that embody an inventor's innovative ideas in a sufficiently important or prominent way. It would be helpful to be able to talk about patent protection in a manner that is not paradoxical and that reserves a semantic space for the important facts that ideas per se are not patentable and that later innovators can often freely appropriate the knowledge generated and disclosed by earlier patentees. What is needed to clear the air is a concept that provides the missing link between innovative ideas and innovative things and that captures how ideas are embodied in patented things. The notion that claimed things have a

12. Because the framework positions properties of things, rather than things in their entireties, as the relevant goods, the import of a later-developed complement or substitute invention is the opposite of the import of a later-developed complement or substitute work in the fair use analysis of copyright. *See infra* note 303.

13. *See, e.g.*, Tun-Jen Chiang, *The Levels of Abstraction Problem in Patent Law*, 105 NW. U. L. REV. (forthcoming 2011) (discussing claims to the idea of curing AIDS at different levels of generality).

14. *See infra* notes 117–18 and accompanying text (discussing idea-only cumulative innovation cases).

"spirit" represents one attempt to provide this missing link, however vague or other-worldly it may seem. The notion of the innovative properties of things is another candidate for the missing link, one that is both more capable of a precise formulation and more firmly rooted in the material world of infringing things than the notion of an inventive "spirit" of things ever could be. The properties of things can be framed as the entities that instantiate ideas in things; things that possess certain properties can be said to embody certain ideas. In sum, a fine-grained focus on the innovative properties of things, rather than a coarse-grained focus on innovative things in their entireties not only helps to solve pragmatic problems in the administration of patent rights (e.g., to factually differentiate the classic and overlooked improvements that courts are already treating differently as well as to normatively justify why these two types of improvements should continue to receive this differential treatment), but it also provides a missing link between ideas and things in the concepts that structure contemporary patent theory.

This Article proceeds in six substantive parts, a coda, and a conclusion. Taking the form of an extended introduction, Part II lays out the problem that motivates this Article and sketches a solution to this problem. It reviews the contemporary theory on patent protection for improvements and points out its blind spot, illustrating the insights that the theory yields when classic improvements are at issue but also highlighting the inapplicability of the theory to overlooked improvements. Part II also offers a high-level overview of how this blind spot can be corrected. What is needed is a shift in the conceptual framework that is used to understand the nature of invention. Properties of things, rather than things in their entireties, must be viewed as the locus of invention in improvement cases. For clarity, Part III defines an "improvement" as the term is employed in this Article.

The following three parts describe the difference between classic and overlooked improvements and demonstrate that overlooked improvements are already routinely treated as "easy" cases in the courts. Part IV defines a classic improvement in terms of successively invented properties: the later-invented property refines the earlier-invented property. Part V does the same for an overlooked improvement: the later-invented property simply compounds with the earlier-invented property in the improved thing. It also illustrates three categories of overlooked-improvement cases in which courts routinely allow earlier-issued patents to encompass later-developed improvements. Part VI reinforces the distinction developed in the previous two Parts, representing it in visual form.

Part VII is the heart of the normative argument. It explains why the principle of proportionality of contribution and reward counsels that classic

improvements should result in the contested infringement cases that they do (in which the conventional theory on improvements gains traction) and why overlooked improvements should result in the "easy" infringement cases that they do (in which the conventional theory is irrelevant). It also illustrates how the economic concepts of complements and substitutes can be used to explain how claim scope is, and should be, crafted in improvement cases if, and only if, properties are identified as the locus of the inventions created by successive inventors.

Part VIII, a coda, takes a step back and briefly considers the implications of the arguments presented in this Article for the peripheral claiming regime. Principally, it is not the on-the-ground scope of peripheral claims that must change. It is rather what we understand peripheral claims to be that must change. The dominant, thing-centric conceptual framework of what constitutes an invention is often defended with the argument that it is a necessary artifact of the contemporary peripheral claiming regime. This argument has no merit. A shift to a focus on properties as the locus of invention in improvement cases is entirely compatible with a peripheral claiming regime. However, this shift does require a concomitant shift in common understandings of what peripheral claims already are and how they already operate. Once again, a paradigm shift in the theory of peripheral claims is needed if this theory is to map onto the reality of peripheral claiming. Part IX concludes.

II. THE PUZZLE AND A QUICK SKETCH OF ITS RESOLUTION

This Part offers an extended introduction. It both discusses a shortcoming in contemporary patent theory on improvements and offers an overview of how to remedy this shortcoming. Section II.A introduces the conventional theory on patent protection for improvements. It identifies factors that are thought to increase or decrease the strength of a patent owner's normative argument for control over a later-developed improvement. Section II.B argues that the conventional theory has a blind spot because it gains traction in explaining the optimal outcomes of some improvement cases (what this Article calls classic-improvement cases) but not others (what this Article calls overlooked-improvement cases). Section II.C provides a high-level overview of the change to the conventional theory that is needed to correct the blind spot. It proposes a shift from an exclusive focus on things in their entireties to the properties of things as the locus of invention in improvement cases.

A. THE CONVENTIONAL THEORY ON PATENT PROTECTION FOR IMPROVEMENTS

In its most basic formulation, the incentive-to-invent justification of patent rights involves free-riders or pirates. By preventing later actors from copying the very technologies that earlier actors invent, patent rights increase the incentives for the earlier actors to invest in the invention and commercialization of technology.[15] Patent protection for improvements adds a significant wrinkle to this justification. When the issue is the reach of earlier-filed patents into later-invented improvements, there are inventors on both sides of the infringement suit.[16] A simple policy of favoring inventors over free-riders is not dispositive. Instead, there are four distinct factors that are conventionally considered in determining the optimal reach of earlier-filed patents into later-developed improvements.[17]

First, the relative importance of the earlier and later inventions should affect the reach of patent protection into improvements. Whether earlier-filed patents should encompass later-developed improvements determines whether improvers owe a portion of their profits to earlier inventors, and it therefore affects the inter-generational division of the rewards of patent protection.[18] To reinforce the proportionality of contribution and reward in

15. *But cf.* Christopher A. Cotropia & Mark A. Lemley, *Copying in Patent Law*, 87 N.C. L. REV. 1421 (2009) (arguing that only a small number of patent cases even involve allegations of copying). For a brief overview of the incentive-to-invent justification of patent rights, see *infra* notes 196–98 and accompanying text.

16. If the same firm generates both the earlier invention and the later improvement, then many of the issues raised in the conventional theory on patent protection for improvements are moot. However, the first and second generation inventors in an improvement scenario will often not be the same party because the earlier inventor is unlikely to be able to keep the information about the earlier invention secret (particularly while simultaneously obtaining patent protection and commercializing it) and different potential improvers likely possess specialized information and capacities. *See* Bessen & Maskin, *supra* note 2, at 620.

17. The normative debates discussed in the remainder of this Section assume that patent protection for later-developed improvements is sometimes desirable, and they focus on when and how much protection for later-developed improvements there should be. As a doctrinal matter, however, a more basic question is still the subject of considerable controversy. Judicial rhetoric in patent cases and scholarship often bolsters a strong fixation theory of literal claim scope under which the literal scope of a patent claim can never grow over time in the manner that is needed to encompass later-developed improvements. *See infra* notes 308–11 and accompanying text.

18. Green & Scotchmer, *supra* note 2, at 21; Scotchmer, *supra* note 2, at 30. Importantly, this inter-generational distribution of profit is not a "mere" distributional issue of the type that is often overlooked in discussions of static efficiency and tangible property regimes. *See* Guido Calabresi & A. Douglas Melamed, *Property Rules, Liability Rules, and Inalienability: One View of the Cathedral*, 85 HARV. L. REV. 1089, 1098–101 (1972) (segregating

patent law,[19] everyday improvements that yield moderate increases in social value should fall within the rights of earlier patentees, but radical improvements that generate large increases in social value should not.[20] Inversely, the greater the social welfare increase attributable to the invention disclosed in the earlier-issued patent, the farther the patent should reach into later-developed improvements.[21]

Second, the nature of the industry in which the improvements occur is viewed as relevant to patent protection for improvements, at least at the margin. Robert Merges and Richard Nelson famously argue that the reach of earlier patents into later-developed improvements should be scrutinized more carefully in industries in which technical advance proceeds in a "cumulative" rather than a "discrete" fashion and in which skepticism about a frictionless market for patent licenses is warranted.[22] Here, the concern is less about ensuring that inventors are rewarded in proportion to their contributions and more about preventing bargaining impasses in which later inventors are unable to acquire the rights from earlier inventors that are needed to continue the inventive process in a competitive fashion.[23]

A third policy concern implicated in patent protection for later-developed improvements addresses the magnitude of patent rents that are available for distribution among successive inventors. Economists interested in cumulative innovation sometimes work with a "quality ladder" as a stylized model of improvement.[24] The quality-ladder model assumes that each

distributional concerns from efficiency concerns). The reward from invention creates incentives, so the distribution of the reward among different generations of inventors is a dynamic efficiency issue under an incentive-to-invent justification of patent rights. Scotchmer, *supra* note 2, at 30.

19. *See* Lemley, *supra* note 2, at 1073 (arguing in the context of patent protection for improvements that "efficiency is best served by some sort of calibration, however rough, between the importance of the invention and the scope of the patent"); *infra* Section VII.A (addressing the policy basis of this proportionality).

20. This argument is often made to justify patent law's reverse doctrine of equivalents. *See* Lemley, *supra* note 2, at 1008–13 (distinguishing between the rights of "significant" and "radical" improvers as a descriptive matter); *id.* at 1065, 1070 (defending this distinction as a normative matter); Merges & Nelson, *supra* note 2, at 909–11 (arguing that earlier inventors should not have rights to hold up radical improvers). The reverse doctrine of equivalents, and its placement of radical improvers beyond the reach of earlier patentees, has also been defended as a means of preventing bargaining breakdown between the owners of blocking patents. Merges, *supra* note 2, at 91–102.

21. Ensuring adequate rewards for especially important inventions underlies pioneer theory in patent law. *See* Lemley, *supra* note 2, at 1072–73.

22. Merges & Nelson, *supra* note 2, at 880–908.

23. *Id.*

24. SUZANNE SCOTCHMER, INNOVATION AND INCENTIVES 149–52 (2004); O'Donoghue et al., *supra* note 2, at 5–7.

successive innovating firm produces an improvement before the expiration of the earlier patent that is of higher quality than the previous one. The "leading breadth" of a patent—that is, the reach of patent scope into improvements—is measured in quality increments: an earlier patent encompasses only the improvements that do not exceed a certain quantum of increase in quality. Under this model, the optimal reach of a patent into improvement depends on the desired magnitude of the overall incentive to innovate to be created by the patent regime. The deeper patents reach into improvements, the larger the supra-competitive profits that are created. Assume an increase from X to 2X in the quality increment that defines the leading breadth of a patent. Making the simplifying assumption that improvements continue to arrive at the same rate, a firm must now share its profits in any given time period through a licensing agreement with a larger number of other firms, but this loss is balanced by the gain that comes from a patent covering the highest quality product (whoever produced it) during a time period that is twice as long. The difference maker is that the quality difference between the highest-quality, patented product and its closest unpatented substitute is twice as great, so the patentees' collective per-period profits will be larger.[25] Thus, under a quality-ladder model of improvement, the reach of a patent into improvements should be greater when the optimal strength of patent-induced incentives is higher,[26] as, for example, would be the case if the sunk costs of innovation in an industry were larger than in other industries.

A fourth policy concern affecting the optimal reach of patents into later-developed improvements is the importance attributed to the prospect function of patent rights. The prospect theory of patent law suggests that patent protection should create incentives to prevent the wasteful duplication that results when the development of nascent technologies into marketable products occurs in an uncoordinated fashion.[27] The more important one believes the prospect function of patent law to be, the deeper the reach of patent protection into later-developed improvements should be.[28]

---

25. *See infra* notes 272–75 and accompanying text (discussing the impact of unpatented substitutes on the private value of a patent).

26. SCOTCHMER, *supra* note 24, at 134, 149–52.

27. The canonical presentation of prospect theory is Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265 (1977), although Kitch's argument is more descriptive than normative.

28. The four factors addressed in the text do not form an exhaustive list of the conventional wisdom that could be brought to bear on patent protection for improvements. For example, James Bessen and Michael Meurer argue that patent claims that reach far beyond the set of things disclosed in a patent and far into later-developed improvements are

B. THE BLIND SPOT

In many improvement cases—cases that this Article refers to as classic-improvement cases—the conventional theory on patent protection for improvement seems to gain traction in explaining how infringement allegations against improvers should be resolved. For an example of a classic improvement, consider the aviation industry just after the turn of the twentieth century, and a stylized telling of Glen Curtiss's improvement on the Wright Brothers' patented invention.[29] The Wright Brothers realized that an airplane could be stabilized by raising and lowering different portions of the wing surface at the same time. They disclosed airplanes that could perform these simultaneous adjustments because the wings had a flexible frame that allowed the entire surface of the wing to be warped.[30] Curtiss then borrowed from the Wright Brothers the notion of an airplane that could be stabilized by raising and lowering different parts of the wing at the same time. In doing so, Curtiss produced an improved airplane that the Wright Brothers had not themselves invented or disclosed in their patent. He invented wings with ailerons—discrete flaps that could move independently of the rest of the wing—that supplanted wing-warping technology.[31] The successive inventions of Curtiss and the Wright Brothers exemplify the type of improvement scenario that is commonly addressed by the improvement debate. In such scenarios, the improvement debate focuses on whether the set of things within the scope of the Wright Brothers' patent claim can grow over time, extending beyond the set of things actually disclosed by the patent (airplanes that use wing-warping technology) and into later-developed improvements (airplanes that use ailerons).[32]

In the context of the Wright Brothers/Curtiss patent infringement suit, the factors addressed in the conventional theory on patent protection for

inherently more "abstract," have fuzzier boundaries, and therefore entail higher social costs. JAMES BESSEN & MICHAEL J. MEURER, PATENT FAILURE: HOW JUDGES, BUREAUCRATS, AND LAYWERS PUT INNOVATORS AT RISK 187–214 (2008) (arguing that abstract software claims—i.e., software claims reaching far into after-arising technologies—provide poor public notice). This argument highlights a cost in the form of poor public notice of earlier-filed patents that reach deep into later-developed improvements.

29. For a fuller historical account of Curtiss's work and its relationship to the Wright Brothers' work, see SETH SHULMAN, UNLOCKING THE SKY: GLEN HAMMOND CURTISS AND THE RACE TO INVENT THE AIRPLANE (2002).

30. U.S. Patent No. 821,393 (filed Mar. 23, 1903).

31. *See* SHULMAN, *supra* note 29, at 133–34 (discussing the historical development of ailerons).

32. The Wright Brothers' patent fight with Curtiss on this precise point is legendary. For one judicial opinion in this dispute, see Wright Co. v. Herring-Curtiss Co., 204 F. 597, 614 (W.D.N.Y. 1913), *aff'd*, 211 F. 654 (2d Cir. 1914).

improvements make sense.[33] That is, there is good reason to believe that they should be relevant to the outcome if patent protection is to optimally promote cumulative innovation, even if any given factor is not dispositive. First, the more significant the Wright Brothers' invention in terms of the social value that it creates, and the smaller the additional increment of social value contributed by Curtiss, the stronger the case for allowing the Wright Brothers' patent to encompass airplanes stabilized with ailerons.[34] Second, the more cumulative the pattern of technological advance and the less confidence inspired by the market for patent licenses, the stronger the case for allowing Curtiss to be free of the Wright Brothers' patent.[35] Third, the more costly the process of creating and commercializing the innovations in the airplane industry, the stronger the need to reduce competition and augment the monopoly power attributable to the patent regime, and thus the deeper earlier patents should reach into later-developed improvements.[36] Fourth, the more prominence one gives to the prospect function of patents, the more certain one becomes that the Wright Brothers' patent should encompass Curtiss's improved airplane.[37]

While the conventional theory has traction when classic improvements are at issue,[38] all improvements are not classic improvements. There is a distinct set of improvement cases—cases that this Article refers to as overlooked-improvement cases—that have been largely ignored in patent scholarship. Overlooked-improvement cases are "easy" cases in the sense that their outcomes are routine and uncontroversial: earlier inventors' patents expand to encompass the later-developed improvements.[39] Most importantly for the argument here, the conventional theory has little, if any, purchase in overlooked-improvement cases. The factors on which the conventional theory focuses are simply unable to explain how overlooked improvement are (and should be) resolved.

For a simple example of an "easy" overlooked-improvement case, consider a hypothetical improvement on the Wright Brothers' patented technology. Imagine that a later inventor makes an unexpected advance in

33. *See supra* Section II.A (listing the factors considered in the conventional theory).

34. *Cf. supra* notes 18–21 and accompanying text.

35. *Cf. supra* notes 22–23 and accompanying text.

36. *Cf. supra* notes 24–26 and accompanying text.

37. *Cf. supra* notes 27–28 and accompanying text.

38. Although the factors considered in the conventional theory are all relevant in classic-improvement cases, a focus on properties as the locus of invention reveals an additional factor that is also relevant. *See infra* notes 242–48 and accompanying text (introducing the concept of a classic improvement's least-general naked property).

39. *See supra* note 6 (defining an "easy" case).

the art of canvas-making, and that he develops a much-improved, revolutionary, and unforeseen canvas that can be stretched over the flexible frame of a wing-warping airplane. In relation to the Wright Brothers' patent, a wing-warping airplane that uses the improved canvas is clearly an improvement. A later inventor has made a change to the things that were earlier disclosed by the Wright Brothers, generating new things that were not made available to the public at the time the Wright Brothers' patent was filed.[40] Yet, the Wright Brothers' infringement suit would be an "easy" case. It would be "easy" in the sense that a court would clearly rule in favor of the Wright Brothers, as the literal scope of a patent in the mechanical arts routinely grows over time to encompass devices made out of after-arising materials.[41] It would also be "easy" in the sense that this doctrinal rule has proven uncontroversial in the patent community. When after-arising material cases in the mechanical arts are recognized as improvement cases at all (and they frequently are not recognized as such), the uncontroversial outcomes are accepted as being in line with common sense: a patent on a new mechanism for a doorknob would reach into doorknobs made of any and all after-arising materials because, well, "a doorknob is a doorknob."[42]

Critically, the conventional theory on the reach of patents into improvements cannot explain why the overlooked cases are "easy" cases. To illustrate this point, tweak the facts of the hypothetical improved-canvas airplane scenario so as to make every factor listed in the conventional theory weigh against allowing the earlier-filed patents to encompass the later-developed improvements.[43] First, adjust the relative importance of the inventions. Assume that the Wright Brothers' patent on wing-warping

40. The Wright Brothers claimed "a normally flat aeroplane" (with "aeroplane" meaning a wing surface) that could be warped. U.S. Patent No. 821,393 (filed Mar. 23, 1903). The improved-canvas wing is an improvement, as the term is used in this Article, because the set of distinct things described by the claim must therefore grow over time after the claim is filed if the claim is to encompass the improved-canvas wing. *See infra* Section III.A.2 (presenting the new-thing criterion of an improvement). If the Wright Brothers' patent had claimed only the frame of an airplane wing, then the combination of the claimed frame and an improved canvas would not be an improvement because the frame in the improved-canvas wing is the same frame that would have been disclosed and claimed by the Wright Brothers. *See infra* notes 110–16 and accompanying text (noting that disclosed-thing cumulative innovation cases are not improvement cases).

41. Robin Feldman, *Rethinking Rights in Biospace*, 79 S. CAL. L. REV. 1, 28 (2005); Michael J. Meurer & Craig Allen Nard, *Invention, Refinement and Patent Claim Scope: A New Perspective on the Doctrine of Equivalents*, 93 GEO. L.J. 1947, 1976–77 (2005).

42. Feldman, *supra* note 41, at 3.

43. *See supra* Section II.A (listing the factors considered in the conventional theory).

technology is a minor improvement over the prior art,[44] and that the improved canvas is important in that it completely revolutionizes the industrial fabric industry.[45] Second, assume that the airplane industry is an industry characterized by cumulative technological advance and problematic markets for patent licenses.[46] Third, assume that invention and commercialization in the airplane industry require little, if any, sunk costs. Fourth, reject the prospect function of patents.[47] The irrelevance of the conventional theory is put on full display because including the assumptions in the hypothetical does not derail the intuition that, well, a wing-warping airplane is a wing-warping airplane, regardless of the canvas employed,[48] and that the Wright Brothers' patent should encompass the improved-canvas airplane.[49] Even if patent protection were to be trimmed back to something resembling a minimalist core, it is hard to imagine a viable patent regime in which the improved-canvas airplane is not within the scope of the Wright Brothers' patent. A case that the conventional theory suggests should be resolved by preventing the earlier-filed patent from encompassing the improvement (or, at the least, should be controversial) is in fact an "easy" case in which the earlier-filed patent does encompass the improvement.

For another example of an overlooked, "easy" improvement case that illustrates the blind spot in the conventional theory on patent protection for later-developed improvements, consider a functionally defined software patent with an apparatus claim.[50] A classic-improvement case might involve an allegedly infringing apparatus programmed with improved software that performs the claimed functions in a manner that is more efficient than the

44. For example, assume counterfactually that someone before the Wright Brothers had figured out the trick to stabilizing an airplane by simultaneously raising and lowering wing surfaces, that the Wright Brothers only invented wing-warping as a means of achieving this end, and that wing-warping was a less valuable technology than the prior art.

45. For example, assume that wing-warping airplanes become commercially viable only when the new, lighter canvas that is strong enough to withstand the wear of constantly being stretched as a wing flexes becomes available at a reasonable cost.

46. These assumptions are not far-fetched. *See* Merges & Nelson, *supra* note 2, at 890–91 (discussing the airplane industry at the turn of the twentieth century as an industry characterized by cumulative technical advance); *cf.* SHULMAN, *supra* note 29, at 169–85 (discussing the patent fights in the airplane industry before the pooling of patents during World War I).

47. Furthermore, allowing the Wright Brothers' patent to encompass the improved-canvas airplane does not seem to entail any greater fuzziness in the meaning of an "aeroplane," so the concern that patents that reach into later-developed improvements are somehow more "abstract" does not apply to overlooked improvements. *See supra* note 28.

48. *Cf.* Feldman, *supra* note 41, at 3.

49. *Cf. infra* Part VII (outlining an economic justification of this position).

50. *See infra* Section V.B.2 (discussing software improvements at greater length).

manner disclosed in the software patent. Here, the conventional theory on patent protection for improvements has merit.[51] Now, assume a different improvement on the same software patent. Assume that the allegedly infringing apparatus is software that performs the specified functions in the precise manner that is disclosed in the patent, but the hardware on which the program runs is after-arising hardware. Like the improved-canvas airplane, the improved apparatus that consists of after-arising hardware is an improvement in the sense that the set of things within the earlier-issued patent must grow over time for literal infringement to lie.[52] Nonetheless, it would be an "easy" case for the courts: software apparatus claims routinely grow over time to describe the identical software running on after-arising hardware.

Furthermore, the case would remain an "easy" case even if one makes all of the assumptions that, under the conventional theory, would support preventing the earlier-filed patent from encompassing the later-developed improvement.[53] In other words, assume that the early software advance is a minor advance and the later hardware advance is a major advance, that the software industry progresses through a cumulative pattern of technical advance and that markets for patent licenses are full of friction,[54] that sunk costs are low in the fields of computer-related technology, and that the prospect function of patents is not important. Again, the intuition that computer software is computer software, regardless of the hardware on which it is running, is strong.[55] Despite the fact that many generations of hardware improvements occur during the twenty-year term of a software patent, no court has ever held that software running on after-arising

51. *See supra* Section II.A (discussing the factors considered in the conventional theory). In brief, the earlier patentee's assertion of rights to exclude others from the improvement should be more carefully scrutinized as the improved software becomes more important in relation to the software disclosed in the patent, the nature of technological advance in the industry becomes less cumulative, the sunk costs of invention in the industry become smaller, and the prospect function of patents becomes less important.

52. *See infra* Section III.A.2 (presenting the new-thing criterion of an improvement).

53. *See supra* Section II.A (discussing the factors considered in the conventional theory).

54. The software industry is already widely viewed as an industry in which technical advance is frequently cumulative. Bessen & Maskin, *supra* note 2, at 612; Julie E. Cohen & Mark A. Lemley, *Patent Scope and Innovation in the Software Industry*, 89 CALIF. L. REV. 1 (2001); Pamela Samuelson et al., *A Manifesto Concerning the Legal Protection of Computer Programs*, 94 COLUM. L. REV. 2308 (1994).

55. *Cf.* Feldman, *supra* note 41, at 3. Furthermore, allowing the earlier-filed claim to encompass the later-developed improvement does not entail any greater fuzziness in the meaning of an "apparatus," so the concern that patents that reach into later-developed improvements are somehow more "abstract" does not apply to overlooked improvements. *See supra* note 28.

hardware is a non-infringing improvement.[56] Software patents as a category are controversial for some commentators.[57] However, assuming that patent protection for software exists as a categorical matter, the fact that an earlier-filed software claim should encompass the software executed on after-arising hardware is anything but controversial. It is difficult to imagine a viable patent regime in which software apparatus claims, if they are permitted, would not encompass the disclosed software running on later-developed hardware.

There is nothing odd or unusual about these examples of overlooked improvements. They are not once-in-a-blue-moon events. Rather, they involve ordinary, work-a-day occurrences. Yet, the scholarly literature on patent protection for improvements has more or less ignored them,[58] and they demonstrate the incompleteness of the contemporary theory on patent protection for improvements. To the extent that overlooked improvements have avoided sustained analytical attention, they have simply been hiding in plain sight.[59] Once the blind spot is brought to our attention and the "easy" cases are acknowledged as facts about the reality of contemporary patent protection that must be accounted for, a puzzle arises. There is a radical disconnect between an everyday, uncontroversial practice of treating

56. *See infra* notes 175–77 and accompanying text.

57. *See, e.g.*, BESSEN & MEURER, *supra* note 28, at 187–214.

58. One common theme in both the doctrine and scholarship relating to the disclosure doctrines is that broader claims are permitted when the claimed technology is more predictable. *See, e.g.*, *In re* Wands, 858 F.2d 731, 737 (Fed. Cir. 1988); Sean B. Seymore, *Heightened Enablement in the Unpredictable Arts*, 56 UCLA L. REV. 127 (2008). To the extent that the mechanical arts are viewed as predictable, this theme might seem to explain the outcome of the overlooked improvement cases, like the improved-canvas airplane. The argument would be that the art is predictable, so the claims in the mechanical arts are allowed to reach deep into later-developed improvements. However, predictability is a red herring in any attempt to explain the optimal reach of earlier-filed patents into after-arising technology. Kevin Emerson Collins, *Enabling After-Arising Technology*, 34 J. CORP. L. 1083, 1094–98 (2009) (discussing the limitations of using predictability to determine the reach of patent protection into after-arising technology). Not all overlooked improvements are predictable. The later-developed material may have been highly unexpected (and thus not predicted), and yet the earlier-filed mechanical patent will encompass devices made from it. Furthermore, predictability cannot differentiate classic and overlooked improvements. The later development of other materials in general may have been expected, but the later development of other means of simultaneously raising and lowering surfaces on an airplane in general would likely have been expected, too.

59. These cases have been able to hide in plain sight because, to employ a Kuhnian frame, they cannot be explained using the dominant conceptual paradigm that identifies things as the primitives of a patentable invention and that is (inaccurately) understood to be a necessary artifact of the contemporary peripheral claiming regime. *See infra* notes 337–40 and accompanying text.

overlooked improvements as "easy" cases and a contemporary theory that suggests the practice should, at least under some circumstances, be highly controversial.[60] To eliminate the explanatory gap between the theory and reality of contemporary patent protection, something has to give. As outlined in the following Section, this Article resolves the puzzle by defending the practice and amending the theory so that it is capable of explaining the practice.[61]

C. THE CORRECTIVE LENS: PROPERTIES, NOT THINGS, AS THE LOCUS OF INVENTION

Given that the conventional theory on patent protection for improvements gains normative traction in the classic-improvement cases but not in the overlooked-improvement cases, the simplest way to correct the blind spot would be to identify an additional factor that distinguishes overlooked improvements from classic improvements. The normative importance of this factor must be so great that the new factor trumps the factors addressed in the conventional theory on improvements that weigh in favor of allowing the improver to escape the earlier-issued patent. This Article adopts this find-a-new-factor approach to correcting the blind spot in the conventional theory. However, in order to identify this factor, it has to undermine and replace one of the most widely-shared conceptual frameworks structuring contemporary understandings of patent law and theory.

Contemporary patent discourse is insistent on the notion that *innovative things* in their entireties are the primitives of the inventions that are protected by patent claims.[62] What is an inventor's patentable invention? Today, the only permissible way to answer this question is to point to the set of things encompassed by the claims to which the inventor is legally entitled. As Jeff Lefstin has noted, "[i]n modern patent parlance, 'the claim,' 'the invention,'

60. As a doctrinal matter, the existence of the "easy" overlooked-improvement cases also undermines the strong fixation theory of literal claim scope. *See infra* Part VIII (presenting and undermining this theory).

61. The opposite tack is also possible, at least in theory. One could argue that infringement cases involving overlooked improvements should not be "easy" cases in which the rights of earlier-filed patent owners routinely expand over time. However, if patents are to structure a market for embodied ideas, this argument would be difficult to defend on a normative level. *See infra* Part VII (defending differential treatment of classic and overlooked improvements as a normative matter).

62. "Primitives" refers to basic units of a system that are "not derived from something else." THE AMERICAN HERITAGE COLLEGE DICTIONARY 1087 (3d ed. 2000). *Cf. infra* note 87 (noting that this Article does address process claims).

and 'the patent' are essentially synonymous."[63] This approach to identifying inventions as sets of things operates on a coarse level of granularity. Precisely what it is about the claimed things that makes them inventive is ignored. Any attempt to identify what an inventor has invented at a level of granularity that is finer-grained than a set of things—that is, any attempt to identify the "spirit" or point of novelty of the patented things that differentiates them from the prior art—is categorically dismissed.[64]

This dismissal of the relevance of the "spirit" or point of novelty of an invention is often justified with dubious reasoning. Sometimes, the dismissal is justified with the inaccurate assumption that a finer-grained approach is incongruous with the modern "peripheral" claiming regime.[65] Sometimes, it may be grounded in the highly questionable twin intuitions that innovative things are stable, real-world entities that make for a good conceptual foundation for patent protection and that the "spirit" of an invention is a nebulous mental construct that is administratively unmanageable.[66]

---

63. Jeffrey A. Lefstin, *The Formal Structure of Patent Law and the Limits of Enablement*, 23 BERKELEY TECH. L.J. 1141, 1145 (2008). Oskar Liivak has similarly noted that "[t]he invention itself has no substantive existence other than as a short-hand for the subject matter that a patentee can claim." Oskar Liivak, Rescuing the Invention from the Cult of the Claim 9 (Feb. 24, 2011) (unpublished manuscript), *available at* http://ssrn.com/abstract=1769270. Liivak also offers a historical explanation for this state of affairs. One purpose of the 1952 Patent Act was to change the way of measuring how much of a technical advance was needed to obtain patent protection. The Act replaced problematic judicial discussions of "the requirement of invention" with the nonobviousness provisions of § 103. A substantive definition of an inventor's invention for the purpose of determining claim scope—substantive in the sense that it does not simply reference the set of things encompassed within a valid claim—is distinct from the concept of the amount of inventiveness required to surmount the nonobviousness threshold. Nonetheless, because both concepts are associated with the word "invention," a substantive definition of what an inventor has invented was the baby that was thrown out with the bathwater of the requirement of invention in post-1952 opinions. *Id.* at 40–42.

64. For an extended analysis of the courts' rejection of the relevance of the "spirit" of an invention, see Bernard Chao, *Breaking Aro's Commandment: Recognizing That Inventions Have Heart*, 20 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 1183 (2010); Mark A. Lemley, *Point of Novelty* 3–9 (Stanford Pub. Law, Working Paper No. 1735045, 2011), *available at* http://ssrn.com/abstract=1735045.

65. For typical language rejecting the use of the "spirit" of an invention in the construction of a peripheral claim, see Ormco Corp. v. Align Tech., Inc., 498 F.3d 1307, 1322–23 (Fed. Cir. 2007) ("This court . . . has rejected a claim construction process based on the 'essence' of an invention."). The assumption that peripheral claims and a focus on the "spirit" of the invention cannot coexist as a logical matter is addressed, and rebutted, below. *See infra* Part VIII.

66. The concept of the innovative properties around which this Article is structured provides an intuitive, real-world grounding for discussions of the "spirit" of an invention. Inversely, it is important not to overstate the case that innovative things are stable, real-world entities. The things at issue in the determination of claim scope and validity are thing-

However, the motivation to ignore the point of novelty of an invention in patent law may also be grounded in an application of the classic rules-standards debate that should not be lightly tossed aside.[67] Identifying the "spirit" of an invention is an information-intensive and error-prone exercise. It takes work to identify the one or more ways in which a patented invention differs from the prior art. Positioning coarse-grained innovative things as the primitives of invention means that less information needs to be considered to decide issues. It is easier to say that a thing is innovative than to say precisely why a thing is innovative, especially if there are several alternative reasons why a thing is innovative. The analysis that follows from focusing on innovative things rather than an innovative "spirit" may, at times, result in greater deviation from the ideal scope of a valid patent claim. However, under some circumstances, the cost of this deviation may be outweighed by benefits of more predictable outcomes and less costly adjudication. Using only innovative things to define patent rights may, in some circumstances, be an efficient proxy for the innovative "spirit" of things, even if the latter more closely reflects a patentee's contribution to technological progress.[68]

Therefore, this Article does not make a blanket claim that it is always important to identify the innovative "spirit" of the things encompassed within a valid patent claim. This Article launches a much more targeted attack. Whatever the merits of this refusal to parse the nature of invention more finely in the context of other patent doctrines, it proves to be highly problematic in the context of patent protection for improvements.[69]

types, not thing-tokens. Types are mental constructs that people carry in their heads, just like the "spirit" of an invention. Kevin Emerson Collins, *The Reach of Literal Claim Scope into After-Arising Technology: On Thing Construction and the Meaning of Meaning*, 41 CONN. L. REV. 493, 514–38 (2008) (illustrating the importance of "thing construction" in patent law). *Cf.* Michael J. Madison, *Law as Design: Objects, Concepts, and Digital Things*, 56 CASE W. RES. L. REV. 381 (2005) (mulling on the importance of the definition of things in intellectual property in general).

67. The primary dimension of the rules-standards debate at issue portrays rules as entrenched generalizations that impose costs in the form of over- and under-inclusiveness in relation to the rule's justification. FREDERICK SCHAUER, PLAYING BY THE RULES: A PHILOSOPHICAL EXAMINATION OF RULE-BASED DECISION-MAKING IN LAW AND IN LIFE (1991). Rules and standards can also be defined by the ex ante and ex post time at which law is made. *See* Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557 (1992).

68. *See* SCHAUER, *supra* note 67, at 145–49 (addressing the efficiency justification of over- and under-inclusive rules).

69. Recent scholarship suggests that interest in a patent doctrine that pays more attention to the "spirit" of an invention (or its heart, core, essence, or point of novelty) may be on the rise. *See* Chao, *supra* note 64, at 1227–38 (arguing that the heart of an invention has relevance in many patent doctrines); Lemley, *supra* note 64 (considering instances in which the point of novelty of an invention does and should have relevance in patent doctrine).

Whenever a coarse-grained, rule-like factual analysis is employed instead of a fine-grained, standard-like one, information is lost. In other words, a more granular picture provides a lower-resolution image and reveals less detail. Similarly, the coarse-grained understanding of invention that positions things as the primitives of what an inventor invents and that structures contemporary patent theory is a low-resolution conceptual framework. Sometimes, the information lost in a coarse-grained analysis is not important, but, when the issue is patent protection for later-developed improvements, it is. It is the root cause of the inability of the conventional theory on improvements to identify and justify the distinction that already exists between the outcomes of contested classic-improvement cases and "easy" overlooked-improvement cases. The factor that differentiates classic and overlooked improvements can only be measured with reference to the "spirit" of the inventions produced by the earlier and later inventors that exist in any improvement scenario. To identify the systematic difference between classic and overlooked improvements, the locus of invention must be identified at a finer-grained level.

To talk about the vague concept of the "spirit" of a set of patented things in a precise and accessible manner, this Article argues that it is useful to identify innovative properties of things, rather than innovative things in their entireties, as the locus of invention. A property of a thing "is what is variously called a feature, quality, attribute or . . . a way that something is."[70]

However, no court or scholar has addressed the role that the "spirit" of an invention has to play in determining patent protection for improvements or the reach of peripheral claims into after-arising technology. Even advocates of the use of a point of novelty approach to patent law in other contexts usually adhere to the strict rule that the determination of the scope of a peripheral claim and a point of novelty approach are incompatible. Chao, *supra* note 64, at 1187 ("The heart of the invention should not be considered when the law needs to determine when something falls within the boundaries outlined by a patent's claims."). In contrast, this Article argues that the point of novelty of an invention must be considered even when using the "all elements" rule and determining the scope of a valid peripheral claim. *See infra* Part VIII.

70. Daly, *supra* note 9. To the same end, consider the following common-sense definition of a "property":

> Asked to describe a given tomato, you might cite its redness, its size and its age. In doing so, some philosophers would claim, you have cited some of the tomato's properties. A property is what is variously called a feature, quality, attribute or (as some philosophers put it) a way that something is. A property is supposed to be an entity that things (including particulars, such as tomatoes or people) have.

Id.; *see also* Chris Swoyer, *Properties*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY § 1.1 (rev. ed. 2000), *available at* http://plato.stanford.edu/entries/properties/ (offering another common-sense definition of a property). Only a rough, working definition of a property is needed here, and only such a definition is possible. A definitive philosophical definition of a

Properties are familiar metaphysical entities. *Being red* is a property, as is *being six inches long* and *being on the edge of a desk*.[71] Properties are finer-grained entities than things are. *Being red* is only one of many properties possessed by a McIntosh apple. Things, in turn, can be defined on an intuitive level by the sum total of the properties that they possess.[72] Properties of things offer a new way of describing what an inventor has invented that is more specific than talk about sets of innovative things.[73] Grossly described, inventors who produce patentable inventions create new things. For example, the inventor of a slow-release pill has created a pill, or set of pills if the slow-release matrix works with many different drugs, that did not previously exist. At a more detailed level, however, what inventors do when they invent things is to create innovative properties of things. They reconfigure material so that it possesses properties that the things in the prior art did not possess, and any property not possessed by the prior art is an innovative property (at least in the sense that it makes the things that possess it novel things). For example, the inventor of the first slow-release version of a pill endows things with innovative properties like *having a particular matrix structure in which the drug is embedded* or *being able to release an active ingredient slowly over time in the human digestive tract*.[74] Identifying the innovative properties as the locus of invention is a more fine-grained way of talking about invention. Stating that an inventor invented a set of things does not convey as much information as stating that an inventor invented the particular properties that make the thing innovative. The reason that talk of properties conveys a finer-grained understanding of invention is that patented things do not have a fully new slate of properties. They are agglomerations of properties, some of which

property would entail a complex ontological discussion that is beyond the scope of this Article. Philosophers who study properties even disagree about whether properties exist. *See generally* Daly, *supra* note 9, § 1.

71. For clarity, this Article uses the stylistic convention of putting all properties recited in the text in italics.

72. The notion that things can be defined by a set of properties is a corollary of "Leibniz's Law," which holds that no two discernable objects have exactly the same set of properties. Peter Simons, *Identity of Indiscernibles*, *in* ROUTLEDGE ENCYCLOPEDIA OF PHILOSOPHY, *supra* note 9 (discussing the identity of indiscernibles).

73. Properties also provide a convenient (but today absent) link between the distinct realms of material things and ideas per se, both of which are important in patent theory. Properties can be taken to be the entities that instantiate or embody innovative ideas in things. *See infra* note 220 and accompanying text.

74. Innovative properties can be either structural or functional. This distinction is important in determining an inventor's rights in classic-improvement cases as functional properties often operate at a higher level of generality than structural properties do. *Cf. infra* notes 242–48 and accompanying text (introducing the concept of the "least-general naked property" of a classic improvement).

existed in the prior art and some of which did not (and are thus the locus of the inventor's invention). Patentable things must have at least one novel and nonobvious property to be patentable, but they also possess many properties that prior-art technologies already possessed at an earlier point in time.[75] For example, the newly invented slow-release pills discussed above are newly invented things that have properties like *being round*, *having active chemical entity X*, and *weighing 0.1 ounces*, all of which are properties of prior art pharmaceuticals as well.

Focusing on properties of things, rather than things in their entireties, as the primitives of what an inventor has invented enables a finer-grained analysis of invention in improvement cases. If the conventional framework is adopted and innovative things are taken to be the primitives of what an inventor invents, then all that can be said about the intrinsic qualities of an improvement in relation to the things disclosed and claimed by the earlier patentee is that the improvement is a new thing that was not made available to the public by the earlier inventor's patent disclosure. In this information-poor environment, one cannot even coherently pose a question about the extent to which the improvements still embody or manifest in some way the invention that justified the earlier-issued patent. To do so would require a concept of an invention that facilitates a comparison between two sets of things: the set of things made available by the earlier patentee and the set of things produced by the later improver. A concept of an invention that is itself a set of things, without more, cannot perform this job.

Thus, in this information-poor environment, the only way to draw distinctions between different improvement cases is to look outward to the context in which the earlier- and later-invented sets of things exist. This outward-looking approach is precisely the one that the conventional theory on patent protection for improvements takes. It looks to consumer demand for the earlier-invented things and the later-improved things.[76] It looks to the patterns of technical advance and licensing in an industry.[77] It looks to the magnitude of the sunk costs of innovation in an industry.[78] It toggles

75. The argument is not that any inventor's achievement can be correctly summarized as a single advance or that patented things possess a sole innovative property. Innovative things may possess multiple, distinct innovative properties. They may also embody multiple, nested innovative properties that describe the technical progress generated by an inventor at varying levels of generality. The multiple levels of generality are critical to understanding classic-improvement cases in terms of successively invented properties. *See infra* Section V.A (defining a classic improvement in terms of successively invented properties).

76. *See supra* notes 18–21 and accompanying text.

77. *See supra* notes 22–23 and accompanying text.

78. *See supra* notes 24–26 and accompanying text.

between different mechanisms through which patent rights can be welfare-enhancing tools.[79]

By zooming in more precisely on innovative properties of things as the inventions of the successive inventors, the information-rich nature of the environment becomes apparent. Even as coarser-grained things change over time, finer-grained properties can remain constant, allowing an earlier inventor's invention to be tracked through later-developed improvements. The intrinsic qualities of the earlier-invented and later-improved things can be compared, and the factor that distinguishes overlooked from classic improvements comes into focus. The difference between the two types of improvements stems from two different ways in which the properties invented by successive inventors relate to each other in an improved thing.

In a classic improvement, the properties invented by the improver *refine* the properties invented by the earlier inventor.[80] The improvement possesses the most general of the earlier inventors' innovative properties, but the improver's innovative properties displace the most specific of the earlier inventor's innovative properties. In other words, the improver's invention *compounds with* the earlier inventor's invention framed in its most general fashion, but it *supplants* the earlier inventor's invention framed in its most specific fashion. For example, Curtiss's improved airplane continued to embody the Wright Brothers' innovative idea about raising and lowering wing-surfaces simultaneously.[81] Curtiss's improvement still possessed the property *having a plurality of surfaces that can be raised and lowered simultaneously*. However, Curtiss's innovative property—*having a flap (aileron) that can move independently of the wing surface*—displaced the Wright Brothers' more specific idea about wing-warping, as planes with ailerons do not possess the property *having a wing surface that is capable of being warped.*

In contrast, overlooked improvements do not involve the refinement of properties. The overlooked improvements involve later-invented properties that only compound with, and do not supplant at all, the earlier-invented properties.[82] For example, the process of improving the airplane to generate the improved-canvas version does not displace the properties that embody the Wright Brothers' technological advance in an airplane. The improved

79. *See supra* notes 27–28 and accompanying text.

80. This relationship of refinement is explored in more detail *infra* Part IV.

81. The following analysis to distinguish classic and overlooked improvements builds on the discussion of the inventions of the Wright Brothers and Curtiss. *See supra* notes 29–32, 40 and accompanying text. The identical analysis can be built on a software example. *See supra* notes 51–52 and accompanying text; *infra* Section V.B.2.

82. This relationship of pure compounding is explored in more detail *infra* Part V.

canvas airplane still possesses the full range of the Wright Brothers' innovative properties at all levels of generality, from *having a plurality of surfaces that can be raised and lowered simultaneously* to *having a wing surface that is capable of being warped*.[83] The improver's innovative properties—including *being made of a later-developed canvas*—only compound with the Wright Brothers' innovative properties. The Wright Brothers' technological advance is wound up in the properties that enable a wing to warp; the Wright Brothers neither invented any canvas nor pioneered advances in the use of the existing canvas in airplane construction. The improved canvas is a distinct advance in a different technological area that is built on an independent foundation of knowledge, and it is embodied in properties of the airplane that are functionally independent of the properties that embody the advance of wing-warping. Thus, despite the fact that the improved-canvas airplane is an improvement that was not disclosed to the public by the Wright Brothers' patent, it embodies the Wright Brothers' invention just as much as the wing-warping airplanes that were actually disclosed by the Wright Brothers' patent.

Importantly, a shift to a focus on properties as the locus of invention is not proposed because precision is an end in and of itself.[84] Adopting a fine-grained conceptual framework in which inventors invent properties of things brings to light a factual difference that distinguishes classic and overlooked improvements. This factual difference does not register in a coarse-grained conceptual framework that takes things in their entireties to be the primitives of what an inventor invents. A focus on properties as the locus of invention is desirable because it is a tool that can do what the contemporary focus on things as the primitives of invention cannot. By allowing courts and scholars to differentiate classic and overlooked improvements, it can reduce the gap between theory and practice and explain an uncontroversial feature of how the contemporary patent regime already works.[85]

III. DEFINING AN IMPROVEMENT

The legal and economic literature on patent protection for improvements defines an improvement in many different ways. To head off confusion,

83. *But cf. infra* note 158.

84. In fact, precision can have costs if the information needed to be more precise is costly to obtain and consider. *See supra* notes 67–68 and accompanying text (framing the choice between things and properties as the locus of invention in terms of the rules-standards debate).

85. A focus on properties as the locus of invention also reveals an additional factor that is relevant when determining the reach of earlier patentees' rights into later-developed classic improvements. *See infra* notes 242–48 and accompanying text (introducing the concept of a classic improvement's least-general naked property).

Section III.A provides the three criteria that define an improvement as the term is used in this Article. To clarify this definition, Section III.B emphasizes that improvement is only one species of the broader genus of means through which cumulative innovation occurs. Section III.C briefly addresses the doctrinal mechanics of how earlier inventors' patent claims can extend to later-developed improvements. The purpose of this short digression into patent doctrine is simply to demonstrate that patent doctrine, at least on its rhetorical surface, does not provide any basis for distinguishing classic and overlooked improvements.

A. THE THREE CRITERIA THAT DEFINE AN IMPROVEMENT

This Article uses the term "improvement" as a term of art that requires a definition.[86] This Section provides this definition by identifying the three necessary criteria of an improvement.

1. The Thing Criterion

The first criterion is that improvements are things—the entities that are described and privatized by patent claims.[87] This criterion is important because the term "improvement" could be used to refer to an idea or increment of newly discovered technological knowledge that makes it possible for humans to conceive and/or make new things. Here, however, it is not. For clarity, this Article refers to these increments of technological knowledge as innovative ideas or advances.

2. The Timing (or New-Thing) Criterion[88]

Improvements are not disclosed or made available to the public by any earlier-issued patent (or other prior art).[89] Improvement stories always feature two inventors, one acting after the other.[90] The first invents a technology and

86. A key feature of the definition is that it avoids reference to the legal outcome of a patent infringement suit. For example, an improvement cannot be defined as a later-developed technology that falls within the scope of an earlier-filed claim. Such a definition would lead to circular reasoning, as the goal of the exercise is to identify which improvements should be within the control of an earlier inventor.

87. The phrase "improvement-as-thing" is therefore redundant, but this Article periodically uses it nonetheless for emphasis. Things include machines, manufactures, and compositions of matter. *See* 35 U.S.C. § 101 (2006). Processes are also patent-eligible subject matter, *id.*, but this Article brackets discussion of process claims and defers it to a later date.

88. Both the second and third criteria illustrate that an improvement can only be defined in relation to a particular patent.

89. *See supra* note 3.

90. It is possible for a single person acting at different times to fill both roles, but the reach of the earlier patent into improvements is not as important in this situation. *See supra* note 16.

patents it, and the second improves on it.[91] The second inventor uses the ideas disclosed in the earlier patent as inputs in an inventive process, and he produces things that are intrinsically new[92] and that are an inventive stride beyond the things disclosed by the earlier inventor. For simplicity, improvements always involve technological advances over the prior art, which at the time of the improvement includes the earlier patent disclosure.[93] The later-developed improvement can therefore never be disclosed by a patent on the earlier-developed, improved-upon technology. An improver makes new things after the filing of the earlier patent. He does not make, use,

91. Despite what its everyday meaning might suggest, an improvement in patent law need not be technically superior to the prior art. It need not be better at satisfying consumer preferences or reducing production costs. It merely needs to be innovatively different from the prior art. Giles S. Rich, *The Principles of Patentability*, 28 GEO. WASH. L. REV. 393 (1960) (explaining that an invention need not be better than the prior art to be patentable).

92. The phrase "intrinsically new" denotes that the inventor must produce a new thing and not simply discover a new purpose for an old thing or put an old thing into a new context that contains after-arising things. *See infra* text accompanying notes 110–16 (noting that disclosed-thing cumulative innovation cases are not improvement cases). For a detailed discussion of an intrinsic property and how it differs from an extrinsic property, see Collins, *supra* note 66, at 520–36.

93. A patent specification directly discloses, and thus makes available to the public, only the small set of things that it describes in full. (Even this is an idealization, however, as things are never described "in full.") However, it also constructively discloses the broader set of things that the person having ordinary skill in the art would have readily thought of, and understood how to make, after reading the disclosure. *See infra* notes 130–33 and accompanying text (discussing patent law's disclosure doctrines). Therefore, many later actors who make minor changes to the precise things disclosed in full in a patent specification do not generate improvements as the term is used in this Article because they produce things that were constructively disclosed. *But cf.* Lemley, *supra* note 2, at 1007–08 (labeling such later actors as "minor improvers"). For convenience, it is simplest to assume that the threshold of what divides a thing that is not constructively disclosed by a specification from a thing that is, and the threshold of what constitutes a patentable improvement that satisfies the nonobviousness requirement, are identical. That is, it is simplest to assume that all improvements involve patentable advances over the prior art, that all improvers can obtain patent protection for their improvements, and that overlapping blocking patents will result if the earlier inventor's claim is allowed to encompass the improvement. *Cf. infra* text accompanying note 297 (distinguishing overlapping and economic blocking patents). However, whether the improver actually seeks patent protection at the PTO is irrelevant to the issue at hand, namely whether the earlier patentee's rights should encompass the improvement. Furthermore, it is possible to imagine that the threshold of an advance that generates a thing that is not constructively disclosed by an earlier patent and the threshold of a patentable advance should not perfectly align. However, this additional wrinkle would not affect the distinction between classic and overlooked improvements explored in this Article. A perfect alignment of the thresholds is therefore assumed to simplify the analysis.

or sell things that were disclosed and made available to the public by the earlier patent upon which he improves.[94]

3. The Embodies-the-Earlier-Advance Criterion

A later-developed thing is an improvement in relation to an earlier-patented technology only if it continues to embody in some way the technological advance that justified the issuance of the earlier patent.[95] This criterion may at first be counterintuitive to readers steeped in patent discourse because the term "embodiment" already has a common meaning in patent rhetoric. It is used as a noun to mean a thing that is disclosed in a patent specification or, more broadly, at least a thing that falls within the scope of a patent.[96] This conventional meaning of "embodiment" is unhelpful here because it reinforces the dominant paradigm in which things are the primitives of what an inventor invents.[97] To the extent that the term "embodiment" should still be used at all as a noun in this Article, it means a property, or a set of properties, of a thing, not a thing in and of itself.[98]

The embodies-the-earlier-advance criterion is needed because it differentiates an improvement from a generic after-arising technology.[99] Without this criterion, every after-arising technology would be an improvement. For example, assume that a first, earlier inventor invents a new drug that cures the common cold, and a second, later inventor invents a new coffee sleeve. The new coffee sleeve is a thing (satisfying the thing criterion),[100] and it is a new thing, not disclosed in the drug patent, that embodies an advance over the state of the art at the time the drug patent was

94. The phrase "later-developed improvement" is therefore redundant, but this Article periodically uses it nonetheless to differentiate the issue of the patent protection that the earlier inventors can get for subsequent improvements from the issue of the patent protection that the later-acting improvers can get for their own improvements.

95. In turn, what it means for a thing to embody an earlier inventor's advance or idea can be taken to be nothing more than what it means for a thing to possess at least one innovative property that is attributable to the earlier inventor. *See supra* notes 70–75 and accompanying text (introducing the concept of a property); *infra* note 220 and accompanying text (positioning a property as the entity that instantiates an idea in a thing).

96. *See, e.g.*, AK Steel Corp. v. Sollac, 344 F.3d 1234, 1244 (Fed. Cir. 2003) (using a "disclosed embodiment" to describe a technology revealed in the specification); Waymark Corp. v. Porta Sys. Corp., 245 F.3d 1364, 1368 (Fed. Cir. 2001) (using "infringing embodiment" to describe a technology that falls within a patent claim).

97. *See supra* notes 62–67 and accompanying text (presenting this dominant paradigm).

98. *Cf. supra* note 96 (noting uses of "embodiment" that refer to things).

99. Although this distinction is conceptually important to understand the mechanics of patent protection, little of economic import turns on the distinction between improvements and after-arising technology that is not an improvement. *See infra* note 123.

100. *See supra* Section III.A.1.

filed (satisfying the timing, or new-thing, criterion).[101] Yet, the concept of an improvement loses its utility if its meaning is so broad that the coffee sleeve is an improvement on the drug patent. Later-developed things that are entirely unrelated to the ideas that justified the issuance of an earlier patent should generically be after-arising technologies, not improvements. The embodies-the-earlier-advance criterion provides an intuitive upper limit on the notion of an improvement that allows improvements to occur only when the later-developed thing continues to embody in some way the technological advance of the earlier patentee.

The economic literature on improvements offers two definitions of an improvement that could be used to provide this limit without reference to the notion of a thing embodying an earlier inventor's advance. However, neither is satisfactory for the purpose of this Article because neither corresponds to how the patent regime actually works.

First, an improvement could be identified with a process-oriented definition under which an improvement exists whenever a later innovation is facilitated by an earlier one.[102] Following this definition, a later-developed thing could only be an improvement on an earlier-developed thing if the later inventor actually knew of and built upon the earlier inventor's work in some way, whether consciously or unconsciously. In other words, the status of a later-developed thing as an improvement would be contingent on a later actor getting a leg up from a prior actor. This definition is fatally over- and under-inclusive with respect to the actual patent regime. There are many cases in which products made by earlier-generation inventors facilitate the inventions of later-generation inventors that are not improvement cases.[103] For example, the earlier invention of a particular type of blast furnace may facilitate the later invention of a new type of metal alloy, but the metal alloy is not an improvement on the furnace. Inversely, there are many improvement cases in which the later inventor is ignorant of the earlier inventor's work, meaning that there is no actual facilitation. Copying is not an element of a cause of action in patent infringement,[104] and thus independent improvement

101. *See supra* Section III.A.2.

102. Scotchmer, *supra* note 2, at 31 (discussing a variety of ways in which earlier inventions can facilitate the development of later inventions).

103. *See, e.g.*, *infra* notes 110–16 and accompanying text (discussing disclosed-thing cumulative innovation cases); *infra* notes 117–18 and accompanying text.

104. *See* DePuy Spine, Inc. v. Medtronic Sofamor Danek, Inc., 567 F.3d 1314, 1328–30 (Fed. Cir. 2009). In contrast, copyright requires copying as an element of infringement. Arnstein v. Porter, 154 F.2d 464, 468–69 (2d Cir. 1946).

is not a defense to patent infringement.[105] Later generations may produce things that are improvements on the things generated by earlier inventors and that infringe earlier patents even if the later generations are entirely unaware of the earlier inventors' efforts.[106]

Second, the economic literature identifies improvements as economic substitutes. An improvement could exist whenever a later-developed thing is an economic substitute for an earlier-developed thing.[107] While it is true that most—but not all—improvements are things that are economic substitutes for earlier-patented things,[108] the inverse statement does not hold true. It is difficult to conceive of a later-developed mouse poison as an improvement on an earlier-developed mousetrap as the term "improvement" is commonly used in patent law.

B. IMPROVEMENT AS A SPECIES OF CUMULATIVE INNOVATION

Technological progress is a cumulative endeavor. The outputs of the work of earlier generations of inventors are inputs into the work of later generations of inventors.[109] Improvements clearly implicate cumulative innovation. Earlier inventors produce innovative things, and later inventors figure out a way to make better things that still embody some of the innovative ideas of earlier inventors. However, it is important not to equate cumulative innovation with improvements or assume that a solution to the problem of patent protection for improvements in patent law is a silver bullet for the problem of distributing rewards among multiple generations of inventors. Improvements are only a species of the broader genus of mechanisms through which cumulative innovation occurs, and cumulative innovation presents many challenges to the patent regime that do not involve

105. One can reasonably argue that independent invention should be a defense to patent infringement. *See, e.g.*, Oskar Liivak, *Rethinking the Concept of Exclusion in Patent Law*, 98 GEO. L.J. 1643 (2010); Samson Vermont, *Independent Invention as a Defense to Patent Infringement*, 105 MICH. L. REV. 275 (2006) (considering the implications of an independent-inventor defense). A component of this argument is that the reach of an earlier patentee's rights into later-developed improvements should be contingent on process, too, and that later-developed improvements should infringe earlier-issued patents only when the later actor has actually learned something from the earlier inventor.

106. Recognizing that independent invention can yield improvements, it is still possible to identify constructive facilitation in all improvement cases. If the later improver had known of and read the earlier patentee's disclosure, the disclosure would have facilitated the improver.

107. *See infra* Section VII.C.1 (defining complements and substitutes).

108. *See infra* notes 276–78 and accompanying text.

109. Rebecca S. Eisenberg, *Patents and the Progress of Science: Exclusive Rights and Experimental Use*, 56 U. CHI. L. REV. 1017, 1055 n.161 (1989) (listing sources from the sociology of science that discuss the cumulative nature of technological progress).

improvements. To clarify what improvements are by illustrating what they are not, this Section briefly addresses two types of cases that implicate cumulative innovation but do not involve improvement.

First, there are cumulative innovation cases in which later innovators use the exact things disclosed and made available to the public by earlier inventors. These cases are *disclosed-thing* cumulative innovation cases. They are not improvement cases because the things that the later innovators make or use do not satisfy the timing (or new-thing) criterion of an improvement.[110] In some disclosed-thing cases, the earlier-disclosed things are components of products that also contain after-arising technologies. For example, earlier-invented smaller chips (A) allow later inventors to design innovative, light-weight devices (B) that are sold to consumers with the chips integrated therein (A+B).[111] After-arising component cases are not improvement cases: the earlier-filed claim to A can encompass the combination of A+B without any expansion in the set of distinct things that it describes.[112] To be an improver in the technical sense that raises the issue of the reach of patent protection into later-developed things, the later innovator must produce an A′ that does not contain an A as a component. In other disclosed-thing cumulative cases, later inventors discover new uses for earlier-invented technologies. The earlier patentees of things may control those things even when the things are being used in later-discovered ways.[113] Again, the earlier patents can encompass the later new-use innovations without having to encompass any new thing, so new-use innovation cases are not improvement

110. *See supra* Section III.A.2.

111. Some after-arising component cases involve infringing technologies that can be intuitively called improvements, even though they are not improvement cases in the technical sense. For example, assume that an earlier inventor patents chemical A, a cleaner. A later inventor may invent chemical B, an additive that does not change A's chemical structure but that increases A's cleaning efficacy. In an everyday sense, the later inventor has created an improvement, as A+B cleans better than A does. In the more technical, patent sense, however, there is no improvement. The combination A+B infringes the earlier claim to A not because A+B is an improvement but because the A in A+B is the same old A that the earlier inventor disclosed in her patent specification. The later inventor has used A as a component in combination with a distinct, complementary, and after-arising thing.

112. The scope of the claim to A does not need to expand to encompass A+B because the new property of A that has been invented by the later innovator—its property of *being in a combination with B*—is an extrinsic property of A and therefore does not define the thing-type A as distinct from other thing-types. *See* Collins, *supra* note 66, at 520–36 (distinguishing between intrinsic and extrinsic properties).

113. *See* A.B. Dick Co. v. Burroughs Corp., 713 F.2d 700, 703 (Fed. Cir. 1983) ("[A] pencil structurally infringing a patent claim would not become noninfringing when incorporated into a complex machine that limits or controls what the pencil can write.").

cases.[114] In yet other disclosed-thing cases, the later inventor may use a tangible instance of an earlier-disclosed thing as an input into technological progress but not as a part of whatever product (if any) is eventually invented. For example, the disclosed thing may be a research tool.[115] Despite the fact that they are wound up in the process of cumulative innovation, the later inventors in these disclosed-thing cases have generated improvements.[116]

Second, there are *idea-only* cumulative innovation cases. In idea-only cases, an idea qua idea—that is, a thought about or a representation of knowledge itself—is both the input into and the output of the later actors' efforts. Later actors use the knowledge generated by earlier inventors and disclosed in patent specifications as an input into further mental progress in technological ideas per se, and they generate new mental knowledge without generating any new things at all. For example, an earlier inventor may patent a molecule that is useful for treating a disease. The patent specification discloses the structure of the molecule. After the patent has been filed, someone else may learn the knowledge of the protein structure disclosed patent specification, appropriate an idea qua idea from the earlier patent, and have a "Eureka!" moment. He may realize that if a molecule with the

114. Again, the scope of the claim to A does not need to expand to encompass A when used in a later-developed manner because the new property of A that has been invented by the later innovator—its property of *being put to a new use*—is an extrinsic property of A and therefore does not define the thing-type A as distinct from other thing-types. *See* Collins, *supra* note 66, at 520–36 (distinguishing between intrinsic and extrinsic properties).

115. *See* SCOTCHMER, *supra* note 24, at 132–33.

116. Because they involve cumulative innovation, disclosed-thing cases raise the specter of first-generation rights impeding second-generation progress. *Cf. infra* Section VII.C.2.a (framing the successive inventions in cumulative innovation as complements). However, whatever tempering of the first-generation rights is required to address the problem cannot be accomplished by tailoring the size of the set of distinct things that falls within claim scope. Armed only with the rights to the set of things that is needed to prevent "pirates" from copying their inventions, patentees have sufficient patent scope to reach the conduct of the later inventors in disclosed-thing cases. Rather, the tempering must be achieved with other policy levers. User-specific defenses to patent infringement, such as the experimental use defense, are the topic of much commentary in research tool cases. *See, e.g.*, Eisenberg, *supra* note 109; Katherine J. Strandburg, *What Does the Public Get? Experimental Use and the Patent Bargain*, 2004 WISC. L. REV. 81. Less powerful remedies, such as the denial of injunctive relief or the lowering of the reasonable royalty, are commonly discussed as ways of dealing with after-arising component cases. *See* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388, 396–97 (2006) (Kennedy, J., concurring) (stating that injunctive relief may not be appropriate when a patented invention is a small component of a larger product); Mark A. Lemley & Carl Shapiro, *Patent Holdup and Royalty Stacking*, 85 TEX. L. REV. 1991 (2007) (addressing the royalty stacking problem that results from many reasonable-royalty damages in multi-component products). The reverse doctrine of equivalents, too, can be used to excuse a later innovator in a disclosed-thing cumulative innovation case from infringement. *See infra* note 140 and accompanying text.

molecular structure revealed in the patent has a particular biological activity, then perhaps the metabolic pathway in a cell must include a particular step in order for the molecule to have that activity. The later actor may go on to discover a previously unknown metabolic pathway. Idea-only cases like this hypothetical clearly involve cumulative innovation: the later actor was spurred along or sped up by the work of the earlier actor. Yet, in sharp contrast to the disclosed-thing cases, the later actor is categorically beyond the reach of the patent rights of the earlier inventor. The reason why later actors are allowed to forgo compensating the earlier actors is that patent protection does not propertize ideas per se, but instead, only grants rights to exclude from sets of things (and processes) that embody innovative ideas.[117] Ideas and advances are bits of knowledge, and newly discovered knowledge qua knowledge must be placed into the public domain as part of the quid pro quo of patent protection.[118] The later actor is free to use the knowledge discovered and disclosed by the earlier patentee without running afoul of patent rights, and the subsequent generation of new knowledge does not infringe, either.

The disclosed-thing and idea-only cumulative innovation cases anchor the two ends of a spectrum. In between them lie cases in which a later inventor uses the ideas, but not the things, disclosed by the earlier patentee as

117. Supreme Court case law addressing patents is full of off-hand references to the fact that ideas per se cannot be propertized with a patent claim. *See, e.g.*, Rubber-Tip Pencil Company v. Howard, 87 U.S. (20 Wall.) 498, 507 (1874) ("An idea of itself is not patentable, but a new device by which it may be made practically useful is."). However, many of these cases do not directly address the fact that ideas, advances, and knowledge themselves, whether in the form of human thought or worldly representations like written texts, cannot be patented. Rather, many of these cases address the issue of patent scope: sets of things that are defined by ideas drawn at excessively high levels of generality cannot be patented, either. *See* Kevin Emerson Collins, Bilski *and the Ambiguity of "An Unpatentable Abstract Idea,"* 15 LEWIS & CLARK L. REV. 37 (2011) (distinguishing two distinct concepts of what it means to patent an idea). The simplest way to demonstrate that an idea per se—in the sense of knowledge itself—is not patentable is to recognize that even "idea free-riders"—who engage in no subsequent invention at all—are off the hook so long as the inventive idea disclosed in a patent is used only qua idea. Billing themselves as experts, idea free-riders can legally profit from conveying the knowledge disclosed in a patent to interested parties. *See* Telectronics Pacing Sys., Inc. v. Ventritex, Inc., 982 F.2d 1520, 1523 (Fed. Cir. 1992) (holding that the dissemination of data about a device falling within a patent's claims is not an infringing activity).

118. Kevin Emerson Collins, *Semiotics 101: Taking the Printed Matter Doctrine Seriously*, 85 IND. L.J. 1379, 1427–30 (2010) (discussing the duality of privatizing claims and publicizing disclosures in patent law); Kevin Emerson Collins, *Claims to Information Qua Information and a Structural Theory of Section 101*, 4 I/S: J.L. & POL'Y FOR INFO. SOC'Y 11 (2008), *reprinted in* PATENT CLAIMS: JUDICIAL INTERPRETATION AND ANALYSIS (ICFAI Univ. Press 2009) (same).

inputs into technological progress and produces new, innovative things. The later innovator's use of the ideas disclosed by the earlier patentee does not infringe the earlier innovator's rights,[119] but the later innovator's production of new, innovative things may infringe the earlier patentee's rights. These intermediate cases can, in turn be grouped into three categories by drawing two lines. The first line marks the distinction between improvements and non-improvement after-arising technologies that are facilitated by the patentee's disclosure. As discussed above, improvements result only when the later-developed thing still embodies in some way the earlier patentee's innovative idea.[120] Inversely, in a non-improvement after-arising technology, the earlier patentee's disclosure may have facilitated the later innovator's work, but the things that the later innovator produces are unrelated to the things protected by the earlier patent. Because patents describe and propertize innovative sets of things, not innovative ideas per se, the earlier patentee never obtains rights that are sufficiently broad to encompass the later innovator's newly invented, unrelated thing.[121] For example, if a later innovator reads an earlier patent on a drug that discloses the metabolic pathway into which a drug intervenes, he may be inspired to develop an entirely different drug that has its effect by intervening in the same pathway but in a different manner.[122] Here, the later innovator has gotten a leg up from an earlier innovator, but the earlier patent does not encompass the later-developed technology. The second line distinguishes two sets of improvements: those that infringe the earlier patentee's rights and those that do not. The question here is whether the later-developed thing embodies the innovative ideas of the earlier patentee in a manner that is sufficiently strong or important to merit including that thing within the earlier patentee's rights.[123] The conventional theory on improvement addresses the optimal

119. *See supra* notes 117–18 and accompanying text (discussing idea-only cumulative innovation cases).

120. *See supra* Section III.A.3.

121. *Cf.* R. Polk Wagner, *Information Wants To Be Free: Intellectual Property and the Mythologies of Control*, 103 COLUM. L. REV. 995, 1000 (2003) (arguing that "even perfectly controlled works" in the sense of inventions that are governed by the maximum allowed amount of intellectual property rights "nonetheless transfer significant information into the public domain" because of the creative connections that later innovators may make).

122. Here, the later innovator is engaging in rational drug design. Michael A. Carrier, *Two Puzzles Resolved: Of the Schumpeter-Arrow Stalemate and Pharmaceutical Innovation Markets*, 93 IOWA L. REV. 393, 402 (2008) (describing "rational drug design" as the process of "working backwards from knowledge of a disease's biochemistry").

123. The line between improvements to which the normative claim is the weakest and non-improvement after-arising technology is a fuzzy one. For example, someone who reads a patent and understands a new chemical's structure may be inspired by that structure to create a new mechanical device, such as a stapler, that employs a similar spatial

position of this line,[124] as does the distinction between classic and overlooked improvements around which this Article is structured.

C. THE MECHANICS OF PATENT PROTECTION FOR IMPROVEMENTS

Contemporary patents contain two distinct types of texts, each of which serves a distinct function. By volume, the bulk of a patent is usually the specification—a text that teaches the public about an invention.[125] The specification often describes why an invention is a technological advance over the prior art, and it provides detailed explanations of particular working examples of an invention. At the end of the specification, patents also contain claims. Claims are short (at least in a relative sense) descriptions of the sets of technological things that constitute the patent owner's legal interest.[126] Contemporary claims are "peripheral" claims because they list a set of properties that a thing must possess to be included in the claim and thereby establish the outer bounds (or periphery) of a patentee's interest ex ante.[127] For example, the first inventor of the coffee sleeve might claim "an insulating band tapered in the shape of a truncated cone to fit the conical outer surface of a disposable coffee cup."[128] The claims usually describe a set of things that is broader than the examples described in detail in the specification. For example, the specification may explain in detail how to make a coffee sleeve that is two inches tall, but valid coffee-sleeve claims may encompass sleeves of many different heights.

While claims can encompass a set of things that reaches beyond the precise things disclosed in full in the specification, patent applicants are not

configuration. It is tempting to think of the stapler as unrelated to the molecule, but the stapler does possess a property *having a certain structural configuration* that was earlier possessed by the chemical. At the end of the day, the fuzzy nature of this line is not problematic as the line carries no legal consequences. Neither improvements in which the normative claim of the earlier patentee is the weakest nor non-improvement after-arising technologies are likely to infringe earlier-filed claims.

124. *See supra* Section II.A.

125. SRI Int'l v. Matsushita Elec. Corp. of Am., 775 F.2d 1107, 1121 n.14 (Fed. Cir. 1985) (en banc) ("Specifications teach.").

126. *Id.* ("Claims claim."). Technically, claims are part of the specification. *In re* Gardner, 480 F.2d 879, 879 (C.C.P.A. 1973). This dual status makes sense because claim language can both define a patentee's legal interest and teach the public about the invention at the same time.

127. Lefstin, *supra* note 63, at 1145 ("[Peripheral claims] recite a set of characteristics, or properties, that define the subject matter encompassed by the patent."). For a longer discussion of the nature of peripheral claims and their compatibility with a focus on properties as the locus of invention, see *infra* Part VIII.

128. For the sake of readability, this and all other hypothetical claims employed in this Article ignore the complex and stilted conventions of claim drafting.

free to claim whatever set of things they please. The Patent Act codifies a number of validity doctrines that constrain the claimable set. Some validity doctrines, including novelty and nonobviousness, work retrospectively to ensure that the claimed set of things is an invention in the colloquial sense, i.e., that it embodies a sufficiently important advance over the prior art to merit patent protection.[129] Other doctrines operate prospectively. They constrain the reach of patent claims into technologies that do represent an advance over the prior art. Among other doctrines, the disclosure doctrines of enablement and written description perform this task.[130] The disclosure doctrines are appropriately named: they limit the scope of a claim to a set of things that is commensurate with the contribution to technological progress that an inventor discloses in the patent specification. Enablement requires that an inventor teach the person having ordinary skill in the art ("PHOSITA") to make and use a set of things without undue experimentation that is commensurate with the claimed set of things at the time the patent is filed.[131] Written description requires an inventor to demonstrate to the PHOSITA that the claimed set of things is commensurate with the claimed technology that was "invented" or "possessed" at the time of filing.[132]

The inquiries specified in the disclosure doctrines can be used to identify a core set of technologies that can be claimed. For simplicity, the set of things that the PHOSITA at the time of filing could actually (1) make and use without undue experimentation and (2) recognize as possessed, is the set of things that a patent *discloses* or *makes available to the public*.[133] This set of

129. 35 U.S.C. §§ 102, 103 (2006) (codifying the novelty and nonobviousness doctrines). The "prior art" has a technical definition that is roughly captured as the publicly accessible technological status quo at the time of an invention. *See* § 102.

130. *Id.* § 112, ¶ 1. Other doctrines that perform this function—or, at least, could, if courts were inclined to use them as policy levers—include: claim construction, Phillips v. AWH Corp., 415 F.3d 1303, 1321 (Fed. Cir. 2005) (en banc); a prohibition on the use of purely functional claim limitations at the point of novelty construed according to their ordinary meanings, Halliburton Oil Well Cementing Co. v. Walker, 329 U.S. 1, 9 (1946); § 112, ¶ 6; and a prohibition on claims to abstract ideas under the patentable subject matter provision of § 101, see O'Reilly v. Morse, 56 U.S. (15 How.) 62, 113 (1853).

131. Genentech, Inc. v. Novo Nordisk A/S, 108 F.3d 1361, 1365 (Fed. Cir. 1997).

132. *See* Ariad Pharms., Inc. v. Eli Lilly & Co., 598 F.3d 1336 (Fed. Cir. 2010) (en banc).

133. The important question is whether the PHOSITA could make and use a particular thing, and recognize it as being possessed, at the time of filing. It is irrelevant whether a claim can remain valid under the enablement and written description doctrines while encompassing the thing. The validity of a claim depends on the commensurability of the disclosure and the claims, and the enablement and written description doctrines commonly sanction the validity of claims that encompass things beyond the set of things that is actually enabled and possessed. *See, e.g.*, Collins, *supra* note 58, at 1093–125 (discussing reasons why

things is larger than the small set of things that the specification discloses in full. It includes at least all things that the PHOSITA could have made and would have thought of making after reading the disclosure. Assuming novelty and nonobviousness, the set of things that is disclosed or made available to the public is the core of patent protection. By definition, improvements lie beyond this core.[134] Whether a patent claim can extend beyond this core to encompass improvements is contingent on both the language employed by the patent drafter and the rules of claim construction, validity, and infringement that the patent regime enforces.

For an earlier patent to literally encompass a later improvement, the claim must be drafted by the patent drafter with broad, generic language that describes the later-developed improvement.[135] The need for the claim language to describe the improvement puts the availability of patent protection for later-developed improvements in part at the mercy of the patent drafter. The patent drafter bears the burden of describing the set of claimed things generically enough that the language describes yet-to-be developed products.[136] A patent drafter who fails to recognize the unnecessary limitations that are in the claim may not obtain protection for an inventor that encompasses later-developed improvements even if, as an normative matter, the case for giving the inventor such protection is strong. To hold constant the variability in claim scope that can be attributed to the skill of patent drafters, this Article assumes that patent drafters always draft the broadest permissible claims.

Assuming that the patent drafter did not make an obvious error, there are a number of doctrines that courts can invoke to sculpt the patentee's protection. Some determine the permissible level of generality at which a claim can be drawn. Claim construction—the process through which judges

enabled claims can reach into after-arising technology). If claims could not encompass any things that were not disclosed by the specification, there would be no literal patent protection for later-developed improvements.

134. *See supra* Section III.A.2.

135. The need to describe the allegedly infringing thing with a claim is not an issue that is particular to improvements. Failure to draft a sufficiently generic claim can also result in a failure to obtain rights to exclude from things that are disclosed and made available to the public by the specification.

136. For example, a later-developed, improved coffee sleeve that is folded in a nonobvious geometry or that is made of a later-developed material would be likely to fall within the scope of the claim to "an insulating band tapered in the shape of a truncated cone to fit the conical outer surface of a disposable coffee cup." However, a later-developed coffee sleeve in the shape of a doughnut likely would not, as a doughnut shape is likely not "tapered in the shape of a truncated cone" and it arguably is not a "band" at all (although these conclusions would be actively debated by the parties during claim construction).

determine the meaning of claim language to the PHOSITA—can expand or restrict the reach of a patent into improvements.[137] The validity doctrines, including the disclosure doctrines of enablement and written description, can effectively narrow claim scope if courts invalidate claims drafted without many limitations as incommensurate and uphold claims drafted with more limitations.[138] In addition, a court can expand or contract patent protection beyond the literal scope of a claim through either the doctrine of equivalents ("DOE")[139] or the reverse DOE.[140]

This overview of the patent doctrine that courts use to dole out protection for later-developed improvements overlooks many nuances, but it is designed to make only a simple point. The distinction between classic and overlooked improvements is nowhere to be found in the relevant patent doctrine that determines the reach of claims into after-arising technology, or at least nowhere on its rhetorical surface. The differential treatment afforded to these two types of improvements can be seen only in the outcomes of cases, not in how courts explain the outcomes. That is, it can be seen in what courts do, but not in what they say.

IV. CLASSIC IMPROVEMENTS AND SUCCESSIVELY INVENTED PROPERTIES

This Part defines a classic improvement—the type of improvement that is wound up in the stories that undergird the conventional theory on patent protection for improvements—in terms of the innovative properties produced by successive generations of inventors. Consider a hypothetical improvement story based on a simple technology. Abby is an earlier inventor

137. When defining the meaning of claim language to the PHOSITA, courts have leeway to look both to dictionary definitions and the particular way in which words are used in the specification (and thus to embodiments disclosed in the specification). Phillips v. AWH Corp., 415 F.3d 1303, 1322–24 (Fed. Cir. 2005) (en banc). The more heavily courts rely on the specification as an interpretive source, the more closely the scope of a claim is likely to be restricted to the disclosed embodiments and the more likely it is to exclude later-developed improvements. More drastically, claim construction is supposed to determine the meaning of the words at the time of filing, and some patent opinions have suggested that words construed at the time of filing categorically cannot describe later-developed technologies. Collins, *supra* note 66, at 550–53 (discussing the fixation of denotational, rather than ideational, meaning during claim construction).

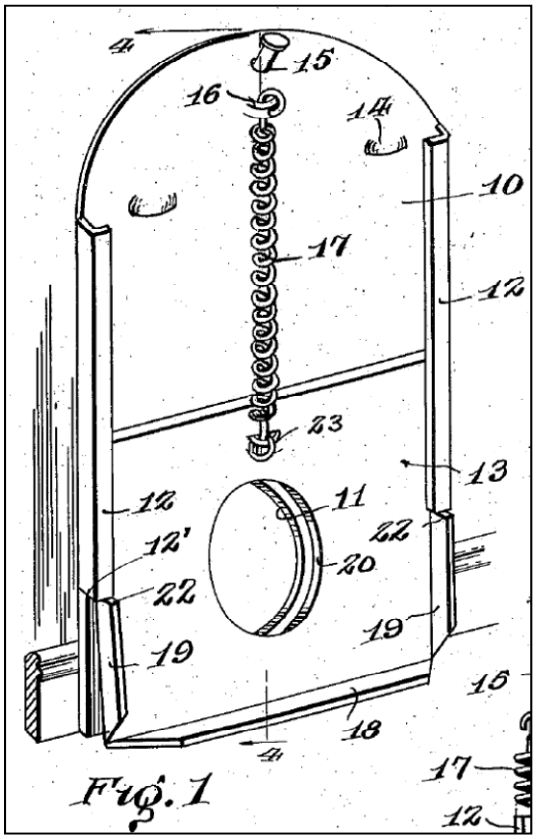
138. *See supra* notes 130–32 and accompanying text.

139. It is widely acknowledged that the DOE can expand a patentee's protection beyond literal claim scope and into after-arising technology. Warner-Jenkinson Co., Inc. v. Hilton Davis Chem. Co., 520 U.S. 17 (1997).

140. Scripps Clinic & Research Found. v. Genentech, Inc., 927 F.2d 1565, 1581 (Fed. Cir. 1991).

in the art of mousetraps. Before Abby's invention, the state-of-the-art mousetrap was an upside-down box over a piece of cheese with a short stick holding one side of the box above the floor. The person attempting to catch the mouse would tie a string to the stick, wait nearby, and pull when a mouse went under the box. Abby invents the first spring-loaded mousetrap: a device that stores potential energy in a spring and that uses the jostling motion caused by the presence of a mouse to release kinetic energy, trapping or killing the mouse. The working example of a mousetrap that Abby actually conceives and discloses in her patent is illustrated in Figure 1:

Figure 1: Abby's Two-Plate Mousetrap[141]



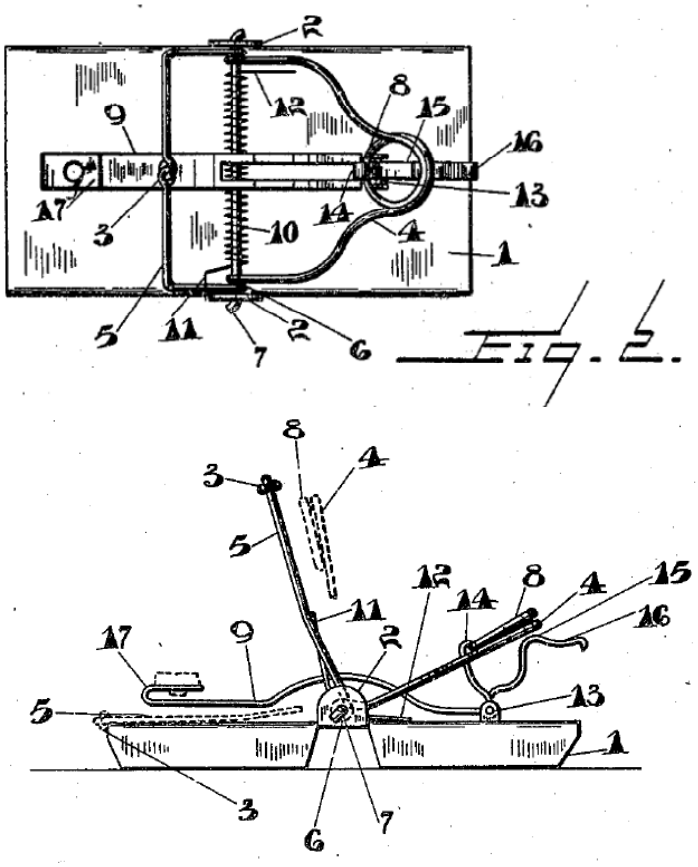
There are two plates, each with a hole in the center and one being able to slide in relation to the other. Cheese is placed in a box with Abby's mousetrap forming one side of that box, or the mousetrap is placed over a mouse hole in the wall. The spring must be stretched from its resting

141. Figure 1 is taken from U.S. Patent No. 2,059,164 (filed Dec. 2, 1935), but the facts of the hypothetical are fabricated to provide a simple teaching tool.

position for the two holes to align so that a mouse can attempt to pass through. Although there is a stop mechanism that can hold the holes in alignment and keep the spring in tension, the jostling motion caused by a mouse passing through the aligned holes destabilizes the stop mechanism, allowing the spring to shift one plate with respect to the other, trapping or killing the mouse.

After Abby files her patent, Bernard invents an improved spring-loaded mousetrap that is, more or less, the standard mousetrap design that one can still buy in the local hardware store today. General familiarity with such mousetraps is presumed:

Figure 2: Bernard's Fixed-Base Mousetrap[142]



Where Abby's trap keeps the spring in tension longitudinally, Bernard's trap places a torsional force on the spring. Where Abby's trap involved

142. Figure 2 is taken from U.S. Patent No. 1,342,255 (filed Apr. 30, 1919), but the facts of the hypothetical are fabricated to provide a simple teaching tool.

sliding plates, Bernard's trap has a wire moving in an arc in relation to the base. Bernard's mousetrap is a patentable improvement over the disclosure of Abby's patent.[143]

The Abby-Bernard hypothetical reaffirms that there are two distinct technological advances made at different points in time whenever the question of the reach of patent scope into improvement is raised. First, there is the advance in technical knowledge that justifies the issuance of the earlier patent whose scope is at issue. The things described by a valid patent claim must embody an advance by definition as a doctrinal matter: the advance explains why the things described by a claim satisfy the retrospective validity requirements of patent law.[144] Second, every improvement embodies at least one advance produced by the improver that occurs after the time the patent is filed.[145]

The Abby-Bernard hypothetical strongly resembles many of the historical examples of earlier patents and later, allegedly-infringing improvements.[146] The family resemblance follows from two facts about the way in which the later advance relates to the earlier advance.

First, considering the successive advances not as embodied in the improvement as properties but rather as ideas per se, the work of the later inventor builds on the work of an earlier inventor in a strong sense of the word. The inventors' contributions to progress are cumulative in that the later inventor must stand on the shoulders of the earlier inventor even to be in a position to make her contribution to technological progress.[147] But for Abby's general idea of a spring-loaded mousetrap, Bernard would not have been in a position to make the advance that he did when he did. But for Abby's shoulders, Bernard would have had to make the more fundamental

143. However, whether Bernard actually seeks patent protection is irrelevant to the hypothetical. The only question at issue is whether the improvement falls within the scope of the earlier-issued patent. The answer to this question does not depend on whether Bernard seeks patent protection.

144. *See supra* text accompanying note 129 (noting the role of the novelty and nonobviousness requirements).

145. *See supra* Section III.A.2.

146. *See, e.g.*, *supra* text accompanying notes 29–32, 40 (discussing the Wright Brothers/Curtiss patent dispute); *see also* Bessen & Maskin, *supra* note 2, at 613 (citing historical examples of classic improvements to define the concept of "sequential" innovation); Merges & Nelson, *supra* note 2, at 884–97 (citing historical examples of classic improvements to define "cumulative" innovation).

147. The shoulder-standing may be constructive rather than actual, given that actual facilitation is not a criterion of an improvement. *See supra* note 106.

advance of the spring-loaded mousetrap himself in order to produce his particular mousetrap.[148]

Second, considering the advances as they are embodied in the improvement as properties, the later advance *refines* the earlier advance. The refinement relationship of the classic improvement is possible because Abby's earlier patented technology embodies a technological advance that can be described at different levels of generality. At a high level of generality, the advance embodied in Abby's mousetrap is the very idea of a spring-loaded mousetrap itself—i.e., the idea of a mousetrap that can store potential energy in a spring and automatically unleash mouse-trapping kinetic energy in response to the presence of a mouse. In other words, the claimed mousetraps embody Abby's general idea because they possess the property *being a device in which the jostling motion of a mouse transforms the potential energy stored in a spring into the kinetic energy required to catch a mouse*. At a lower level of generality, the advance embodied in Abby's mousetrap is the idea of storing potential energy in an elongated spring and releasing it as kinetic energy in the form of one plate with a hole that slides in relation to another. Thus, a mousetrap embodies Abby's newly-discovered, specific idea in part because it has the property *being made of sliding plates*.

Because the properties that instantiate Abby's advance can be described at different levels of generality, the properties that instantiate Bernard's later advance in the improvement can relate to properties that instantiate Abby's advance in multiple ways at the same time. The after-arising advance that gives rise to Bernard's improvement is a new mechanism for storing potential energy and using kinetic energy to catch a mouse. In other words, a mousetrap embodies Bernard's newly discovered idea because it has the property *being made of a wire that can move in an arc in relation to a fixed base*. In relation to Abby's most general innovative property, Bernard's innovative property compounds with, or adds itself to, Abby's property.[149] The improvement-as-thing—that is, the improved mousetrap produced by Bernard—possesses both the property *being a device in which the jostling motion of a mouse transforms the potential energy stored in a spring into the kinetic energy required to catch a mouse* (Abby's general property) and the property *being made of a wire that can move in an arc in relation to a fixed base* (Bernard's innovative property).

148. The metaphor of one inventor standing on another's shoulders is not entirely accurate, as the platform upon which the later inventor builds is only part of the earlier inventor's contribution. A classic improvement is perhaps more akin to a piggy-back ride: the later inventor gets the advantage of some of the earlier inventor's height, but not all of it.

149. "Compound" is used here loosely in the pharmacological sense of the word—to mix two entities together. It is not used in the financial sense of compound interest.

However, the property that instantiates Bernard's advance in the improvement supplants or replaces the property that instantiates Abby's more specific advance. By storing potential energy in a torsional force, Bernard's mousetrap no longer stores energy in a longitudinal force. Because it has the property *being made of a wire that can move in an arc in relation to a fixed base*, the improvement no longer has the property *being made of sliding plates*. Any single, indivisible mousetrap will embody either Abby's specific idea or Bernard's specific idea, but not both.[150]

V. OVERLOOKED, "EASY" IMPROVEMENTS AND SUCCESSIVELY INVENTED PROPERTIES

The conventional theory on patent protection for improvement implicitly focuses on classic improvements, but not all improvements fit the mold of a classic improvement. Elaborating on the hypothetical presented in Part IV, *supra*, Section V.A defines an overlooked improvement in terms of the innovative properties generated by successive inventors. Section V.B offers an illustrative list of three scenarios in which overlooked improvements are likely to occur.

A. AN EXAMPLE AND ITS GENERALIZATION

Taking Abby's invention of the spring-loaded mousetrap presented above as the earlier-patented invention,[151] assume that Bob, too, produces an improved mousetrap. Bob invents a nonobvious metal alloy that makes cheaper, better springs.[152] Bob then manufactures sliding-plate mousetraps that follow the precise arrangement of mechanical parts that Abby discloses in her specification, except that he makes his springs out of his after-arising alloy.

Like Bernard's mousetrap, Bob's mousetrap is clearly an improvement. It is an after-arising thing that was not disclosed or made available to the public

150. It is possible for a mousetrap to embody both specific ideas in the sense that Abby's mousetrap and Bernard's mousetrap can be glued together to form a double-wide mousetrap with two trigger mechanisms. However, this physical aggregation of the earlier- and later-invented things presents an after-arising component issue, not an improvement issue. *See supra* text accompanying notes 110–16 (discussing disclosed-thing cumulative innovation cases that do not involve improvements).

151. *See supra* Part IV.

152. Any one of a number of different advances could underlie Bob's discovery. He may have been the first to conceive of a molecule with a particular chemical structure, or he may have been the first to figure out how to make a long-desired compound. This distinction is of importance in determining the reach of a patent into improvement in some contexts, but it is irrelevant in the hypothetical presented in the text.

by Abby's patent specification and that does not contain a thing disclosed by Abby's patent as a component; it still embodies Abby's technical advance in some way.[153] Yet, Bob's mousetrap differs from Bernard's mousetrap—and all classic improvements—in two important ways.

First, considering the successive advances not as embodied in the improvement as properties but rather as ideas per se, the improver does not need to stand on the shoulders of the earlier patentee to achieve her technical advance. Bernard, the classic improver, got a boost from Abby in order to be in a position to achieve his advance.[154] Bob does not. Bob achieves an advance in metallurgy. The technical barrier confronting Bob would be the same whether Bob produces his advance before or after Abby makes her advance in mousetrap technology.[155] As idea generators who contribute to technological progress, Bob and Abby stand side by side, not one on the shoulders of the other. However, the improvement-as-thing exists only because of the accumulation of the two advances. To risk stretching a metaphor too far, the improvement can be held aloft only by the concerted effort of both Bob and Abby. The improvement is the result of cumulative innovation only in the weak sense that two advances both must occur for the improvement to be produced, but neither one of the advances in knowledge builds on a platform provided by the other. Each advance rests on its own technological bottom in a different art.[156] It is possible to imagine getting to

153. *See supra* Section III.A (defining an improvement).

154. The boost may have been constructive. *See supra* text accompanying note 106.

155. Interestingly, however, if Bob were to invent first, Abby's later-developed mousetrap would technically not be an improvement on Bob's patent claim. Bob would likely claim a new composition of matter, which is an extremely narrowly framed thing that would be literally present as disclosed by Bob's earlier patent in Abby's later-improved mousetrap. If Bob's advance were to come first, Abby's after-arising advance would create a variant of the component problem, not an improvement problem. *See supra* notes 110–16 and accompanying text (distinguishing the after-arising component problem from the improvement problem). This asymmetry—the need for Abby's patent to grow in scope over time to encompass an improved thing but the lack of a need for Bob's patent to grow in scope over time to encompass the same improved thing—demonstrates one of the biases that the nature of things introduces into a peripheral claiming regime. *See* Collins, *supra* note 66, at 514–36 (discussing the importance of "thing construction" in patent law).

156. Professor Tim Holbrook has argued that the DOE should more readily encompass after-arising technologies when the later advance occurs in a field of endeavor that is different from the field of endeavor of the patent. Timothy R. Holbrook, *Equivalency and Patent Law's Possession Paradox*, 23 HARV. J.L. & TECH. 1, 37–40 (2009). Professor Holbrook justifies this argument on a "fairness principle." *Id.* at 7. This Article argues that the divergence of the technological fields of the successive advances is a relevant factor—but not the only factor—when determining the conditions under which literal claim scope—not simply the DOE—should encompass after-arising technology. It also argues that the distinction can be explained as a matter of efficiency, not fairness. *See infra* Part VII.

the same endpoint (that is, producing the same technological thing) with the two advances occurring in the opposite order. This inversion is not possible in a classic improvement, because the earlier inventor's advance is a foundation for the later inventor's advance.[157]

Second, considering the advances as they are embodied in the improvement as properties, the earlier and later advances are both fully embodied in the improvement-as-thing. In the improved mousetrap, the properties that instantiate Bob's after-arising advance only compound with, and do not supersede in any way, the properties that instantiate Abby's earlier-patented advance. This relationship of pure addition exists regardless of the level of generality at which Abby's advance is framed. Bob's improved mousetrap embodies his after-arising advance (the idea of the new alloy) as well as both Abby's earlier-patented general advance (the idea of a self-actuating, spring-loaded mousetrap) and her earlier-patented specific advance (the idea of an elongated spring connected to a sliding plate). Bob has displaced some properties of the things that constitute Abby's patented technology—e.g., the property of *being made out of an earlier-existing metal*—but only those properties that have nothing to do with the properties that embody Abby's innovative ideas. Bob did not make an advance in the arrangement of the mechanical components in a mousetrap design; Abby did not make an advance in the molecular structure of the mousetrap parts.[158]

In sum, the sets of properties that instantiate the successive inventors' innovative ideas in an improvement are effectively independent of one another. One set can be altered without mandating any change in the other set. A thing can have two distinct sets of properties, each of which is capable of being altered or changed within certain parameters without having a significant impact on the other. When the successive advances wound up in improvement stories are manifest in properties of improved things that are

157. One can imagine a possible world in which Bernard's mousetrap is produced before Abby's. In this possible world, however, Bernard must make a different technological advance than he did in the actual world. He must do what Abby did in the actual hypothetical, namely generate the general idea of the spring-loaded mousetrap.

158. If Abby's advance were characterized as the idea of making a spring-loaded mousetrap out of then-existing metals, then Bob's innovative properties could be viewed as part-superseding properties. Thus, the distinction between classic and overlooked improvements may be one of degree rather than kind. However, the characterization of Abby's advance as the idea of making a spring-loaded mousetrap out of then-existing materials is misleading. It elides the things that Abby invented with the contribution to the storehouse of knowledge that she made. To describe the thing that Abby made in full, it is necessary to note that she worked with existing metals, but her use of existing metals is in no way necessary to describe her advance, i.e., her marginal contribution to the progress of technical ideas.

independent variables, the advances can compound with each other (and not supplant each other) in the improvement.

B. THREE REASONS FOR PROPERTY INDEPENDENCE

Things are complex entities. There is no single, catch-all explanation for why two sets of properties of a thing are effectively independent variables and thus for why successive advances that alter those properties lead to an overlooked improvement. This Section offers three distinct underlying reasons for property independence in the things claimed by a patent: claimed things with naturally independent properties, claimed things with properties that have engineered independence, and claim language that recites prior-art context.[159] In the course of identifying these three groups of overlooked improvements, this Section also provides evidence to demonstrate that they are "easy" cases in the sense that courts routinely allow earlier-filed patents to encompass overlooked improvements without so much as raising a caution flag.[160]

1. Things with Naturally Independent Properties

Some properties of an indivisible thing are, within certain bounds, naturally independent. A simple example in the mechanical arts is the property of shape and the property of materiality.[161] One can make two things that differ in shape without requiring any difference in materiality; one can make two things that differ in materiality without requiring any difference in shape.[162] If successive advances yield first the geometry and then the materiality of an improved mechanical device, there is no refinement

159. These three categories are intended as an illustrative, not exhaustive, list of conditions that are fertile for the development of overlooked improvements.

160. *See supra* note 6 (defining an "easy" case). In fact, courts may not even recognize overlooked improvements as being later-developed things at all. *See infra* note 177.

161. In the chemical and biochemical arts, however, materiality and shape are not independent. The materiality of a molecule is its atoms, and a molecule's atoms determine its shape. The fact that there are fewer properties that are naturally independent variables when claims describe inventions on a molecular scale explains in part why claims to mechanical inventions are widely viewed as reaching farther into after-arising technology than claims to chemical and biochemical inventions are. *Cf.* Spectra-Physics, Inc. v. Coherent, Inc., 827 F.2d 1524, 1533 (Fed. Cir. 1987) (suggesting that narrow disclosures can enable broad claims in the mechanical arts). There are fewer overlooked improvements when claims describe inventions on a molecular scale because molecules do not have sets of intrinsic properties that can vary independently.

162. There are limits to the independence. For example, claimed shapes or arrangements of parts may be defined in part by the functions that they perform, and not all materials allow the shapes or arrangements to perform the required functions. A mousetrap with a spring made of cheese would be an inoperative mousetrap.

or supplanting of the successively invented properties. The properties compound with each other in the improvement-as-thing, yielding an overlooked improvement. Bob's improvement on Abby's patented invention is an overlooked improvement for this reason.[163] Similarly, assume that an earlier inventor invents and patents a new mechanical device—say, a plastic gizmo that controls Venetian blinds—having some nonobvious interrelation among its parts and that a later inventor invents a new material—say, a more durable, cheaper type of plastic. Venetian blinds gizmos that are improved because they are made out of the new plastic are overlooked improvements.[164] To the same end, imagine an earlier patent on a pill that is formed into a new shape that is easier to swallow and the later invention of a new chemical that is an effective pharmaceutical. A pill that is made in the patented shape and that contains the new chemical is an improvement over the patent that discloses the pill shape. It was not disclosed to the public in the earlier-filed patent on the pill shape; the literal scope of a claim to a "pill" or "pharmaceutical compound" molded into the specified shape must expand over time to encompass pills made out of the after-arising chemical.[165] More specifically, the after-arising pill is an overlooked improvement. The properties that embody the earlier-patented advance (geometry) and the properties that embody the later-developed advance (materiality) are, within certain limits, naturally independent.

Overlooked improvements that arise from successive advances that are embodied in naturally independent properties routinely wind up within the literal scope of earlier-filed claims. For example, the literal scope of a patent on an advance in the mechanical arts routinely extends into mechanical devices that are improved because they are made out of after-arising materials.[166] This rule reflects the common-sense position that a patent on a doorknob encompasses after-arising, improved doorknobs made out of newly invented materials because a "doorknob is a doorknob"—i.e., it is still

163. *See supra* Section V.A.

164. *Cf.* Collins, *supra* note 58, at 1111–22 (elaborating on this hypothetical).

165. *See supra* Section III.A (defining an improvement).

166. *See* Feldman, *supra* note 41, at 28; Meurer & Nard, *supra* note 41, at 1976–77. Of course, a claim in the mechanical arts cannot literally read on a device made of any after-arising material if materiality is a strict claim limitation. For example, a claim to a "plastic widget" cannot literally read on a widget made out of an after-arising metal, even if it might be able to grow over time to encompass a widget made from an after-arising plastic. *Cf. infra* note 327 and accompanying text (discussing improvements to a hypothetical claim to a "plastic widget").

the earlier inventor's invention on every relevant level of generality—regardless of the material out of which it is made.[167]

2. Things with Properties Engineered To Be Independent

In some technologies, the independence of two groups of properties of a single, indivisible thing is engineered, not natural. For example, consider the computer industry in which hardware and software are often developed separately.[168] The functional properties of a programmed computer are, within certain limits, independent of the physical properties of the computer that executes the program.[169] The gates and switches can be shuffled and reshuffled, and yet the computer—whatever its final internal configuration—can still perform the same software-scripted functions.[170] Software can run on all types of computers with markedly different physical architectures. In fact, "[p]resent-day computers are built of transistors and wires, but they could just as well be built, according to the same principles, from valves and water pipes, or from sticks and strings."[171] The independence of the functional capacities of software and the physical characteristics of computers is not a natural phenomenon. It exists only because it has been engineered. It is only because the computer industry has developed a set of technical standards and intermediary technologies that software can run on a wide variety of hardware.[172]

Because software and hardware have engineered independence, computer-related technologies give rise to many overlooked improvements. Assume an earlier "apparatus" claim to a software invention—that is, a claim to a physical computer that has been programmed with newly developed software.[173] Now, assume the later development of new computer hardware on which the software can be executed. The improved thing—that is, the after-arising hardware executing the earlier-claimed software—is clearly an

167. Feldman, *supra* note 41, at 3.

168. Microsoft Corp. v. AT&T Corp., 550 U.S. 437, 450 (2007) ("Software . . . is a stand-alone product developed and marketed 'for use on many different types of computer hardware . . . .' ").

169. *See* W. DANIEL HILLIS, THE PATTERN ON THE STONE, at ix (1998) ("Computers are understandable because you can focus on what is happening at one level of the hierarchy without worrying about the details of what goes on at the lower levels.").

170. Again, the reshuffling can only be done within limits. Many configurations of gates and switches—including anything that I could make—cannot be exchanged for a functioning computer.

171. HILLIS, *supra* note 169, at viii.

172. *See, e.g.*, *id.* at 56–58 (discussing interpreters and compilers).

173. Apparatus claims to software inventions are commonplace. *See, e.g.*, Arrhythmia Research Tech., Inc. v. Corazonix Corp., 958 F.2d 1053 (Fed. Cir. 1992).

improvement. It is an after-arising thing that is not disclosed or made available to the public in the specification of the software patent and that does not contain any disclosed thing as a component; it still embodies in some way a technical advance that is attributable to the software inventor.[174] Yet, the later advance is embodied in a set of properties of the improved thing that, within bounds, has engineered independence in relation to the set of properties that embody the earlier advance.

Overlooked improvements that can be traced to properties with engineered independence also routinely wind up within the literal scope of earlier claims. Here, the proof of a negative proposition is required. The effective life-span of computer hardware is extremely short, as hardware becomes outdated within several years of its purchase. The life of software and software patents, however, is much longer. Software patents have a duration of twenty years from the date on which the patent is filed, and their functionally-defined claims mean that they encompass many iterative versions of the software programs marketed to the public.[175] Therefore, running earlier-developed software on after-arising hardware must be commonplace. If software patents were made obsolete every time a new generation of computer hardware arrived on the market, then all of the arguments made today about the detrimental effects of software patents[176] would be moot as an economic matter, yet they clearly are not. If software apparatus claims did not literally encompass software being run on after-arising hardware, one would expect judicial opinions finding noninfringement of software patents for this reason. However, none of these cases exist. No argument of non-infringement of a software apparatus claim has ever been accepted by a court based on the fact that the allegedly infringing device incorporates later-developed hardware. In sum, software apparatus claims literally encompass overlooked improvements routinely in the contemporary patent regime.[177]

---

174. *See supra* Section III.A (defining an improvement).

175. Some software claims may become obsolete before their terms expire because the market no longer demands the claimed functionality.

176. *See, e.g.*, BESSEN & MEURER, *supra* note 28, at 187–214.

177. In fact, after-arising hardware programmed with earlier-patented software is likely not even recognized as an improvement. Software apparatus claims are commonly drafted with "means-for" limitations, such as "a means for choosing a random number." Means-for limitations (and all other purely functional claim limitations) are construed under the special rules of claim construction set forth in 35 U.S.C. § 112, ¶ 6 (2006). It is black-letter law that means-for claim limitations cannot encompass after-arising technology. *See* Chiuminatta Concrete Concepts, Inc. v. Cardinal Indus., Inc., 145 F.3d 1303, 1310–11 (Fed. Cir. 1998). However, given the rapid evolution of computer hardware, it is likely that means-for limitations in computer software patents routinely encompass after-arising hardware

The inverse scenario, too, gives rise to overlooked improvements that routinely fall within literal claim scope in the contemporary patent regime. Imagine an earlier claim to hardware, and the later development of a new software operating system. The hardware patentee brings a suit against someone who is running the new operating system on the precise hardware disclosed in the patent specification. It is black-letter law in the area of software patents that a machine programmed with new software is a new machine for the purposes of novelty and nonobviousness,[178] so a computer programmed with the nonobvious software is an after-arising thing in relation to the disclosure of the earlier hardware patent. Again, however, no argument has ever been addressed in the Federal Circuit that earlier-patented computer hardware is beyond the reach of the hardware patent just because it is running later-developed software.

3. Claims with Prior-Art Context Limitations

Sometimes, the compounding of successively invented properties in an improvement is not grounded in the independence of the properties of any single, indivisible thing. Rather, the compounding derives from the way in which claims frame the things. In a peripheral claiming regime, inventors often file both independent claims describing the thing that they have invented and dependent claims that recite as limitations both that thing and some of the prior-art context in which it is found. For example, the inventor of an eraser may claim both an "eraser" in isolation and an "eraser attached to a pencil." Generically formulated, an inventor who has invented thing A may claim both "A" and "A+B," with B being a thing that is divisible from A and part of the prior-art context in which A is often found. By filing a claim to "A+B," the inventor has framed the things claimed by the patent to include more physical matter. As more context limitations are added to the claim language, the claimed thing becomes physically larger and the claim scope becomes smaller.[179]

executing the disclosed software. The point here is not that the literal interpretation of black-letter law on the construction of means-for claims should prevent the reach of software patents into later-developed hardware. Rather, the point is that the courts have not yet realized the implications of a per se bar on the literal infringement of after-arising technology. Some overlooked improvements raise such "easy" cases that they simply have not been recognized as improvements at all.

178. *In re* Bernhart, 417 F.2d 1395, 1400 (C.C.P.A. 1969).

179. Patent applicants file claims with prior-art context limitations as a safety net. Even if a claim to "A" turns out to be obvious, the claim to "A+B" may be nonobvious because combining A with B may generate unexpected properties. Additionally, there are several reasons why patent applicants might recite prior-art context limitations as the broadest, independent claims of a patent. The problem may be a conceptual error on the part of a

Claims with prior-art context limitations give rise to overlooked improvements whenever it is the prior-art context rather than the invented thing that is improved. Assume that A is a newly invented thing, that B is a prior-art limitation, and that the inventor claims both "A" and "A+B." A later improver comes up with an improved version of B—call it B′. With respect to the claim to "A," the invention of B′ is a later-developed component, not an improvement. The claim to "A" can read on A+B′ because the A present in the combination of A+B′ is the same old A disclosed in the earlier patent.[180] However, with respect to the claim to "A+B," the invention of B′ does yield an improvement. As defined by the claim, the relevant thing that needs to have been disclosed in the earlier patent specification is A+B′, not simply A. The set of things described by the claim "A+B" must grow over time after the claim has been filed if A+B′ is to infringe. The set of things encompassed within a claim to an "eraser" need not expand over time after filing if it is to encompass an eraser attached to an after-arising pencil, but the set of things encompassed within a claim to an "eraser attached to a pencil" does.

Improvements that alter the properties of the not-inventive-yet-claimed context of an earlier invention wear their status as overlooked improvements on their sleeves. Property independence corresponds to thing independence—or, more accurately, to divisible sub-thing or divisible part independence. If the claimed things are composed of more than one distinct sub-thing or part, the properties of the independent distinct sub-things or parts can vary independently. The earlier advance may be embodied in one part of the claimed thing, and the later advance may be embodied in a distinct part.

There is no black-letter law stating that patents can expand over time to include improvements on prior-art, context limitations recited in a claim. However, to the extent that a disclosure doctrine would be called upon to invalidate the claim to "A+B" discussed above[181] as overbroad because it encompasses the allegedly infringing technology A+B′, the result would seem

claim drafter. Sometimes, B is so ingrained in prevailing conceptions of how A is used that it may not occur to the claim drafter to claim "A" apart from "A+B." *Cf.* Larami Corp. v. Amron, 27 U.S.P.Q.2d 1280 (E.D. Pa. 1993) (holding a claim to a new water-gun trigger mechanism was not infringed because the claim recited a gun "having a chamber therein for liquid" and the allegedly infringing technology had an external water reservoir). Patent drafters may also include context limitations because claims that describe larger physical entities may result in larger damages. Lemley, *supra* note 64, at 25–27.

180. *See supra* text accompanying notes 110–16 (discussing disclosed-thing cumulative innovation cases).

181. *See supra* text accompanying note 180.

to violate a basic principle of patent law. When the allegedly infringing technology is A+B′, the argument that the claim to "A" is enabled and possessed but that the claim to "A+B" is not would mean that an independent claim remains valid as a dependent claim is invalidated. However, dependent claims are widely presumed to be enabled if the independent claims from which they depend are enabled.[182]

For an anecdotal example of a claim that likely has a prior-art, context limitation, consider the Federal Circuit's opinion in *Superguide Corp. v. DirecTV Enterprises*.[183] The *Superguide* court addressed the scope and validity of claims based on the invention of interactive electronic television programming guides—the ones that replaced the supplements in the Sunday paper.[184] The court construed the meaning of the claim term "regularly received television signal,"[185] a likely example of a prior-art context limitation (as the inventor did not claim to have invented regularly received television signals). At the time the patent was filed, the regularly received signals were analog, but they were digital by the time of infringement. While a concurrence argued that the claim was limited to interactive electronic television programming guides that employed analog signals,[186] the majority allowed the claim to encompass digital signals.[187] This facet of the *Superguide* holding allowed literal claim scope to expand over time to encompass an overlooked improvement. The claim recited a prior-art context limitation. The allegedly infringing technology was an improvement because of an after-arising alteration of this context.

VI. VISUAL REPRESENTATIONS OF THE DISTINCTION

In a classic improvement, the properties that instantiate the later inventor's ideas in part compound with, and in part supplant, the properties that instantiate the earlier inventor's ideas. Figure 3 roughly represents this relationship[188]:

182. *See* Lefstin, *supra* note 63, at 1170–74 (discussing the paradox of non-enabled dependent claims).

183. Superguide Corp. v. DirecTV Enters., 358 F.3d 870 (Fed. Cir. 2004).

184. *Id.* at 873.

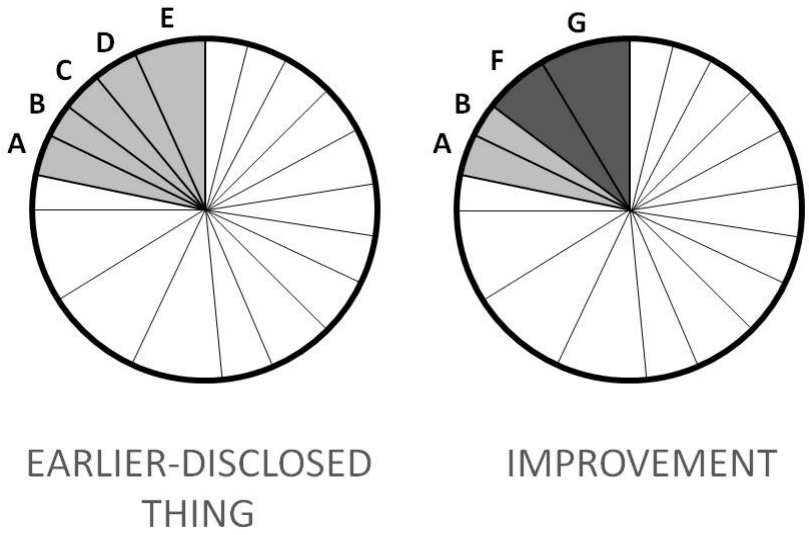
185. *Id.* at 876.

186. *Id.* at 897 (Michel, J., concurring).

187. *Id.* at 876–81 (majority opinion).

188. The representation is rough in part because it fails to capture the fact that there is a hierarchy among properties of different levels of generality, i.e., that some more general properties are entailed by the presence of more specific properties.

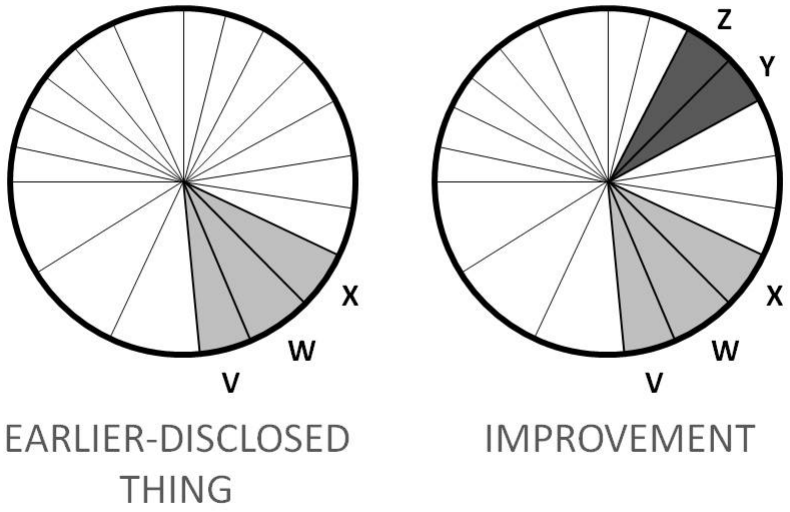
Figure 3: Classic Improvement



In Figure 3, the circles are things, and the pie-shaped wedges are the properties that comprise the things. On the left are the things disclosed and claimed by an earlier patent. Properties A, B, C, D, and E all instantiate the earlier inventor's advance. The other unlabeled properties do not embody the inventor's ideas. On the right is the improvement. Properties F and G instantiate the improver's ideas in the improvement. Properties A and B persist in the improvement. They compound with properties F and G. Properties C, D, and E do not persist in the improvement. They have been supplanted by properties F and G.

In contrast, as Figure 4 illustrates, the properties that instantiate the later advance in an overlooked improvement do not supplant any of the properties that instantiate the earlier advance:

Figure 4: Overlooked Improvement



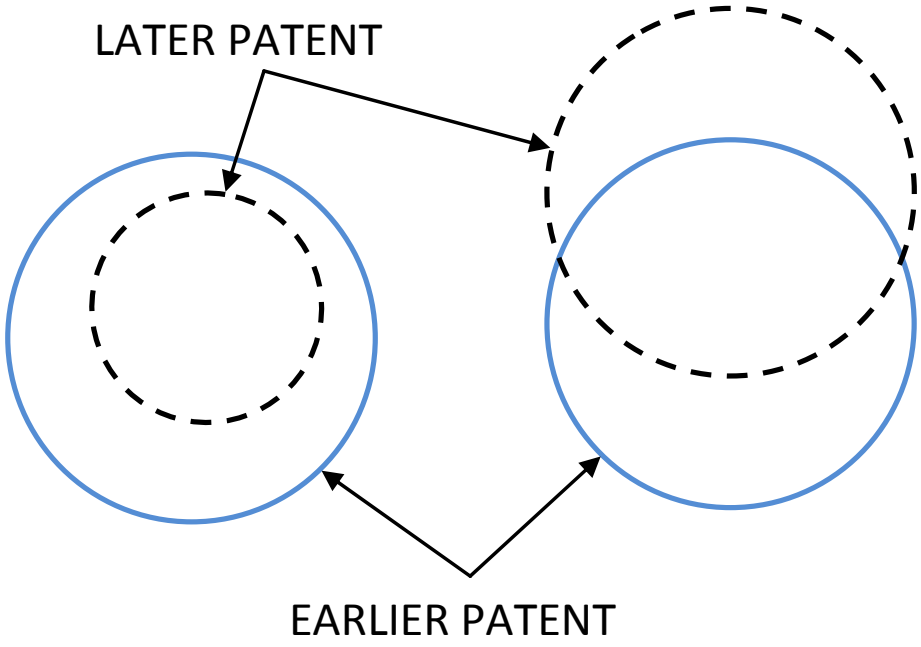
In Figure 4, the properties V, W, and X are all of the properties that instantiate the earlier inventor's advance in the things disclosed by the earlier patent, and they persist in the improvement. The properties Y and Z instantiate the improver's advance in the improvement. They compound with properties V, W, and X in the improvement, displacing only properties that did not instantiate the earlier patentee's ideas.

Assuming that the later inventor also seeks patent protection,[189] Figure 5 illustrates that classic and overlooked improvements implicate different types of blocking patents[190]:

189. *See supra* note 4 (noting that an improver can patent an improvement).

190. Both of these types of blocking patents are overlapping blocking patents. *See infra* note 297 (distinguishing overlapping and economic blocking patents).

Figure 5: Relationship to Blocking Patents



Classic improvements often involve nested blocking patents (on the left of Figure 5), in which the improver's patent scope is entirely subsumed within the earlier inventor's patent scope. In a classic improvement that involves refinement, the earlier inventors' more general properties persist in the improvement and compound with the improver's properties.[191] If the earlier patentee is able to obtain a claim that encompasses the later-developed improvement, then the claim will reference the more general properties, and it will be drafted at a sufficiently high level of generality so to encompass all of the improver's patentable things.[192] Overlooked improvements are more likely to involve offset blocking patents (on the right of Figure 5). The earlier inventor and later improver usually work in distinct technological fields,[193] so

191. If the earlier inventor is denied a patent at that relatively high level of generality and is limited to a more specific level of generality that corresponds to the specific properties that are supplanted in the improvement by the improver's properties, then the earlier inventor's rights do not encompass the improvement at all, and there are no blocking patents.

192. *See infra* notes 242–48 and accompanying text (noting the relationship between the permissible level of generality of a valid peripheral claim and whether classic improvements infringe earlier-filed patents).

193. *See supra* text accompanying notes 154–57.

the later improver's invention is likely to have some embodiments that do not require the use of the earlier inventor's properties. An improved canvas can be used in devices other than airplanes;[194] Bob's after-arising metal can be used in devices other than spring-loaded mousetraps.[195]

VII. WHY CLASSIC AND OVERLOOKED IMPROVEMENTS MERIT DISTINCT TREATMENT

This Part uses a conceptual framework that identifies properties as the locus of invention to analyze the optimal reach of earlier-filed patents into later-developed improvements. Section VII.A presents a normative goal of the patent regime: patent claims should track in some way the set of things that embody inventors' innovative ideas. Section VII.B uses a focus on properties as the locus of invention to explain why this goal means that overlooked-improvement cases should be "easy" and classic-improvement cases should be contested. It also introduces the concept of the least-general naked property that should be added to the conventional theory on patent protection for improvements in classic-improvement cases. Section VII.C reveals a conceptual bonus of focusing on properties as the locus of invention. The innovative properties generated by successive inventors can be identified as pure complements in the overlooked-improvement cases and as part-complement, part-substitute mixtures in the classic-improvement cases, allowing these well-established economic concepts to have a newfound relevance to crafting claim scope.

A. CRAFTING A MARKET FOR *IDEAS* FROM RIGHTS THAT GOVERN *THINGS*

This Section establishes the normative concern that, in the following Section, is used to drive the analysis of patent protection for improvements. It states two basic principles of patent law and derives a logical corollary from them. First, patent rights should create a market for innovative ideas—that is, a market in which inventors are rewarded in rough proportion to the social value of their ideas, as measured by consumers' willingness to pay. Second, patent rights propertize sets of things, not ideas per se. For a patent regime that respects both of these principles to function effectively, these two principles entail a third: the sets of things governed by patent rights must track the sets of things that embody inventors' ideas.

194. *Cf. supra* text accompanying note 40.
195. *Cf. supra* text accompanying note 152.

The first principle is that patents are a market-based solution to the problem of insufficient incentives for the generation of innovative ideas. Under an incentive-to-invent justification for patents, patents augment the inefficiently low incentives for individuals to invest in generating the innovative ideas that are needed to produce welfare-enhancing, innovative products.[196] The ideas required to produce technologically innovative things are often costly to generate. Yet, once the ideas are produced, competitors are often able to gain access to them cheaply and quickly by examining the products that inventors must sell in order to profit from their ideas. Under these conditions, rational actors may not invest in generating the innovative ideas. They may see down the road that marginal-cost pricing in a competitive market will deprive them of the ability to recoup the costs sunk into idea generation, so they choose to spend their time and money on endeavors other than research and development.[197] To address this incentives problem, the patent regime establishes property rights (or, at least, property-like rights)[198] in inventions. Such rights are not the only means of addressing inefficiently low incentives to invent.[199] However, they have the virtue of enlisting the distributed intelligence of the market to make decisions that the government is arguably not very competent to make because it lacks the required information.[200] If patentees get rights to exclude others from socially valuable uses of their innovative ideas, patent rewards will be proportional to the social value of inventors' ideas, with social value being measured by consumers' willingness to pay.[201] This proportionality, in turn, means that the

196. The incentive-to-invent justification of patent rights described in the following sentences is well established. *See, e.g.*, SUBCOMM. ON PATENTS, TRADEMARKS, AND COPYRIGHTS OF S. COMM. ON THE JUDICIARY, 85TH CONG., AN ECONOMIC REVIEW OF THE PATENT SYSTEM (Comm. Print 1958) (Study No. 15 prepared by Fritz Machlup, Johns Hopkins Univ.) (presenting a historical overview of several justifications of the patent regime, including the incentive-to-invent justification).

197. This before picture is not as bleak as it is sometimes made out to be. Some of the costs of invention can be recouped even without patent protection. *Cf.* Wesley M. Cohen, Richard R. Nelson & John P. Walsh, *Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)* (Nat'l Bureau of Econ. Research, Working Paper No. W7552, 2000), *available at* http://ssrn.com/abstract=214952 (providing empirical data of firms' use of patents and other mechanisms for appropriating profits from invention).

198. *See infra* note 210 (noting a debate over the extent to which property in tangible resources and intellectual property are birds of a feather).

199. *See* Brian D. Wright, *The Economics of Invention Incentives: Patents, Prizes, and Research Contracts*, 73 AM. ECON. REV. 691 (1983) (discussing prizes and government contracts for research).

200. *Id.* at 691–92; Scotchmer, *supra* note 2, at 30.

201. The alignment of the magnitude of patent rewards and the social value of patented inventions is commonly cited as a beneficial feature of patent protection. *See, e.g.*, FED.

magnitude of patent rents will reflect what consumers want, "thus channeling productive efforts" of inventors "in directions most likely to enhance consumer welfare."[202] Because of this proportionality—that is, because "the profit available from exclusive control of the innovation will be correlated with its social value"—a patent regime encourages potential innovators to perform the important task of "screen[ing] their ideas by comparing cost to some measure of expected social value."[203] In sum, to harness the power of the rational individuals' drive to maximize private welfare and use it to drive inventive activity toward the production of the most socially valuable ideas, patents must structure a market for ideas, i.e., they must create property rights that reward inventors in proportion to the social value of their ideas. For a patent regime to use a market to incentivize innovation and achieve allocative efficiency among possible research and development expenditures, patents on innovative ideas that generate larger welfare gains must yield larger profits for their inventors, all else being equal.[204]

TRADE COMM'N, THE EVOLVING IP MARKETPLACE: ALIGNING PATENT NOTICE AND REMEDIES WITH COMPETITION 2 (2011), *available at* http://www.ftc.gov/os/2011/03/110307patentreport.pdf ("By aligning the patentee's market reward with consumer preferences, competition in product and technology markets encourages investment in those inventions that are more likely to be valued by consumers."); Carl Shapiro, *Patent Reform: Aligning Reward and Contribution*, 8 INNOVATION POL'Y & ECON. 111, 113 (2008) ("[E]conomic efficiency is promoted when the rewards provided to patent holders are aligned with their actual social contributions.").

202. William Fisher, *Theories of Intellectual Property*, *in* NEW ESSAYS IN THE LEGAL AND POLITICAL THEORY OF PROPERTY 168, 178–79 (Steven R. Munzer ed., 2001) (citing Harold Demsetz, *Information and Efficiency: Another Viewpoint*, 12 J.L. & ECON. 1 (1969) (discussing a utilitarian theory of intellectual property based on optimizing patterns of productivity)). This argument is sometimes marshaled to argue that intellectual property should aspire to perfect internalization. *See, e.g.*, PAUL GOLDSTEIN, COPYRIGHT'S HIGHWAY 178–79 (1994). However, ensuring proportionality between an inventor's ideas and his rewards, but not perfect internalization, has some value in optimizing patterns of productivity. Although rational actors will not undertake some welfare-enhancing innovative projects because the private returns will not cover the costs of invention when internalization is not perfect, proportionality still effectively channels efforts among the technological endeavors that rational actors will undertake. Proportional, but not perfect, internalization also tempers a market for ideas in a way that, on balance, likely promotes the efficient generation of ideas. *See infra* notes 209–16 and accompanying text. Furthermore, assuming that perfect internalization is not possible in endeavors that do not lead to intellectual property rights, it is likely necessary to prevent over-investment in the generation of technological ideas. Arnold Plant, *The Economic Theory Concerning Patents for Inventions*, 1 ECONOMICA 30, 31–32 (1934). *Cf.* Glynn S. Lunney, *Reexamining Copyright's Incentives-Access Paradigm*, 49 VAND. L. REV. 483, 655–56 (1996) (raising this argument in relation to copyright law).

203. *See* SCOTCHMER, *supra* note 24, at 97.

204. The "all else being equal" caveat is important. *See infra* notes 209–16 and accompanying text (discussing reasons for diverging from proportionality of contribution and reward).

The second principle is that patents propertize sets of things. Although patents structure a market for innovative ideas, they do not propertize ideas per se. They do not meter the mental consumption of newly discovered knowledge itself: one cannot infringe a claim to a newly discovered mousetrap by thinking about the ingenious mechanism that links the cheese to the deadly consequences that await the hungry mouse, nor can one infringe by communicating to the public how to understand the mechanism, how to make it, and how to use it, even if one profits from the communication.[205] Rather, patents grant inventors rights to exclude others from making, using, selling, offering to sell, or importing innovative things.[206] One must make, use, sell, offer to sell, or import an instance of a thing that is actually capable of catching mice to infringe the patent.

Viewed in combination, these two principles reveal an important design feature of the patent regime that, in turn, offers guidance about how patent scope should be determined. Patent rights are a kind of drive-chain technology linking two gears, each of which is located in a distinct ontological realm. Rights to exclude from the specified uses of sets of innovative things (which exist in the material, extra-mental world) are created in order to create a market for innovative ideas (which exist, at least in part, in human minds).[207] This trans-realm design feature—an indirection built into the heart of patent law—has important implications for patent scope. For the drive-chain linkage to work properly, patent rights governing a set of things must reward an inventor in proportion to the value of his innovative ideas, and the set of things governed by a patent must track the set of things that embody an inventor's ideas.[208] As explored in the following Section, this

205. *See supra* notes 117–18 and accompanying text (noting that idea-only cumulative innovation cases never result in infringement).

206. 35 U.S.C. § 271(a) (2006). *But cf. supra* note 87 (noting that method claims are also eligible for patent protection).

207. The fact that the patent regime uncomfortably straddles the realms of ideas and things can be seen in the fact that the term "invention" in patent law is ambiguous. In one sense, inventions are widely recognized as ideas or mental realizations. *See, e.g.*, Pfaff v. Wells Elecs., 525 U.S. 55, 60 (1998) (stating that, in the context of the on-sale bar of 35 U.S.C. § 102(b), "[t]he primary meaning of the word 'invention' in the Patent Act unquestionably refers to the inventor's conception rather than to a physical embodiment of that idea"). It is in this sense of an invention that, according to legend at least, inventions can occur when a light bulb goes off in our minds when we are in the shower. In another sense, inventions are commonly identified as the things that are described by patent claims. *See* § 271(a) (defining direct infringement).

208. The indirection at the heart of patent law means that patent protection does not align the private value of a patent to an inventor with the social value of an innovative idea per se. Rather, patent protection tracks the social value of embodied ideas—ideas as they exist embodied in things. The value of ideas per se that are not embodied in things escapes

conclusion is an important driver of the different patent protection doled out to classic and overlooked improvements.

Structuring a market for embodied ideas is a goal of patent protection, but it is not the ultimate or only goal. It is a means to the end of promoting technological progress. The patent regime is not a purely formalist system in which all rules must reinforce the proportionality of the social value of an embodied idea and the private value of a patent. There are often good reasons to stray from strict proportionality.[209] For example, issues related to the cost of producing innovations may be relevant. A strict focus on the proportionality of the social value of ideas and a patentee's reward would mean that the costs of supplying the ideas required to produce innovative technology would be irrelevant to patent scope determinations.[210] It is

---

the patentees' control. In some cases, researchers who come up with valuable factual discoveries about the world (i.e., ideas per se) without inventing any nonobvious thing are denied patent protection altogether, despite the value of their contribution to technical progress. In patent lingo, these cases are often discussed as cases in which patent applicants are denied all patent protection because they claim a "law of nature" in the abstract rather than an application of the "law of nature." *See* Diamond v. Diehr, 450 U.S. 175, 185, 191–92 (1981). *But see supra* note 117 (noting that many discussions of unpatentable abstract ideas involve claim scope, not claims describing the use of knowledge per se). In other cases, even when some form of patent protection is available to the earlier innovator, the value of ideas per se is often an externality from the perspective of the patent-owning inventor. *See supra* notes 117–18 and accompanying text (discussing idea-only cumulative innovation cases).

209. For example, the indirection principle at the heart of patent protection and its consequence of leaving knowledge per se beyond the reach of patent rights, *see supra* note 208, can readily be defended as a feature, rather than a bug, of patent protection.

210. The supply-driven vision of what patents do (they cover fixed costs) and the demand-driven vision (they allow inventors to eat a part of what they kill or profit in proportion to their contribution in increasing social welfare as measured by willingness to pay) are sometimes at odds. WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 300 (2003) (noting the schism between the magnitude of patent rewards and the costs of supplying ideas). The two can be brought into alignment by assuming proportionality between the social value of an invention and the cost of supplying the ideas that are required to produce it, but this assumption often does not reflect reality. Broadly speaking, the schism between supply- and demand-side concerns underlies many of the economic debates over the extent to which patent law, and intellectual property more broadly, should be thought of as a property regime. The stronger the focus on providing sufficient incentives to overcome the problems of supplying innovative ideas, the less intellectual property rights act like property rights and the more they should sanction free-riding when appropriate. The stronger the focus on rewarding inventors in proportion to the value of their innovative ideas, the more intellectual property rights act like property rights. *Compare* Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 TEX. L. REV. 1031 (2005) (arguing against the property metaphor in intellectual property based on a supply-side goal of providing sufficient incentives), *with* John F. Duffy, *Intellectual Property Isolationism and the Average Cost Thesis*, 83 TEX. L. REV. 1077 (2005) (arguing that intellectual property should be treated like property and patents should be proportional to the value of

therefore conceivable to imagine a deviation from the proportionality principle in order to provide more incentives when the costs of supplying innovative ideas is high and fewer incentives when the costs of supplying innovative ideas is low.[211] Tempering the proportionality principle when costs of supplying innovative ideas are low and the market reward for those ideas is high also prevents wasteful races.[212] A strict focus on proportionality not only detaches the scope of today's patents from the costs of supplying yesterday's patented inventions, but it also ignores the costs that today's patents impose on tomorrow's innovation. It is therefore also conceivable to imagine a deviation from a strictly enforced proportionality principle when patented inventions are common inputs into future technological progress.[213]

---

an inventor's embodied ideas in order to promote efficient allocation of research and development dollars).

211. Supply-oriented arguments are sometimes called "inducement" arguments—that is, arguments in which patent rights are calibrated so as to induce the expenditures required to supply inventions. The nonobviousness requirement is the most common locus of inducement requirements. *See* Graham v. John Deere Co., 383 U.S. 1, 9–10 (1966); Michael Abramowicz & John F. Duffy, *The Inducement Standard of Patentability*, 120 YALE L.J. 1590 (2011); Tun-Jen Chiang, *A Cost-Benefit Approach to Patent Obviousness*, 82 ST. JOHN'S L. REV. 39 (2008). Inducement arguments are not as common in relation to the doctrines that curtail the reach of patent scope into after-arising technology and improvement, but they are sometimes made. *See supra* notes 24–26 and accompanying text (discussing the quality-ladder model for patent protection for improvements); DAN L. BURK & MARK A. LEMLEY, THE PATENT CRISIS AND HOW THE COURTS CAN SOLVE IT 150–53 (2010) (arguing that broad protection in biotechnology is appropriate to provide sufficient incentives for risky, costly research).

212. *See* Yoram Barzal, *Optimal Timing of Innovations*, 50 REV. ECON. & STATS. 348 (1968).

213. The general concern is that patent rights on "the basic tools of scientific and technological work," Gottschalk v. Benson, 409 U.S. 63, 67 (1972), may do more to slow down post-invention progress than they did to speed up pre-invention progress. This general concern is an agglomeration of several distinct specific concerns. In part, the concern is about stealing from Peter to pay Paul. Today's patents on the inputs into tomorrow's research make tomorrow's research more expensive, a questionable move given that research may already be underproduced because of the positive externalities that it entails. *See* Brett M. Frischmann & Mark A. Lemley, *Spillovers*, 107 COLUM. L. REV. 257, 279 (2007) (discussing the "demand side" justification of intellectual property rights with spillovers). In part, the concern is about excessively broad claims that give a single firm excessive control over the direction of future research. *See* Merges & Nelson, *supra* note 2, at 908 (arguing in favor of competition rather than coordination in follow-on invention). Similar concerns have been voiced about the control that even a small, bottleneck claim that describes an essential input into many different avenues of research gives to a single patentee. *See* Eisenberg, *supra* note 109. In part, the concern is also about the anticommons or thicket problems that may result when an excessive number of fragmented rights on basic tools must be gathered together to perform research or develop innovative products. Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCIENCE 698 (1998); Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, 1 INNOVATION POL'Y & ECON. 119 (2000).

Finally, a market for embodied ideas is a means to the end of providing incentives to generate valuable technological ideas, but the patent regime may at times be focused on achieving other goals. In some contexts, the principal function of the patent regime may be to foster the efficient management of technological prospects,[214] the commercialization of already-invented technologies,[215] or the disclosure of already-discovered technical information.[216] Deviations from a market for embodied ideas and its proportionality principle may be required for the patent regime to achieve these alternative goals. Nonetheless, an important factor in patent protection, all other things being equal, is creating a market for embodied ideas. For such a market to exist, the private value of patent rights to patent owners should have a rough form of proportionality to the social value of an inventor's ideas, and, in turn, the set of things governed by a patent should track the set of things that embody an inventor's ideas.

B. THE IMPLICATIONS OF FOCUSING ON PROPERTIES AS THE LOCUS OF INVENTION

Assuming that patent claims should track the set of things that embody an inventor's ideas,[217] the important question in administering a patent regime is how the set of things that embody those ideas, and that therefore fall within the scope of a claim, should be identified. Yet, oddly and despite its centrality to the mechanics of a patent regime in which rights to exclude from things are designed to create a market for ideas, what it means for a thing to embody an idea has never been fully examined. One approach to answering this question is effectively to avoid answering it. Things could be assumed to embody ideas in some unspecified manner, and the set of things that embody ideas could be taken to be coextensive with the set of things disclosed and made available by the patent disclosure. Here, innovative things in their entireties are what embody innovative ideas, and things are the primitives of invention.[218] What Abby, Bernard, and Bob have all invented is simply the set of novel and nonobvious things that they disclosed and made available to the public at the time they filed their patents.[219]

---

214. *See* Kitch, *supra* note 27.

215. *See* Ted Sichelman, *Commercializing Patents*, 62 STAN. L. REV. 341 (2010).

216. *See* Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 480–81 (1974).

217. *See supra* Section VII.A.

218. *See supra* notes 62–67 and accompanying text (describing the paradigm in which things are the primitives of what an inventor invents).

219. *See supra* Part IV and Section V.A (presenting hypotheticals involving Abby, Bernard, and Bob).

Another approach to the question, however, is to answer it. Particular innovative properties of things can be taken to be the entities which instantiate innovative ideas in things. Properties can be framed as convenient intermediaries between the ontologically distinct realms of ideas (the entities for which incentives are sought) and things (the entities governed by patent rights) that patent law connects together as an instrumental matter.[220] It is because a thing possesses inter alia the property *being made of sliding plates that can capture mice* that the thing embodies Abby's inventive idea; it is because a thing possesses inter alia the property *being made of a wire that can catch mice and move in an arc in relation to a fixed base* that the thing embodies Bernard's inventive idea; it is because a thing possess inter alia the property *being made of a specific metal alloy* that the thing embodies Bob's innovative idea. Here, properties are the locus of invention.

In some situations, the difference between the thing- and property-centric approaches to invention is only a matter of semantics, and the coarse- or fine-grained nature of the analysis is irrelevant. For example, when a patentee sues someone who has made, used, or sold things that were actually disclosed and made available to the public by the patent at the time of filing, the scope of a patent will be the same regardless of whether things or properties are identified as the locus of invention. At the time a patent is filed, the set of innovative things that an inventor discloses is coextensive with the set of things that possess the inventor's innovative properties.

In improvement cases, however, the granularity at which the locus of invention is identified matters. In particular, it affects the ability of patent scope to track the set of things that embody an earlier inventor's ideas as later-developed improvements are discovered and produced. Even as coarser-grained things change over time, finer-grained properties can remain constant, allowing an earlier inventor's invention to be tracked through the ensuing later-developed improvements.[221] If innovative things are taken to be the primitives of what an inventor has invented, it is impossible to differentiate among a range of improvements and to say that some of the improvements embody an earlier inventor's ideas more than others do. The earlier inventor's invention is a set of things that the patent disclosed and made available to the public at the time of filing, and no improvements fall within that set because all improvements are new things that the earlier

220. *Cf. supra* notes 207–08 and accompanying text (describing the patent regime as a drive chain that links these distinct realms).

221. *See supra* Section II.C (arguing that a focus on properties permits a finer-grained analysis than a focus on things does).

inventor's specification did not disclose and make available to the public.[222] Precisely how, or in what way, they are different cannot be ascertained (or is irrelevant) because there is no "spirit" of an invention to track through the line of post-filing improvements. In contrast, if properties are viewed as the locus of invention, there is a way of tracking the persistence of an invention in post-filing improvements. Even if coarse-grained things change over time, certain finer-grained properties may persist in an unchanged state.[223]

Section VII.B.1 demonstrates that the properties invented by the earlier inventor persist in their entirety in overlooked improvements, meaning that earlier-filed patent claims should routinely encompass overlooked improvements and overlooked-improvement cases should be the "easy" cases that they are. Although an overlooked improvement is a new thing, the properties that instantiate the earlier inventor's ideas in the improvement are old and unchanged. In contrast, Section VII.B.2 demonstrates that the properties invented by the earlier inventor persist in classic improvements, but not in their entirety, meaning that classic-improvement cases should be the contested cases that they are today. Classic improvements continue to embody the earlier inventors' ideas in some ways, but not in others, making control of classic improvements a highly fact-specific inquiry. Section VII.B.2 also demonstrates that a property-centric framework enables the identification of the least-general naked property of a classic improvement and argues that this factor should be added to the conventional theory on patent protection for improvements in classic-improvement cases.

---

222. *See supra* Section III.A.2 (presenting the timing (or new-thing) criterion of an improvement).

223. If patents are intended to structure a market for ideas, *see supra* Section VII.A, and properties are the entities that instantiate ideas in things, *see supra* note 220 and accompanying text, the notion that patent interests should encompass at least the set of things that possess in full the properties invented by an inventor is practically a truism. However, at least for the purpose of constructing a patent regime, there is no metaphysically correct way to determine the level of granularity at which ideas are embodied in things. The open question is whether a focus on things or properties produces a patent regime that better serves the desired normative ends. *Cf. supra* notes 67–68 and accompanying text (framing the choice between things and properties as the locus of invention in terms of the rules-standards debate). Getting into the "spirit" of innovative things and focusing on properties as the locus of invention explains the otherwise inexplicable distinction between classic and overlooked improvements. As explored below in Sections VII.B.1 and VII.B.2, the proof of the value of a focus on properties as the locus of invention in improvement cases lies in its explanatory power and normative purchase. A conceptual framework in which properties are the locus of invention does what a thing-centric framework cannot. It both points out the distinction between classic and overlooked improvements and justifies the differential treatment that each type receives, with classic improvements giving rise to contested cases and overlooked improvements giving rise to "easy" cases.

1. Overlooked, "Easy" Improvements

To review, a focus on properties as the locus of invention reveals that an overlooked improvement still possesses all of the properties invented by the earlier inventor.[224] The improver's advance is embodied in a set of properties that is functionally independent from the set of properties that embodies the earlier advance,[225] so the addition of the improver's innovative properties to the improvement does not displace the earlier inventor's innovative properties from the improvement. More concretely, using the facts of the Abby-Bob hypothetical,[226] the properties invented by Abby (i.e., the properties addressing the mechanical configuration of the parts of a mousetrap) are fully present in Bob's improved mousetrap in which the spring is made out of an after-arising metal. They are present to the same extent that they ever were in the mousetraps that Abby disclosed in her own, earlier patent.

A focus on properties as the locus of invention also explains the "easy" nature of the overlooked-improvement cases. If properties are the entities that instantiate ideas in things,[227] then overlooked improvements still fully embody the earlier inventor's ideas because they possess all of the properties invented by the earlier inventor. In turn, if patent claims should track the set of things that embody an inventor's idea,[228] then the earlier inventors' patents should encompass the later-developed improvements. The conventional theory on improvements has only a minor role to play, if any, in determining the reach of earlier-filed patents into overlooked improvements[229] because the need for some rough form of proportionality between private patent rewards and the social value of innovative ideas simply trumps the factors addressed in the conventional theory.[230] Importantly, one cannot glean this

224. *See supra* Section V.A.

225. *See supra* Section V.B (discussing factual scenarios that give rise to functionally independent sets of properties).

226. *See supra* Section V.A.

227. *See supra* note 220 and accompanying text.

228. *See supra* Section VII.A.

229. *See supra* Section II.B (demonstrating that overlooked improvements lie in a blind spot of the conventional theory on patent protection for improvements).

230. At first glance, a non-infringement holding might seem simply to mean that an inventor cannot internalize any of the social welfare derived from a set of technologies that fully embody his technical advance. However, the distortion of the market for embodied ideas would be much more severe. Assuming that improvements are substitutes for the things the earlier inventor made available to the public in his disclosure, as they usually are, a non-infringement holding leads to erosion of the patentee's profits on the things that he did actually disclose and make available to the public in his specification. *See infra* notes 272–78 and accompanying text. A distinct, evidentiary reason why patent protection should perhaps

insight if things are the primitives of invention. Overlooked improvements, like all improvements, are new things that were not disclosed and made available to the public by the earlier inventor's patent. If things are the primitives of invention, then there is no way to track the continuity between an earlier-invented set of things and a later-invented set of things. Only by focusing on innovative properties is this continuity revealed. Although an overlooked improvement is a new thing, the properties that instantiate the earlier inventor's ideas in the improvement are old and unchanged. Particular finer-grained properties invented by an earlier inventor can remain constant over time even as the coarser-grained things change. The change in the things resides in properties of the things other than those properties that instantiate the earlier inventor's ideas.

2. Classic, Contested Improvements

Again to review, classic improvements are different from overlooked improvements when the focus is on properties as the locus of invention. Unlike in an overlooked-improvement case, there is no simple yes-or-no answer to the following question in a classic-improvement case: "Does the improvement still embody the earlier inventor's innovative ideas?" The answer is both yes and no. When parsed in terms of the properties attributable to the earlier inventor that persist in the improvement, classic improvements are a mixture that presents an uneasy middle ground. They continue to embody the earlier inventor's ideas in some ways, but not in others. The properties that instantiate the earlier inventor's more general advances are still present in the improvement, but the properties that instantiate the earlier inventor's more specific advances have been supplanted from the improvement by the properties that instantiate the improver's advance.[231] More concretely, using the facts of the Abby-Bernard hypothetical,[232] Bernard's improved mousetrap still embodies Abby's general idea because its possesses a property like the property *being a device in which the jostling motion of a mouse transforms the potential energy stored in a spring into the kinetic*

more readily encompass overlooked improvements is that overlooked improvements are unlikely to be the result of independent invention. The after-arising advances that give rise to overlooked improvements are usually in arts that are different from the art of the earlier patent. *See supra* text accompanying notes 154–57. It is therefore extremely unlikely that the later inventor independently invented the earlier invention. Bob, an expert in metallurgy, is unlikely to have thought up Abby's mousetrap design on his own. Although patent law does not provide an independent-invention defense as a doctrinal matter, *see supra* notes 104–05 and accompanying text, the increased likelihood of copying strengthens the earlier patentee's normative claim to overlooked improvements.

231. *See supra* Part IV.

232. *See id.*

energy required to catch a mouse. However, Bernard's improved mousetrap no longer embodies Abby's more specific idea because it does not possess the property *being made of sliding plates*. Abby's more specific property has been supplanted by Bernard's innovative property *being made of a wire that can move in an arc in relation to a fixed base*. Whether an earlier inventor's innovative property persists in the improvement depends on which innovative property is the focus of attention.

Given this partial persistence of the earlier inventor's properties, a focus on properties as the locus of invention explains the contested nature of classic-improvement cases under the conventional theory of patent protection for improvements.[233] If properties are the entities that instantiate ideas in things,[234] then classic improvements still embody some of the earlier inventor's ideas, but not all of them. In turn, if patent claims should track the set of things that embody an inventor's ideas,[235] then the reach of the earlier inventors' patents into classic improvements is legitimately contested.[236] The need to include or exclude the improvement from the earlier patentee's rights in order to structure a market for embodied ideas is not such an overriding concern that it must trump other normative considerations. In this borderline situation, the outward context is important and, hence, the conventional theory that addresses the context in which an improvement comes into being becomes relevant.[237] The earlier patentee's case for infringement is weaker when the classic improvement is very important in relation to the earlier-patented technology,[238] when the pattern of technical

233. *See supra* notes 29–37 and accompanying text (demonstrating that the conventional theory on patent protection for improvements has relevance in the classic-improvement cases).

234. *See supra* note 220 and accompanying text.

235. *See supra* Section VII.A.

236. Overlooked improvements are "easy" because an improvement that possesses *all* of the properties invented by an earlier patentee should infringe the earlier patent, despite the fact that the improvement is a new thing that was not made available to the public by the earlier patent. *See supra* Section VII.B.1. Classic improvements, however, raise a different threshold question: Should the presence of *any* property invented by an earlier patentee in a later-developed improvement, regardless of the generality of that property, be a sufficient condition for infringement? If one were to answer this question in the affirmative, then all classic improvements, too, would be "easy" cases as improvements by definition still embody at least one property that instantiates an earlier inventor's ideas. However, there are good reasons to be skeptical of such an expansive approach to patent protection for improvements that furthers a goal of perfect internalization. *See supra* note 202.

237. *See supra* notes 76–79 and accompanying text (noting that the conventional theory on improvements can only consider the outward context in which a later invention occurs, not the intrinsic relationship between the earlier- and later-invented things).

238. *See supra* notes 18–21 and accompanying text.

advance in the industry is cumulative and the market for patent licenses is full of friction,[239] when the need for incentives in the industry is low,[240] and when the prospect function of patents plays only a minor role in the justification of patent rights.[241]

A focus on properties as the locus of invention does more than simply highlight the importance of the conventional theory in classic-improvement cases. By examining improvement at a finer-grained level, it also reveals a factor that should be critical in determining the optimal reach of an earlier-filed patent into classic improvements and that should be added to the conventional theory on patent protection for improvements. This additional factor goes unnoticed when things are taken to be the primitives of what an inventor invents because it implicates the nature of the earlier inventor's property that persists in the classic improvement. In particular, it implicates the level of generality of the property that persists in the improvement. In a classic improvement, the most specific of the earlier inventor's innovative properties is uniformly supplanted by the later inventor's innovative property. However, the most specific level of generality at which the earlier inventor's properties persist in the improvement will vary from case to case. The generality of this property—call it the earlier inventor's least-general naked property[242]—is a critical factor to consider in determining whether earlier-filed patents should encompass later-developed classic improvements.

To understand the concept of the least-general naked property in a classic improvement, consider a third hypothetical, once again involving Abby as the earlier inventor in an improvement scenario.[243] Abby is still the first inventor of a spring-loaded mousetrap, and she discloses and makes available to the public a sliding-plate mousetrap. This time, however, Barry creates an improvement. Barry makes a mousetrap in which sheets of wire mesh, rather than plates, slide with respect to each other. The wire mesh is newly engineered for the improved mousetrap. The openings in the wire mesh are large enough for a mouse to pass through. When the trap is triggered and the sheets of wire mesh slide with respect to each other, they catch a mouse, just like the sliding plates with holes did in Abby's mousetrap. However, the wire mesh is less expensive to produce, and it also works better

239. *See supra* notes 22–23 and accompanying text.

240. *See supra* notes 24–26 and accompanying text.

241. *See supra* notes 27–28 and accompanying text.

242. This property is "naked" in the classic improvement because all of the earlier inventor's less general, i.e., more specific, properties have been stripped away.

243. *See supra* Part IV (describing Abby's earlier invention).

because some "give" in the wire mesh ensnares the mouse more snuggly and securely than sliding plates with holes do.

Just like Bernard's improvement, Barry's improvement is a classic improvement. His later-developed innovative property is something like *being made of sliding wire meshes with "give."* In the improvement, Barry's innovative properties supplant the most specific of Abby's innovative properties, namely the property *being made of sliding plates with holes*. However, Abby's more general properties persist in the improvement. For example, Abby's more general innovative property *being made of sliding planar elements* persists, as both a mesh and a plate are planar elements.

What factual differences between Bernard's classic improvement and Barry's classic improvement are relevant in assessing whether Abby has a stronger normative claim to patent rights that encompasses either one or the other?[244] One difference is the level of generality of Abby's least-general naked property. In the Abby-Barry hypothetical, Abby's least-general naked property is something like *being made of sliding planar elements*. In the Abby-Bernard hypothetical, it is something like property *being a device in which the jostling motion of a mouse transforms the potential energy stored in a spring into the kinetic energy required to catch a mouse*. In a relative sense, the least-general naked property is much more specific in the Abby-Barry hypothetical. The set of mousetraps that possess this least-general naked property is a subset of the set of mousetraps that possess the least-general naked property in the Abby-Bernard hypothetical.

A coarse-grained focus on things as the primitives of invention does not allow the least-general naked property of a classic improvement to be identified, but a finer-grained focus on properties as the locus of invention does. Furthermore, allowing an earlier inventor's patent claim to encompass a classic improvement merits more skepticism when the earlier inventor's least-general naked property is a more general property. The higher the level of generality of the least-general naked property of a classic improvement, the weaker the earlier patentee's normative claim to patent rights that encompass the improvement. On an intuitive level, this proposition makes sense because it formalizes the notion that some improvers, e.g., Barry,

244. For the sake of convenience, assume that the conventional theory on patent protection for improvements, *see supra* Section II.A, does not differentiate between Bernard's and Barry's inventions. Assume that the relative importance of the two inventions is the same. The mousetrap industry is the same in both, so the same pattern of technical advance and concerns about friction in the market for patent licenses are present.

borrow more of their ideas from earlier inventors than others, e.g., Bernard, borrow.[245]

As an economic matter, this proposition allows the scholarship on improvements to tap into a well-established body of scholarship on the optimal scope of a valid claim. Through an array of doctrines, patent law curtails the level of generality at which patentees can draft claims to protect their inventions.[246] Thus, the inventor of the first cure for the common cold cannot claim generally "a drug capable of curing the common cold" but is instead limited to a more specific claim that is in some way limited by the structure of the molecule that the inventor has created. The costs of claims drafted in the most general or abstract of language are well documented, and it is widely believed that these costs outweigh whatever benefits (in terms of incentives to invent or coordinate) such claims would entail.[247] Paying attention to the generality of the earlier inventor's least-general naked property in a classic improvement allows the body of scholarship and case law on the permissible generality of claim scope to be brought to bear on improvements. The more general the least-general naked property in a classic improvement, the more general the language in the claim that the patentee would need to be granted in order to have rights that encompass the improvement.[248] If properties are identified as the locus of invention and attention is paid to the earlier inventor's least-general naked property in a classic-improvement case, the normative concerns about excessive general claims and earlier inventors' rights to classic improvements merge into the same concern.

Of course, this merger is not a silver bullet for courts grappling with classic-improvement cases. The highest level of generality at which claim scope should be sanctioned is notoriously difficult to identify, and identifying the least-general naked property that gives an earlier inventor rights that encompass an improvement will therefore be equally difficult to identify. The points to be made here are only that the theory on patent protection for improvements has to date failed to incorporate the levels-of-generality debate

245. Importantly, this intuition cannot be articulated with precision if things are taken to be the primitives of what an inventor invents.

246. *See supra* notes 130–32 and accompanying text.

247. For a recent articulation of this argument that expressly builds on Judge Hand's levels-of-generality argument in copyright law, see Chiang, *supra* note 13.

248. The connection between the least-general naked property of a classic improvement that can be controlled by an earlier patentee, on the one hand, and the permissible level of generality at which a peripheral claim can be drawn, on the other hand, demonstrates the strong conceptual connection between peripheral claims and properties of things. *See infra* notes 315–17 and accompanying text.

and that this oversight can, and should be, corrected by recognizing the properties of things, rather than things in their entireties, as the locus of invention in improvement cases.

C. REFRAMING COMPLEMENTS AND SUBSTITUTES IN PATENT LAW

The previous Section demonstrated the explanatory value and normative purchase of paying attention on a fine-grained level to innovative properties, rather than only on a coarse-grained level to innovative things as indivisible wholes, when fine-tuning the reach of patents into improvements. This Section reprises this argument, bringing new conceptual tools to bear on the problem. It illustrates that the economic concepts of complements and substitutes can be used to explain why courts do and should treat overlooked improvements as "easy" cases and classic improvements as contested cases. This move is both novel and important because, for the first time, it internalizes the analysis of these well-known economic concepts within patent law proper, i.e., the patent law that determines claim scope and validity.[249]

This Section proceeds in three steps. Section VII.C.1 introduces the economic concepts of complements and substitutes. Section VII.C.2 reviews the ways in which complements and substitutes are employed today in scholarship related to patent law. Importantly, these uses of the concepts do not allow courts to identify a distinction between classic and overlooked improvements or craft claim scope. Section VII.C.3 demonstrates how complements and substitutes can be used to achieve these ends if, and only if, properties are viewed as the locus of invention and are labeled as the goods at issue.

1. Complements and Substitutes

In the everyday sense of the word, "substitutes" are goods that can replace or fill in for each other because they satisfy the same consumer need.[250] Nails and industrial strength glue for bonding wood are substitute goods in this common-sense way: I use either one, but probably not both, to join pieces of wood. In contrast, the everyday meaning of "complements" is a set of goods that are two parts of a whole and that consumers tend to consume together because they desire the whole.[251] Hammers and nails are

249. I have previously offered a rough sketch of this argument in the context of the enablement doctrine. Collins, *supra* note 58, at 1111–24.

250. THE AMERICAN HERITAGE COLLEGE DICTIONARY, *supra* note 62, at 1354 (defining a substitute as "one that takes the place of another; a replacement").

251. *Id.* at 284 (defining a complement as "something that completes, makes up a whole, or brings to perfection").

complements: I practically need a hammer to use a nail for its most common purpose and vice versa.

To determine whether goods are complements or substitutes as a technical, economic matter, economists measure the goods' cross-price elasticity of demand.[252] Two goods are substitutes if a decrease in the price of one good results in a decrease in demand of the other good and, inversely, an increase in the price of one good results in an increase in demand for the other good.[253] This technical definition usually maps onto the common-sense definition of a substitute. If consumers are willing to use either one good or the other to fulfill their needs, then a decrease in the price of one will drive consumers toward that good and away from the other. The cheaper nails are, the less likely I am to buy an industrial strength glue when either one or the other can be used to achieve the desired goal of attaching pieces of wood.

In contrast, two goods are complements as a technical matter if a decrease in the price of one good results in an increase in the demand for the other good and, inversely, an increase in the price of one good results in a decrease in the demand for the other good.[254] Again, there is a link between the common-sense and technical definitions of a complement. Consumers tend to consume complementary goods together because the consumer's willingness to pay for the combination of the two goods is more than the sum of the consumer's willingness to pay for the two goods individually.[255] In other words, there is a "synergy" between the two goods."[256] Because consumers place greater value on consuming the complementary goods together, the price that drives consumer-purchasing decisions is in part the price of the bundle of goods. A decrease in the price of one good in the bundle decreases the price of the bundle as a whole, meaning a consumer will tend to consume more of the bundle and thus more of the other good. All things being equal, the cheaper hammers are, the more nails I will consume.[257]

252. ROBERT S. PINDYCK & DANIEL L. RUBINFELD, MICROECONOMICS 36 (7th ed. 2008) ("A *cross-price elasticity of demand* refers to the percentage change in the quantity demanded for a good that results from a 1 percent increase in the price of another good.").

253. *Id.*

254. *Id.* at 24–25.

255. SCOTCHMER, *supra* note 24, at 144; Shapiro, *supra* note 201, at 122–23.

256. Shapiro, *supra* note 201, at 122.

257. Complements and substitutes span a spectrum from being perfect complements and substitutes to not being complements or substitutes at all. Perfect complements and substitutes exist if the goods are consumed together or substituted for each other, respectively, at a one-to-one ratio. PINDYCK & RUBINFELD, *supra* note 252, at 76–77.

2. The Three Existing Frames for Identifying Complements and Substitutes

In contemporary scholarship on improvements, the concepts of complements and substitutes have been used in three different manners. One looks at the boost that earlier innovators give to later innovators in the process of cumulative innovation and labels the successive innovations as technological complements. A second looks at the things produced by earlier and later inventors and queries whether they are complements or substitutes in the eyes of consumers. A third identifies patent rights as either complements or substitutes. None of these approaches to identifying complements and substitutes, however, offers a conceptual framework that can distinguish classic and overlooked improvements.

a) Cumulative Innovation and Complements

When innovation is cumulative and later innovations build on earlier ones, the leg up given to the later innovation can be a significant part of the social value created by the earlier innovation.[258] The leg up can come in many forms. The earlier innovation may be necessary to develop the later innovation, it may reduce the cost of achieving the later innovation, or it may speed up the later innovation.[259] Regardless of the form that the boost takes, the process of cumulative innovation can be characterized as the successive discovery of complementary innovations.[260]

Although all cumulative innovation involves the successive invention of complements from this process-oriented viewpoint, there is no single, unified implication of this economic fact for patent protection for improvements. Earlier innovators can give boosts to later innovators through a diverse array of mechanisms, and these different mechanisms result in different allocations

258. SCOTCHMER, *supra* note 24, at 127; Scotchmer, *supra* note 2, at 31. The notion of an earlier innovation giving a post to a later innovation is related to the process-oriented definition of an improvement that focuses on intergenerational facilitation. *See supra* notes 102–06 and accompanying text. James Bessen and Eric Maskin have defined this type of innovation as "sequential" innovation. Bessen & Maskin, *supra* note 2, at 612 (defining sequential innovation as a process in which "each successive invention builds on the preceding one"). They also use the term "complementary" as a term of art and in a manner that differs from its use in this Article. *Id.* at 612 (defining complementary innovative paths as non-redundant innovative paths).

259. SCOTCHMER, *supra* note 24, at 127; Scotchmer, *supra* note 2, at 31. The fraction of the earlier innovation's value that lies in the boost can also vary. *See* Green & Scotchmer, *supra* note 2, at 22 (discussing an earlier innovation that has no direct value to consumers).

260. Shapiro, *supra* note 201, at 124–25.

of rights among the earlier and later inventors.[261] Many of these mechanisms through which cumulative innovation occurs do not involve improvements at all, at least as the term is used in this Article. In research tool cases, neither the products nor the knowledge generated by the later innovator is likely to fall within the scope of the earlier innovator's claim, but the later innovator must compensate the earlier patentee for the use of the very research tool disclosed in the patent.[262] In cases in which the earlier-patented invention is improved by the later creation of a new component, the very things disclosed in the earlier patent must still be made, sold, and used as part of the "improved" technology.[263] In idea-only cumulative innovation cases, the scales tip in the other direction. The earlier innovator's rights do not reward him for the boost that he gives to later innovators, as the knowledge disclosed in a patent is free for all to use qua knowledge.[264] When the mechanism through which cumulative innovation occurs is improvement, the presumption is that the earlier patent disclosure gives a boost to the improver to develop the improvement.[265] The improvement possesses some property that was invented by the earlier patentee,[266] and the improver was able to use the knowledge of this earlier-invented property disclosed in the specification as a platform upon which to pursue further technological progress.[267] All improvement cases implicate these intergenerational boosts; all improvers have benefited from the value created by earlier patentees.

261. The root cause of the different mechanisms yielding different allocations is the fact that patents treat the sets of things an inventor invents, not the ideas per se he generates, as property (or something akin thereto). *See supra* notes 205–08 and accompanying text.

262. *See supra* note 115 and accompanying text. In addition to the fact that research tools and the innovations that they facilitate are complements, research tools are often implicated in another, distinct complementary relationship. Two (or more) research tools of the same generation that are both inputs into the next generation of innovation are complements. SCOTCHMER, *supra* note 24, at 144. The need to acquire licenses to a large number of complements to achieve the next generation innovation can lead to licensing difficulties. *See supra* note 213.

263. *See supra* notes 111–12 and accompanying text. The inclusion of a large number of complementary components in a single good can lead to a royalty stacking problem, *see supra* note 116, and raise Cournot-complement problems, *see infra* notes 279–80 and accompanying text.

264. *See supra* notes 117–18 and accompanying text.

265. Because independent invention is not a defense to patent infringement, the boost from the earlier patentee to the later infringer should be understood as a constructive boost: if the later improver had known of and read the earlier patentee's disclosure, the improver would have gotten a boost. *See supra* note 106.

266. *See supra* Section III.A.3.

267. Unlike in a case involving a research tool or a later-developed component, the improver never needs to use a thing that was actually disclosed and made available to the public in the earlier patent.

Therefore, although the notion that cumulative innovation involves successively complementary innovations teaches us something about improvements broadly writ, it does not help to distinguish classic and overlooked improvements.

b) Successively Invented Things as Complements or Substitutes

Rather than looking to the process through which cumulative innovation occurs and identifying earlier boosting innovations and later boosted innovations as complements, it is also possible to focus on the things produced by earlier and later innovators and to label them as either complements or substitutes. Things that are either complements or substitutes are often invented in succession. Because patent rewards are market-based and filtered through willingness to pay, the successive invention of complement and substitute things has important ramifications for the private value of patents over time. Even if the set of things encompassed within a patent remains rigidly fixed and the later-developed thing falls outside of the scope of the earlier-filed patent, a later-developed substitute decreases an earlier patentee's profits, whereas a later-developed complement may increase an earlier patentee's profits.[268]

When the later-developed thing is a complement for the earlier-patented thing, the private value of the earlier patent can increase when the later-developed thing is marketed.[269] For example, assume that an earlier inventor invents and patents the hammer and that a later inventor invents an improved nail that does not bend as easily. The later invention of the improved nail increases the utility of hammers to consumers. For many consumers, the value of a hammer is determined by the fact that the hammer plays a role in the process of pounding nails into wood. When improved nails become available, the value of a hammer increases as it now plays a role in a process that is more valuable to consumers because the pounding of nails into wood is easier to accomplish and involves fewer bruised thumbs and less waste (in the form of fewer bent nails). A shift in the availability of technology beyond the scope of a patent makes the patented technology

268. For simplicity, the following discussion assumes that a patentee is initially able to exercise some monopoly power or enjoy some supra-competitive profits. *But cf.* Kenneth W. Dam, *The Economic Underpinnings of Patent Law*, 23 J. LEGAL STUD. 247, 249–51, 268–70 (1994) (noting that a patent does not allow its owner to exercise monopoly power or charge supra-competitive prices if consumers are indifferent between patented and unpatented technologies).

269. *See* Shapiro, *supra* note 201, at 122–25 (defining the concept of technical complementarity as the social value of the combination being greater than the social values of the two inventions considered separately).

more valuable to consumers.[270] In a regime in which rewards are determined by a market for embodied ideas and willingness to pay,[271] this increase is an expected change in the value of patent rights over time.

In contrast, when the later-developed thing is a substitute for the thing disclosed and claimed by the earlier patent, the private value of the earlier patent is diminished when the later-invented thing is marketed.[272] The profit that a patentee can realize from a patent cannot be determined simply by understanding intrinsic value or utility of the technology that falls within the scope of the patent to a consumer. For example, assume that an earlier inventor patented the nail, that a later inventor invents an industrial-strength glue, and that, at least for some purposes, consumers are indifferent as to which product they prefer. Also assume that the owner of the nail patent enjoyed some monopoly power before the invention of the industrial-strength glue. When the industrial strength glue reaches the market, the value of nails to consumers does not change. Nails are still capable of doing what they did before the invention of the glue. However, because nails and the glue are substitutes, the owner of the nail must compete on the basis of price with the glue producers, so the later-developed substitute decreases the private value of patent rights on the earlier-invented thing.[273] Again, in a regime in which rewards are determined by a market for embodied ideas and

270. The portion of this increase in social welfare that occurs upon the later development of a complement that an earlier patent owner can internalize depends, in part, on whether the later-developed complement is patented. *See* Michael Kremer, *Patent Buyouts: A Mechanism for Encouraging Innovation*, 113 Q.J. ECON. 1137, 1156 (1998) (noting that the value of a patent is higher if a non-infringing complementary good is not patented and is instead in the public domain).

271. *See supra* Section VII.A.

272. PINDYCK & RUBINFELD, *supra* note 252, at 434–35 (noting that firms with some monopoly power face more elastic demand curves, and earn smaller profits, when there are more and closer substitutes); SCOTCHMER, *supra* note 24, at 103–07 ("The demand curve [for a patented technology] will be more elastic, and generally lower, if close [non-infringing] substitutes are allowed in the market."). It is for this reason that some economists model the "leading breadth" of a claim that sets the reach of a patent right into improvements as a measure of the effective term of a patent. *See supra* notes 24–26 and accompanying text (discussing a "quality ladder" model of improvement). The profit-reducing effect of the entry into the market of substitute, non-infringing goods is a core tenet of models of monopolistic competition. *See* Christopher S. Yoo, *Copyright and Product Differentiation*, 79 N.Y.U. L. REV. 212, 238–39 (2004) (describing the effect of the entry of a new, substitute work into the market as a "backwards" shift in the demand curve).

273. How robust that competition will be is dependent upon whether the glue, too, is patented. *See* Kremer, *supra* note 270, at 1154 (noting that a patent on Prozac is worth less if the patent on Zoloft is made available to all comers at no price).

willingness to pay,[274] this decrease is an expected change in the value of patent rights over time.[275]

Improvements of all types are usually substitutes for the things that were disclosed and made available to the public at the time of the filing of the earlier patents.[276] More importantly, both classic and overlooked improvements are usually substitutes in this manner. Both Curtiss's improved airplane (a classic improvement) and the improved-canvas airplane (an

274. *See supra* Section VII.A.

275. Because later innovators can obtain rents by designing around an earlier patent and transforming a monopoly (presuming it exists) into a duopoly, patent law creates an incentive to design around existing patents, whether through the creation of improvements or the exploitation of gaps in poorly drafted claims. Whether this incentive to design around is socially beneficial is the subject of a robust debate. Courts generally view the incentive as beneficial. *See* State Indus. Inc. v. A.O. Smith Corp., 751 F.2d 1226, 1236 (Fed. Cir. 1985). The scholarly opinion is mixed. *Compare, e.g.*, Bessen & Maskin, *supra* note 2, at 613 ("[A]n important role of patents is to encourage innovative activity on the part of others who would otherwise be inclined merely to imitate."), *with* Michael Abramowicz, *The Uneasy Case for Patent Races over Auctions*, 60 STAN. L. REV. 803, 827 (2007) (emphasizing the duplicative, wasteful efforts entailed in design around).

276. *See* Green & Scotchmer, *supra* note 2, at 20 ("[C]ompetition from improved products could undermine the original innovator's profit."); O'Donoghue et al., *supra* note 2, at 2 (noting that design-around shortens the effective, but not legal, term of a patent because improvements are substitutes for the earlier-patented things). Some scholarship on cumulative innovation assumes that improvements increase the value of earlier-patented things and therefore increase an earlier patentee's profits, but this scholarship seems to use a broader definition of improvement than this Article does. *See supra* Section III.A (defining an improvement). For example, a second-stage innovation that produces a new application of a first-stage innovation that serves a market unrelated to the market served by the first-stage innovation is not an after-arising substitute. *See* Green & Scotchmer, *supra* note 2, at 20. This second stage innovation seems to presume a later-discovered use for an earlier-patented thing, and this type of cumulative innovation does not entail improvement. *See supra* note 113 and accompanying text. Similarly, a second-stage innovation can be thought of as an improvement that enhances, but does not compete with, the earlier innovation. Bessen & Maskin, *supra* note 2, at 620. The facts that most closely fit this model seem to be the discovery of an after-arising complement that has no use except when used together with the earlier-patented and earlier-disclosed things. *See supra* notes 111–12 and accompanying text. Nonetheless, even following the narrow definition of an improvement that structures this Article, not all improvements will be substitutes for the things made available by earlier patents. For example, consider an earlier patent on a slow-release technology for pharmaceuticals. *See supra* text accompanying note 74. If the earlier patent claim describes a "pill" or "drug," a pill with an after-arising active ingredient is an improvement: the after-arising pill is a thing; the earlier patent did not disclose or make available the after-arising pill to the public; the after-arising pill still possesses a property that was invented by the earlier inventor. *See supra* Section III.A. Yet, the after-arising drug in its slow-release formulation may be a substitute for some slow-release drugs that were disclosed by the patent, a complement for others, and neither a complement nor a substitute for yet others. For simplicity, this Section addresses only later improvements that are substitutes for the things disclosed by the earlier patent.

overlooked improvement) are substitutes for the airplanes that the Wright Brothers taught the public how to make.[277] Both Bernard's mousetrap (a classic improvement) and Bob's mousetrap (an overlooked improvement) are substitutes for the mousetraps disclosed by Abby's patent at the time it was filed.[278] Thus, looking at successively invented things as complements or substitutes does not allow any distinction between classic and overlooked improvements to be drawn.

c) Patent Rights as Complements or Substitutes

Finally, in the analysis of the pro- or anticompetitive effects of a patent cross-licensing or pooling agreement, legal doctrine and economic scholarship identify patent rights as either complements or substitutes.[279] If the patent rights are substitutes, then an exclusive agreement to cross-license the patents, or to pool the patent rights together, may in effect be an anticompetitive agreement between natural competitors not to compete on price. Consumers want to use either one patented technology or the other, so taking away the competition between the patent owners can be problematic. However, if the patent rights are complements, patent licenses may well be procompetitive because they eliminate Cournot-complement or thicket problems.[280] Consumers want to use both sets of patented technologies together, so one-stop shopping at the least reduces the transaction costs of patent licensing.

Courts identify patent rights as either complements or substitutes through two distinct layers of analysis. If the two patents overlap in the sense that they both describe the self-same, indivisible thing, the rights are complements.[281] To practice the technology that is in the overlap, both rights

277. *See supra* text accompanying notes 29–32, 40 (discussing these inventions).

278. *See supra* Part IV and Section V.A (presenting and analyzing the hypothetical inventions of Abby, Bernard, and Bob).

279. *See* 2 HERBERT HOVENKAMP ET AL., IP AND ANTITRUST: AN ANALYSIS OF ANTITRUST PRINCIPLES APPLIED TO INTELLECTUAL PROPERTY LAW § 34.2c, at 34-6 to -10 (2d ed. 2004); U.S. DEP'T OF JUSTICE (DOJ) & FED. TRADE COMM'N (FTC), ANTITRUST GUIDELINES FOR THE LICENSING OF INTELLECTUAL PROPERTY § 5.5 (1995), *available at* www.usdoj.gov/atr/public/guidelines/0558.htm (discussing the anticompetitive and procompetitive effects of cross-licensing and patent pooling arrangements); Shapiro, *supra* note 213, at 119.

280. DOJ & FTC, *supra* note 279, § 5.5 (noting that cross-licensing and pooling arrangements "may provide procompetitive benefits by integrating complementary technologies, reducing transaction costs, clearing blocking positions, and avoiding costly infringement litigation"). On Cournot complements, see SCOTCHMER, *supra* note 24, at 144–46; Shapiro, *supra* note 213, at 122–26.

281. Two patents that both describe the self-same, indivisible thing are overlapping blocking patents. *See infra* note 297 (defining overlapping and economic blocking patents).

must be obtained. If the patents do not overlap in this manner, then the status of the patent rights as complements or substitutes turns on whether the distinct things described by the patents are complements or substitutes. For example, if the two patents claim hammers and nails, respectively, then the patent rights are complements.[282] However, if the two patents claim a nail and an industrial-strength glue, then the patent rights are substitutes.

The antitrust analysis of patent rights as complements or substitutes offers no insight into the optimal depth of patent rights when improvements are the contested margin. Whether the patents owned by the earlier and later inventors are complements or substitutes hinges on a court's decision to allow the earlier-issued patent to encompass the improvement. If the earlier patent does not encompass the improvement, the patent rights are substitutes. The patents do not overlap to describe the self-same, indivisible thing, and the distinct things claimed by each patent are substitutes. However, if the earlier patent does encompass the improvement, then the patent rights are complements because the earlier and later patents do overlap and describe the self-same, indivisible thing.[283] In sum, trying to use the antitrust analysis of complementary and substitute patents to craft claim scope requires putting the cart before the horse. A court's decision on how to craft claim scope is a necessary input. It must be taken as a given to avoid circular reasoning.

3. Successively Invented Properties as Complements and Substitutes

The three frames in contemporary scholarship for identifying complements and substitutes in sequential innovation neither help courts to distinguish classic and overlooked improvements nor provide insight into the optimal reach of a patent on an earlier-developed technology into later-developed improvements.[284] This Article has developed a fourth frame: the properties of things invented by successive innovators can be identified as complements, substitutes, or a mixture of the two. When properties are the relevant goods, a systematic distinction between contested classic

282. Although two patents describing distinct sets of complementary things do not overlap, they are nonetheless blocking patents of a sort. *See infra* note 297 (defining overlapping and economic blocking patents).

283. Improvement scenarios are commonly described as giving rise to complementary patents, despite the fact that improved things are often substitutes for the earlier-developed and earlier-patented things. *See* John F. Duffy, *The Marginal Cost Controversy in Intellectual Property*, 71 U. CHI. L. REV. 37, 48–49 (2004); Kremer, *supra* note 270, at 1156–57. This characterization is only correct if the scope of the earlier patent encompasses the later-developed improvement.

284. *See supra* Section VII.C.2.

improvements and "easy" overlooked improvements can be discerned, and the concepts of complements and substitutes can be operationalized in patent doctrine so as to help courts craft optimal claim scope.

The properties of things can be identified as either complements or substitutes, just like things can. The standard intuitive and economic definitions apply.[285] Two properties are complements if consumers desire both of them together in a single thing.[286] The properties of *containing pharmaceutical Z* and *being enteric coated* are complementary properties if consumers want pills that both contain chemical Z and have an enteric coating. When consumers desire the two properties bundled together in a single thing, an increase in the price of the right to use one property leads to a higher price for the bundle and a decrease in demand for the other property.[287] In contrast, two properties are substitutes if consumers desire either a thing with one property or a thing with the other property, but not a thing with both. The properties *containing pharmaceutical compound X* and *containing pharmaceutical compound Y* are substitute properties if consumers desire either compound X or compound Y to combat an illness, but not both at the same time or in combination. When consumers desire either a thing with one property or a thing with the other, an increase in the price of the right to use one property leads to an increase in the demand for the other property.

Once established, the concepts of complementary and substitute properties provide a new way of describing the distinction between overlooked and classic improvements. The very concepts of properties that compound with and supplant each other that were employed to differentiate classic and overlooked improvements translate directly into the concepts of properties that are complements and substitutes, respectively. In an

285. *See supra* Section VII.C.1 (offering both intuitive and economic definitions of complements and substitutes).

286. Technically, properties can also be complements if consumers desire both of them together and each property is embodied in a distinct thing. However, the issue of improvements centers on complementary properties of a single, indivisible thing.

287. In order to talk about properties as complementary and substitute goods, it is necessary to imagine a hypothetical patent regime in which patent rights encompass all things that possess the property. Complements and substitutes are identified by the cross-price elasticity of demand of two goods—an idea that is hard to interpret if one does not have to pay to use the goods. This hypothetical patent regime that enables discussions of properties as complements or substitutes is different from the actual patent regime in that the scope of actual patent rights does not track the presence of all properties infinitely far into after-arising technology. *See supra* notes 242–48 and accompanying text (noting that the persistence of properties defined at high levels of generality does not necessarily mean that an earlier-filed patent should encompass a later-developed improvement).

overlooked improvement, the properties that instantiate the advances of the successive inventors are pure complements. The successively invented properties simply compound with each other, making the improvement a single, indivisible thing that naturally bundles together the properties that instantiate the earlier and later advances.[288] Consumer demand for the improvement is demand for the bundle of the earlier- and later-developed properties, and it is sensitive to the total cost of the rights to use both properties. An increase in the price of the right to use one property increases the price of the bundle as a whole, and it thereby decreases the consumption of the other property. The later invention of a metal from which one can make longer-lasting springs (Bob's invention) makes the invention of a mechanical design for a spring-loaded mousetrap (Abby's invention) more valuable to society.[289] In other words, Bob's property *being made of an after-arising alloy* complements all of Abby's innovative properties that pertain to the mechanical arrangements of the components. In contrast, in the process of refinement that gives rise to a classic improvement, the properties that instantiate the advances of the successive inventors are a mixture. They are part complement and part substitute.[290] The properties that instantiate the later advance in the improvement compound with, and thus complement, the properties that instantiate the earlier inventor's more general properties. The later invention of a better way to use a spring in a spring-loaded mousetrap (Bernard's invention) makes the earlier invention of the notion of a spring-loaded mousetrap itself (Abby's most general invention) more valuable to society. In other words, the specific property *being made of a wire that can move in an arc in relation to a fixed base* complements the property *being a device in which the jostling motion of a mouse transforms the potential energy stored in a spring into the kinetic energy required to catch a mouse*. These same later-developed properties supplant, and thus substitute for, the properties that instantiate the earlier inventor's more specific properties. The later invention of a better way to use a spring in a spring-loaded mousetrap (Bernard's invention) can swap in for the earlier invention of the notion of a sliding-plate mousetrap (Abby's more specific invention). In other words, the specific property *being made of a wire*

288. *See supra* Part IV. This bundling is visually apparent in Figure 4, *supra*. In an overlooked improvement, the later-invented property supplants, and thus is a substitute for, a property of the earlier-patented things that is not one of the earlier patentee's innovative properties.

289. If things rather than properties are the primitives of invention, the overlooked improvement cannot be viewed as a bundle of the two inventions. Successively invented properties can be complements even though the things are actually produced by subsequent inventors.

290. *See supra* Section V.A. The mixture is visually apparent in Figure 3.

that can move in an arc in relation to a fixed base is a substitute for the property *being made of sliding plates*. In sum, the concepts of properties that compound with and supplant each other that were employed to differentiate classic and overlooked improvements translate directly into the notions of properties that are complements and substitutes, respectively.[291]

The true payoff of being able to identify successively invented complementary and substitute properties is a retooled normative explanation of why claim scope should be sculpted (as it is today, albeit without any recognition of this fact) so that overlooked improvements give rise to "easy" cases and classic improvements give rise to contested cases. There is no economically meaningful distinction between the earlier and later inventors of complementary distinct things such as a hammer and an improved nail, on the one hand, and the earlier and later inventors in an overlooked-improvement case, on the other. The former may have invented distinct complementary things and the latter complementary properties that are possessed by a single, indivisible thing, but this is not a difference that should matter. The later development of a complementary property for an earlier-invented property makes the earlier-invented property more desirable to consumers in the exact same way that the later development of a complementary thing for an earlier-invented thing makes the earlier-invented thing more valuable to consumers.[292] For example, in the Abby-Bob hypothetical,[293] Bob's later discovery of a better metal for making springs only increases the utility of the mechanical design for a mousetrap invented by Abby to consumers. The inventive property *being made of sliding plates* (Abby) is made more valuable by the later invention of the property *being made of a particular metal* (Bob). In a patent regime in which rewards are

291. The possibility of purely substitute innovative properties is not relevant to a discussion of improvements. Successive inventors who generate properties that are purely substitutes do not produce improvements. If the later inventor's advances completely supplant the earlier inventor's advances in the later-produced thing, then the later-produced thing no longer embodies the earlier inventor's ideas and it is not an improvement. *See supra* Section III.A.3. For example, imagine a variant of the Abby-Bernard hypothetical, *see supra* Part IV, in which spring-loaded mousetraps are already well-known in the art. In this world, Abby's mousetrap might still be patentable, but the only properties that instantiate Abby's ideas in the claimed mousetraps would be the more specific properties like *being made of sliding plates*. If Bernard invents his mousetrap after Abby does, Bernard's mousetrap would not be an improvement. The properties that a mousetrap must possess to instantiate Bernard's inventive ideas—properties like *being made of a wire that can move in an arc in relation to a fixed base*—cannot coexist with the properties that instantiate Abby's ideas in a single, indivisible, and functional mousetrap.

292. *See supra* notes 269–71 and accompanying text (discussing the effect of a later-developed complement on the value of an earlier patent).

293. *See supra* Section V.A.

determined by the market value of embodied ideas and willingness to pay,[294] Abby's patent rights should become more valuable (or, at the least, remain unchanged) because of Bob's invention. However, if Abby's patent rights were to not encompass the overlooked improvement, then Bob's invention would decrease the value of Abby's patent because Bob's mousetrap, viewed as a thing, is a substitute for the mousetraps disclosed by Abby, and Abby must compete on price with Bob's mousetrap.[295] This outcome would add insult to injury: the development of the overlooked improvement increases the value of the set of things that embody an inventor's ideas, but it would trigger a radical decrease in the patentee's profit.

The successive inventors in an overlooked improvement scenario should have rights that mimic the rights given to the earlier inventor of a hammer and a later inventor of an improved nail. In order to achieve this parity, earlier-issued patents must routinely encompass overlooked improvements. The earlier inventor of a hammer and a later inventor of an improved nail can obtain patents that block the hammer-and-improved-nail bundle. For Abby and Bob to have blocking patents with respect to the bundle of their innovative properties, the scope of Abby's earlier-filed patent must be construed so as to encompass Bob's later-developed improvement.[296] It is true that the creation of blocking patents in an overlooked improvement scenario requires the scope of the earlier-filed patent to grow over time as an earlier-filed patent must describe the self-same thing that an improver invents only at a later point in time. In contrast, patents that block the purchase of a bundle of distinct things can come to pass without any post-filing shifts in the set of things that fall within claim scope.[297] However, to forbid the

294. *See supra* Section VII.A.

295. *See supra* notes 272–75 and accompanying text (discussing the effect of a later-developed substitute on the value of an earlier patent).

296. The recognition of overlooked-improvement cases as cases involving successively invented complements—or, more precisely, the "easy" nature of overlooked improvement cases that follows from this recognition—does have the potential to aggravate the royalty stacking problem that exists when goods consumed by consumers include many distinct patented technologies. *See supra* note 116. However, in most situations this is a necessary cost of ensuring that patents structure a market for embodied ideas.

297. The distinction between successively invented complementary things and overlooked improvements leads to a distinction between two types of blocking patents. *Cf.* Collins, *supra* note 66, at 510 n.66 (distinguishing overlapping and economic blocking patents). The blocking patents involved in overlooked-improvement cases are *overlapping* blocking patents as the patents overlap and the self-same, indivisible thing falls within the scope of both patents. For overlapping blocking patents to exist in improvement scenarios, the set of things described by the earlier-filed patent must grow over time. In contrast, the blocking patents at issue when there are successively invented complementary things are *economic* blocking patents because, although the patents do not overlap and encompass the

earlier-filed patent in an overlooked-improvement scenario from growing over time would be to allow a formalistic notion that peripheral claims are "fixed" to triumph over the economic reasoning.[298] Failing to allow an earlier-filed patent to encompass an overlooked improvement would be economically akin to holding that the use of an earlier-patented hammer to pound on an after-arising nail is a non-infringing use of the hammer with which the owner of the hammer patent must compete on price. It would distort the proportionality of contribution and reward to an unacceptable extent.[299] It would be a nonsensical economic result in which a formalism about patents protecting only the things that a patentee actually disclosed and made available to the public at the time of filing eviscerates the ability of the patent regime to structure a market for embodied ideas.[300]

In contrast to overlooked improvements and their purely complementary, successively invented properties, classic improvements and their part-complement, part-substitute properties have no simple, intuitive parallel in terms of complements and substitutes in the world of successively invented things. To the extent that the later inventors' properties compound with and complement the earlier inventors' more general properties, the private value of the earlier inventor's patent should increase upon the later invention of the improvement. To achieve this end, the successive inventors

---

self-same thing, consumers experience the patents as blocking because each one reads on one of two things that consumers want to consume together. If there is a first patent that encompasses A and a second patent that encompasses B, the technology bundle A+B is subject to economic blocking patents. *See* Lemley, *supra* note 2, at 1010 n.87 (noting that patents on distinct complementary goods are a form of blocking patents).

298. *Cf. infra* notes 308–11 and accompanying text (discussing the strong fixation theory of literal claim scope).

299. Couched in the rhetoric of the rules-standards debate, the rule-like option of examining only innovative things and ignoring the more particular properties that make those things innovative would make the rule unacceptably over- and under-inclusive with respect to the justified outcome. *See supra* notes 67–68 and accompanying text (framing the choice between things and properties as the locus of invention in terms of the rules-standards debate).

300. In fact, if one pushes on the distinction between overlapping and economic blocking patents, *see supra* note 297, the distinction often falls apart because inventions can be claimed with differing amounts of matter included within the description. *See supra* Section V.B.3 (discussing claims with limitations that describe the prior art). The hypothetical involving the hammer and improved nail intuitively seems like the successive invention of distinct complementary goods. Provided that the earlier patent claims a hammer and the later patent claims the improved nail, the patents at issue are economic blocking patents. However, if the earlier hammer inventor files a dependent claim in his patent that recites both a hammer and a nail as limitations, then a later claim by the inventor of an improved nail gives rise to overlapping blocking patents, as both patents read on the hammer-and-improved-nail bundle.

should have rights that mirror the rights of the successive inventors of distinct complementary things, and the earlier-filed patent must encompass the improvement. However, to the extent that the later inventors' properties supplant and thus substitute for the earlier inventors' more specific properties, the private value of the earlier inventor's patent should decrease upon the later invention of the improvement. To achieve this end, the successive inventors should have rights that mirror the rights of the successive inventors of distinct substitute things, and the earlier-filed patent must not encompass the improvement. Patent protection in the intuitive realm of successively invented distinct things offers no halfway house to match the in-between, part-complement and part-substitute status of a classic improvement.[301] The issue in play in classic-improvement cases is whether the complementary or substitute nature of the successively-invented properties should be given greater weight, as there is no middle ground.[302] It is the need to simply ignore either the part-complement or the part-substitute nature of the successive inventions, and to treat successive inventors as though they invented either pure complements (which should give rise to blocking patents) or substitutes (which should not), which makes classic-improvement cases legitimately contested.[303]

---

301. The part-complement, part-substitute mixture is not the same phenomenon as imperfect complements and substitutes. *See supra* note 257. A possible analogy might involve an earlier patent that discloses compound A and the later discovery of an improvement A′ that is capable of both increasing the efficacy of A (being a complement) and being taken instead of A (being a substitute). In this scenario, however, each individual consumer experiences A and A′ as either complements or substitutes. In the classic improvement, the properties consumed by each consumer are a part-complement, part-substitute mixture of the earlier- and later-invented properties.

302. Patent regimes in many foreign countries include a "dependency license" provision under which later improvers can obtain a compulsory license to practice earlier inventions. John H. Barton, *Patents and Antitrust: A Rethinking in Light of Patent Breadth and Sequential Innovation*, 65 ANTITRUST L.J. 449, 457–58 (1997); Merges, *supra* note 2, at 102–05. If a court is competent to assess the value of successive improvers' contributions to the overlooked improvement, this solution may provide a middle ground for classic-improvement cases.

303. At first glance, the proposal that patents should reach into improvements with later-developed complementary properties seems to contradict the common understanding of the implications of complements in the fair use analysis in copyright law. In *Ty, Inc. v. Publications International*, the Seventh Circuit stated a general rule that "copying that is complementary to the copyrighted work (in the sense that nails are complements of hammers) is fair use, but copying that is a substitute for the copyrighted work (in the sense that nails are substitutes for pegs or screws) . . . is not fair use." 292 F.3d 512, 517 (7th Cir. 2002). Thus, under the fair use doctrine of copyright law, later-developed complements are less—not more—likely to wind up within the rights of the earlier author. The seemingly opposed nature of the conclusions on complements drawn in the improvements analysis in this Section and the fair use analysis in copyright can be explained by recognizing that the good at issue is being framed differently in each situation. In the improvements context, it is

VIII. CODA: RETHINKING THE "PERIPHERAL" IN PERIPHERAL CLAIMS

This Part illustrates the compatibility of a conceptual framework in which properties are the locus of invention in improvement cases and the contemporary peripheral claiming regime. Upon first impression, such a framework may appear to be incompatible with a peripheral claiming regime, and to mandate instead a central claiming regime, because of its invocation of the "spirit" of a set of claimed things and the attention that it pays to the point of novelty of a patented invention.[304] This Part argues that a focus on properties as the locus of invention in improvement cases is compatible with a peripheral claiming regime, although it does require some common misconceptions about the nature of peripheral claims to be recognized and abandoned. One of the principal goals of this Article is to argue that a conceptual shift is needed to reduce the gap between patent theory and the reality of patent protection.[305] A similar conceptual shift in the theory of what constitutes a peripheral claim is needed if the theory of peripheral claims is to map onto the contemporary reality of peripheral claims. A focus on properties as the locus of invention is compatible with peripheral claims, but peripheral claims turn out not to be what they are commonly presumed to be.

Today's patent claims are commonly described as peripheral because they describe the full set of things that are literally encompassed within a patent claim, right out to its periphery or outer boundaries.[306] Historically, patent

the set of properties contributed by the later actor that is at issue. In the fair use context, it is the allegedly infringing work as a whole that is at issue. Translated into the patent context, this latter framing would require analysis of the later-developed things as complements to or substitutes for the earlier-disclosed and earlier-claimed things. *See supra* Section VII.C.2.b. The later development of a non-infringing complementary thing will increase the private value of the earlier patent or copyright, suggesting that the harm from allowing later-developed complementary things to be non-infringing will not be significant. In contrast, the later development of a non-infringing substitute thing can undermine the ability of the owner of either a patent or a copyright to obtain rents on the things that he actually did invent or author, and this result should be allowed only when the later actor has outmoded the earlier inventor's or author's contribution. When framed as things, patent improvements are usually substitutes for the earlier-disclosed and earlier-claimed things. *See supra* notes 276–78 and accompanying text. Therefore, the concern that motivates differential treatment of complements and substitutes in the fair use analysis does not come into play when addressing complementary and substitute properties in improvements under patent law.

304. *See supra* notes 64–68 and accompanying text.

305. *See supra* text accompanying note 11.

306. Warner-Jenkinson Co. v. Hilton Davis Chem. Co., 520 U.S. 17, 27 n.4 (1997). *But cf. supra* note 139 (noting that the DOE expands a patentee's interest beyond the outer boundaries of a peripheral claim).

claims were not always administered through the demarcation of a periphery. In the first part of the nineteenth century, patent claims were "central" claims.[307] Central claims publicize only an archetypal example of a patented invention, leaving the outer bounds of the claim undefined. In a central claiming regime, infringement determinations are made by comparing an allegedly infringing technology to the publicized example and assessing whether the two are more or less similar than a legal threshold of similarity that determines the scope of the patentee's rights. The outer boundary of the claim is never specified.

Two themes echo through many contemporary discussions of peripheral claims. Both mistakenly presume that peripheral and central claiming must be a strict dichotomy, and both define peripheral claims by the absence of the primary qualities of central claims.

The first theme is fixation. It is black-letter law that the scope of a peripheral claim is fixed ex ante in the sense that the meaning of claim terms to the PHOSITA must be rooted in time on the date on which the claim is filed.[308] Commonly, this black-letter requirement for fixation is interpreted in a strong fashion to mandate fixing the set of claimed things to the set of things made available to the public on that date and, inversely, the exclusion of after-arising technologies from claim scope.[309] This strong interpretation of the fixation requirements treats the distinction between central and peripheral claims as a true dichotomy: central claims do not allow the PHOSITA to identify the full set of claimed things ex ante, but peripheral claims do. Two normative arguments are often made to defend this strong, thing-centric notion of fixation. First, fixation of the set of claimed things is a means to the end of ensuring effective public notice. The post-filing shift in the scope of a claim that is required for a claim to encompass after-arising technology of any kind, it is argued, destabilizes the meaning of a claim and undermines the public-notice benefits of peripheral claims.[310] Second,

307. For extended, and in some respects divergent, discussions of the distinction between central and peripheral claims, see Dan L. Burk & Mark A. Lemley, *Fence Posts or Sign Posts? Rethinking Patent Claim Construction*, 157 U. PA. L. REV. 1743 (2009); Jeanne C. Fromer, *Claiming Intellectual Property*, 76 U. CHI. L. REV. 719 (2009).

308. Phillips v. AWH Corp., 415 F.3d 1303, 1313 (Fed. Cir. 2005) (en banc).

309. Collins, *supra* note 66, at 501–10 (discussing the strong fixation theory of literal claim scope). For scholarly commentary favoring the strong fixation theory in principle, but not in name, see Charles W. Adams, *Allocating Patent Rights Between Earlier and Later Inventions*, 54 ST. LOUIS U. L.J. 55 (2009); Christopher A. Cotropia, *"After-Arising" Technologies and Tailoring Patent Scope*, 61 N.Y.U. ANN. SURV. AM. L. 151 (2005); Holbrook, *supra* note 156; Timothy R. Holbrook, *Possession in Patent Law*, 59 SMU L. REV. 123 (2006).

310. *See* Mark A. Lemley, *The Changing Meaning of Patent Claim Terms*, 104 MICH. L. REV. 101, 112–22 (2005).

fixation of the set of things is viewed as necessary to create a market for ideas and to limit the patentee's reward so that it is proportional to his contribution to technological progress. As expressed in a recent Federal Circuit opinion on claim construction, allowing a claim to encompass later-developed technology "compromises two fundamental tenets of the patent system: first, that the applicant must be the 'inventor' of the things covered by the patent claims, and second, that the right to exclude will be no broader than the inventor's enabling disclosure."[311]

The second theme is what can be called the self-sufficiency of peripheral claims. The "spirit" of an invention—that is, the features of the invention that differentiate it from the prior art—is taken to be irrelevant to the doctrinal mechanics of the infringement and validity analyses in a peripheral claiming regime, making the language of the claim alone sufficient to communicate the information to the public that is required for the public to have notice of the claim's scope.[312] Identifying precisely how and why a patent is different from the prior art requires an information-intensive and contested comparative analysis. By defining the outer boundaries of an invention through a descriptive text, a peripheral claim is presumed to avoid this comparative analysis and thereby make patent rights less costly and more certain. The limitations of a claim become a simple checklist. If all of the claim limitations are present in an accused device, the device infringes; if all of the claim limitations are present in a prior-art technology, the claim is invalid as anticipated.[313] To the extent that it holds true, this self-sufficiency theme starkly distinguishes peripheral claims from central claims. The similarity analysis in a central claiming regime is necessarily performed in

311. Superguide Corp. v. DirecTV Enters., 358 F.3d 870, 898 (Fed. Cir. 2004) (Michel, J., concurring). The strong, thing-centric fixation argument is also voiced during analyses of the enablement and written description that require the "full scope" of the claim to be enabled or possessed, respectively. Collins, *supra* note 58, at 1088 (discussing "full scope" rhetoric in enablement cases).

312. Ormco Corp. v. Align Tech., Inc., 498 F.3d 1307, 1322–23 (Fed. Cir. 2007) ("This court . . . has rejected a claim construction process based on the 'essence' of an invention."). Of course, claim language is not truly self-sufficient because many interpretive sources, including the patent disclosure, can be called upon to determine the meaning of patent claims. *Phillips*, 415 F.3d at 1311–19. The self-sufficiency theme more narrowly assumes that the point of novelty of a claim is irrelevant and that the rules of claim construction do not require a claim to be construed in light of the "spirit" of the invention. The patented technology clearly must be different from the prior art to survive the validity analyses under § 102 and § 103, but the way or ways in which a patented technology differs from the prior art broadly writ need never be isolated, identified, and catalogued.

313. This "all elements" rule of infringement (and anticipation) is part of the bedrock of contemporary patent law. *See, e.g.*, TechSearch, L.L.C. v. Intel Corp., 286 F.3d 1360, 1374–75 (Fed. Cir. 2002).

light of the "spirit" of the invention. To infringe in a central claiming regime, the allegedly infringing technology must be similar to the features of the disclosed, archetypal embodiment that distinguish it from the prior art.[314]

These peripheral claim themes shore up both the notion that innovative things should be viewed as the primitives of what an inventor invents and its corollary, the refusal to countenance inventive properties of things as the locus of invention. However, a conceptual framework in which innovative properties are the locus of invention is compatible with a peripheral claiming regime. On a conceptual level, peripheral claims are not only compatible with a focus on the properties of things, but, more strongly, they rely on the properties of things to perform their boundary-drawing function. The descriptive language of a peripheral claim defines the set of things encompassed within the claim by describing particular properties that things must possess to be included within the claimed set.[315] More pragmatically, the best proof of the compatibility of a focus on properties as the locus of invention and peripheral claims is in the pudding of contemporary patent protection. The contemporary claiming regime is widely regarded as a peripheral claiming regime—especially when the default rules of claim construction and literal infringement apply[316]—yet the brunt of the work in this Article performed by the conceptual framework in which properties are the locus of invention is descriptive. The framework simply explains what the contemporary patent regime is already doing in classic- and overlooked-improvement cases.[317] Unless one argues that contemporary claims are not in fact peripheral (or that classic and overlooked improvements do not receive differential treatment), then it is problematic to argue that peripheral claiming and a focus on properties as the locus of invention are incompatible. What paying attention to properties as the locus of invention does illustrate, however, is that the dominant conceptual paradigm of what it means for a claim to be peripheral is sorely in need of revision. Peripheral claims cannot be defined in sharp contradistinction to central claims; a workable peripheral claiming regime is not everything that a central claiming regime is not.[318] The

314. *See* Burk & Lemley, *supra* note 307, at 1746.

315. Lefstin, *supra* note 63, at 1145 ("[Peripheral claims] recite a set of characteristics, or properties, that define the subject matter encompassed by the patent.").

316. The contemporary patent regime does include rule sets other than this default—such as the rules of means-plus-function claiming and the doctrine of equivalents—that are based on the principles of central claiming. *See infra* note 318.

317. *See supra* notes 84–85 and accompanying text.

318. Recent scholarship on patent claims has highlighted that the contemporary claiming regime is in fact a hybrid, part-peripheral-and-part-central regime. Burk & Lemley, *supra* note 307, at 1771–78; Fromer, *supra* note 307, at 735–41. By arguing that contemporary

two themes that run through contemporary discussions of peripheral claims need to be recognized as misleading and abandoned. To the extent that judicial rhetoric defines peripheral claims by the strict fixation of the claimed set of things on the date of filing or the irrelevance of the point of novelty in the determination of claim scope, there is a significant gap between what courts are saying and what they are doing.

The very existence of the "easy" overlooked-improvement cases is an Achilles heel of the strong variant of the fixation theme, under which peripheral claims fix the set of things that falls within the scope of a claim on the date the claim is filed. These cases definitively disprove this strong theory of fixation as a descriptive matter, demonstrating that literal claim scope already grows over time to encompass after-arising technology on a routine basis.[319] Furthermore, the two policy reasons often given for adhering to a strong variant of the fixation theory and preventing claim expansion over time are not persuasive.[320] First, as I have explored at length elsewhere, the fixation of meaning that is required to ensure reasonable public notice of claim scope does not entail limiting the set of things encompassed within the claim to the set of things of which the PHOSITA was aware on the date the

peripheral claims cannot be defined in sharp contradistinction to central claims, this Part reinforces the notion that the contemporary claiming regime is already a hybrid of sorts. However, the nature of the hybridity unveiled here is new. The hybridity at issue is turned from an either-or hybridity into a both-and hybridity. The argument in recent scholarship is that the patent regime contains many rule sets and that some of these rule sets (other than the default rules that govern literal claim construction) employ the principles of central claiming. In this Article, the argument is that peripheral claiming, even at its most peripheral when the default rules of claim construction and literal infringement apply, is not what it is commonly presumed to be. In other words, even when the "all elements" rule applies, *see supra* note 313, the set of things literally described by a claim can grow over time and the "spirit" of an invention is relevant to determinations of claim scope and infringement.

319. Advocates of a strong fixation theory of literal claim scope often suggest that the DOE provides the necessary protection for later-developed improvements. *See* Cotropia, *supra* note 309, at 185–201; Holbrook, *supra* note 156, at 36–45. However, as an empirical matter, the DOE does not seem to be performing this function. The recent decline of the DOE has been well-documented, and the doctrine is today rarely dispositive of infringement. John R. Allison & Mark A. Lemley, *The (Unnoticed) Demise of the Doctrine of Equivalents*, 59 STAN. L. REV. 955 (2007); David L. Schwartz, *Explaining the Demise of the Doctrine of Equivalents*, 26 BERKELEY TECH. L.J. 1157 (2011). It is therefore difficult to argue as a descriptive matter that the DOE shoulders the entire burden of protecting a patentee's interests in after-arising technology. Furthermore, the examples of courts treating overlooked improvements as "easy" cases, *see supra* Sections V.B.1 and V.B.2, all involve later-developed improvements that, today, are presumed to fall within the literal scope of earlier-filed claims.

320. *See supra* notes 310–11 and accompanying text (presenting these two arguments).

claim was filed.[321] It is possible to fix a stable periphery with language even without complete knowledge of the set of things that, after the expiration of a twenty-year patent term, may populate that set. The meaning of whatever claim language Abby uses to claim her invention does not have to shift from its filing-date meaning in order to be capable of describing Bernard's and Bob's improvements that are not invented until after Abby's filing date.

Second, while it is true that patents should structure a market for embodied ideas and limit inventors' rewards so that they are proportional to their contributions to technological progress, it is counterproductive to measure a patentee's contribution solely by looking at the innovative things that he made available to the public at the time of filing.[322] Even if patent protection is to be trimmed back to what a patent-minimalist would likely prefer, adopting the line between technologies that are known on the date a claim is filed and after-arising technologies as the line that defines literal claim scope distorts proportionality to an unacceptable extent. The fine-grained properties of a thing that embody an inventor's ideas may remain steadfast and unchanging even as the coarse-grained things that possess

321. Provided that courts employ ideational rather than denotational meaning, there can be play between what claim language means to a person (the meaning-scope of a claim) and the set of things to which the claim language refers for that person (the thing-scope of a claim). Collins, *supra* note 66, at 536–53 (distinguishing ideational meaning, which allows this play, from denotational meaning, which does not). Meaning-scope can remain fixed even as thing-scope expands over time. *Id.* For example, the meaning of "bachelor" is determined by the relationship between the word "bachelor" and other words and concepts in the English language, such as "male" and "unmarried." Thus, the meaning of the term "bachelor" would not be destabilized if a race of extraterrestrials, in which there are also unmarried males, were to be discovered and the set of things known to be described by the term "bachelor" were to be expanded in an unexpected way. For an example from a classic patent case, consider the claim at issue in *In re Hogan*, 559 F.2d 595 (C.C.P.A. 1977), a controversial case implicating the reach of an earlier-filed claim into later-developed improvements. The claim at issue described "[a] normally solid homopolymer of 4-methyl-1-pentene." *Id.* at 597. The ideational meaning of this language can remain fixed over time, referring to all molecules that are made up of long chains of a single repeated unit (4-methyl-1-pentene) and that are solid under normal conditions. Growth in the set of distinct things that are known to fit this description over time need not entail a shift in the ideational meaning of the claim language. This observation about the stability of language should not be taken to imply that allowing the *Hogan* claim to encompass after-arising homopolymers is optimal patent policy. What it does imply is, more narrowly, that the stability of the meaning of the claim language over time should not be invoked to categorically prevent improvements from falling within literal claim scope.

322. *Cf.* Superguide Corp. v. DirecTV Enters., 358 F.3d 870, 898 (Fed. Cir. 2004) (Michel, J., concurring) (arguing "that the applicant must be the 'inventor' of the *things* covered by the patent claims") (emphasis added).

those properties change over time.[323] Patentees will be rewarded in proportion to the value of their ideas only if improvements that fully possess the innovative properties invented by an earlier inventor fall within the scope of the inventor's claim.[324] To achieve this result, the set of things within the scope of a claim must sometimes grow over time after a patent has been filed.

Similarly, the distinct treatment afforded to classic and overlooked improvements undermines the self-sufficiency theme and the argument that a peripheral claiming regime always allows decision makers to remain ignorant of the particular inventive features of a patented technology when determining claim scope and validity.[325] The self-sufficiency theme is a useful trope when dealing with technologies that were known at the time a patent is filed. However, knowledge of the particular properties of an earlier-claimed technology that are innovative is critical in defining whether an improvement is a classic or overlooked improvement.[326] Therefore, the claim language, in isolation, does not convey enough information to demonstrate how the claim should grow over time to encompass improvements. A court must look to something that resembles the "spirit" of the invention—an entity that this Article has proposed can be identified by the innovative properties of a claimed set of things. Consider a simple claim to a "plastic widget." Should this peripheral claim encompass an improvement that possesses the exact mechanical structure of a widget disclosed in the specification but that is made out of an after-arising plastic? The self-sufficiency thesis fails because the answer depends on the point of novelty of the patented invention. If the "spirit" of the plastic-widget invention lies in the properties of the widget that embody the mechanical configuration of the widget, then the claim will and should routinely grow over time to encompass the improvement (as the improvement is an overlooked improvement). However, if the "spirit" of the plastic-widget invention lies in the invention of plastic itself (and the claim is therefore a dependent claim in which the widget design is the added limitation), then the expansion of the claim over time that is needed to

323. *See supra* Section II.C (arguing that a focus on properties permits a finer-grained analysis than a focus on things does). The coarse-grained thing may change because some of its properties other than those that embody an earlier innovator's ideas may change.

324. In other words, looking only at the things that an inventor disclosed and made available to the public at the time of filing is an unacceptably crude proxy for the set of things that the inventor should control. *See supra* notes 67–68 and accompanying text (framing the choice between things and properties as the locus of invention in terms of the rules-standards debate).

325. *See supra* notes 312–14 (summarizing the self-sufficiency theme).

326. *See supra* Part IV and Section V.A.

encompass a widget made out of an improved plastic does and should raise a highly contested issue (as the improvement is a classic improvement).[327] In sum, precisely how the claimed set of things grows over time can only be determined by examining the "spirit" or point of novelty of the invention that is disclosed in the specification, even in a peripheral claiming regime.[328]

In sum, a focus on properties as the locus of invention in improvement cases is compatible with a peripheral claiming regime, but, in order to grasp this compatibility, the fixation and self-sufficiency themes that are commonly used to explain the nature of a peripheral claim must be abandoned. Peripheral claims cannot be defined in sharp contradistinction to central claims. Rather, three new themes need to be adopted. First, the fixation of a peripheral claim requires only the stabilization of a linguistic description, not an exhaustive tally of the members of the set of described things.[329] Unexpected or unforeseeable technological developments may produce improvements during the term of a patent, and linguistically stable claim language can describe these improvements without undermining effective public notice.[330] Second, reiterating a theme that has often been explored before, the level of generality at which valid peripheral claims are allowed to be drawn is one of the key policy levers in fine-tuning the incentives created by patents. Concern over the level of generality of a claim meshes seamlessly with a focus on properties as the locus of invention, as peripheral claims define sets of things by listing the properties that things must possess to be members of the claimed set. When classic improvements are at issue, it is the level of generality at which a claim can be drawn that determines how general the least-general naked property of an improvement must become before the

327. The same exercise can be performed using a claim to a "programmed computer." The set of after-arising technologies that literally infringe the claim hinges on whether the software or the hardware is the point of novelty. *Cf. supra* Section V.B.2 (discussing overlooked improvements in the computer arts).

328. In a recent case, the Court of Appeals for the Federal Circuit broke from its refusal to recognize the importance of the point of novelty in validity doctrines and stated that the point of novelty of an invention must be fully enabled by the specification. Automotive Techs. Int'l v. BMW of N. Am., Inc., 501 F.3d 1274 (Fed. Cir. 2007) ("Although the knowledge of one skilled in the art is [] relevant [in the enablement analysis], the novel aspect of an invention must be enabled in the patent."). The "easy" nature of overlooked improvements suggests that the converse is also true: an enabled claim can encompass after-arising technologies that the specification did not teach the PHOSITA how to make and use at the time of filing if the difference between the disclosed and after-arising technologies does not lie at the point of novelty of the patented invention.

329. *See supra* note 321 and accompanying text.

330. The DOE also has a role to play in granting patentees rights to exclude from after-arising technologies. *See supra* note 319. However, it is not a patentee's sole recourse for patent rights that encompass improvements.

earlier patentee's claims can no longer encompass the improvement.[331] Third, and perhaps most controversially, the point of novelty of an invention is sometimes relevant when determining claim scope, even in a peripheral claiming regime in which an "all elements" rule applies.[332]

More precisely, the second and third themes are interconnected. An accused device must possess elements that correspond to all of the limitations of a claim, but the point of novelty must be considered when identifying the subset of claim limitations whose level of generality must be policed to prevent the claim from becoming impermissibly abstract. The level of generality at which claim language is drawn is only a concern when the claim language at issue describes an innovative property of the claimed things, i.e., a property that the inventor of the patent at issue invented and that is implicated in differentiating the claimed invention from the prior art so as to make the claimed things novel and nonobvious. If the claim language describes a property of the innovative things that is functionally independent of the properties that embody the inventor's innovative ideas, then the claim language can be construed in an extremely general manner without over-rewarding the inventor.[333] The "spirit" of an invention or the point of novelty of a claim matters in improvement cases, even in a peripheral claiming regime.[334] It allows courts to pay attention to the claim limitations whose level of generality must be strictly policed, on the one hand, and to ignore the claim terms whose level of generality need not be policed, on the other.

The need for revision in contemporary understandings of the nature of peripheral claims brings us back to where we began. This Article set out to correct a blind spot in contemporary theory. The overlooked improvements were being overlooked; they were hiding in plain sight.[335] Now, it is finally possible to hazard a guess about why they were able to hide in plain sight. The two recurring themes about peripheral claims discussed above[336] demonstrate that there is a dominant conceptual paradigm of peripheral claiming in action today that focuses exclusively on innovative things as the locus of patentable invention. It is precisely the dominance of this conceptual

331. *See supra* notes 242–48 and accompanying text.

332. *See supra* notes 325–28 and accompanying text.

333. In other words, overlooked improvements should routinely fall within the scope of earlier-issued claims. *See supra* Section VII.B.1.

334. Both claim construction and the disclosure doctrines are implicated in curtailing the permissible level of generality of a claim. *See supra* note 130 and accompanying text. Therefore, both of these areas of patent doctrine must pay attention to the point of novelty in improvement cases.

335. *See supra* text accompanying note 59.

336. *See supra* notes 308–14 and accompanying text.

paradigm that has made the overlooked improvements so easy to overlook. As Thomas Kuhn has argued in the context of scientific progress, facts that do not fit well with the dominant conceptual paradigm are often overlooked during periods of "normal science," because the conceptual paradigm serves as a mental screen that filters the way in which we see the world:

> Closely examined . . . [the] enterprise [of normal science] seems an attempt to force nature into the preformed and relatively inflexible box that the paradigm supplies. No part of the aim of normal science is to call forth new sorts of phenomena; indeed those that will not fit the box are often not seen at all.[337]

The overlooked improvements, and their distinction from classic improvements, are phenomena that do not "fit in the box" of the contemporary paradigms of what constitutes a peripheral claim. "Normal patent theory," if you will, did not have as its goal the identification or explanation of these phenomena. Once the phenomena are put openly on the table, however, the need for a paradigm shift (in Kuhn's terminology) becomes self-evident.[338] Bluntly put, it is time to wake up and smell the coffee. The conceptual structures should shift to reflect the facts on the ground, not the other way around. Innovative things can no longer be taken to be the primitives of what an inventor invents. At least when addressing patent protection for improvements, it is the innovative properties of claimed things that must be viewed as the locus of the invention—a proposition that is axiomatic once it is recognized that, first, innovative properties are the entities that instantiate innovative ideas in things[339] and, second, the patent regime strives to structure a market for innovative, embodied ideas.[340]

IX. CONCLUSION

Scholarship on the reach of patent scope into improvement has been farsighted. It has identified and addressed one type of improvement—what this Article terms a classic improvement—but it has failed to notice another common type of improvement that is close at hand and in plain sight—what this Article terms an overlooked improvement. Unlike classic improvements, overlooked improvements are "easy" to deal with in patent infringement

337. THOMAS KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS 24 (1962).

338. *See generally id.* (arguing that scientific progress does not proceed through uniform, gradual accretion but rather occurs through periods of normal science separated by disruptive paradigm shifts).

339. *See supra* note 220 and accompanying text.

340. *See supra* Section VII.A.

cases as earlier-filed patents routinely encompass later-developed improvements. The conventional theory on patent protection for improvements cannot explain why the overlooked-improvement cases are "easy" cases. The only way to explain the different treatment doled out to the different types of improvements is to break with the dominant conceptual paradigm which takes things to be the primitives of what an inventor invents. At least in the context of improvements, it is important to get into the "spirit" of innovative things and identify the innovative properties of things that are the locus of invention. This conceptual framework that focuses more finely on innovative properties, rather than more bluntly on innovative things, allows invention to be studied at a higher resolution and a finer granularity. It reveals information that is lost in a lower-resolution analysis when things are taken to be the primitives of invention, and it is this very information that is needed to explain how classic and overlooked improvements are distinct as a descriptive matter and why this distinction is important as a normative matter. When innovative properties are treated as the locus of invention, the successive inventions of an earlier patentee and a later improver can be identified as either pure complements (giving rise to overlooked improvements) or as a part-complement and part-substitute mixture (giving rise to classic improvements). When the issue of patent protection for improvements is viewed through this conceptual lens that focuses on innovative properties and that allows a "spirit" of an invention to be recognized, the concepts of complements and substitutes can, for the first time, be brought to bear to explain how courts should craft claim scope.