29:1 BERKELEY TECHNOLOGY LAW JOURNAL

2014

Pages

1

to

376

Berkeley Technology Law Journal

Volume 29, Number 1

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.

Copyright © 2014 Regents of the University of California.
All Rights Reserved.

Berkeley Technology Law Journal
University of California
School of Law
3 Boalt Hall
Berkeley, California 94720-7200
btlj@law.berkeley.edu
http://www.btlj.org

BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 29 NUMBER 1 SPRING 2014

TABLE OF CONTENTS

ARTICLES

GAME OVER FOR FIRST SALE 1
Stephen McIntyre

WHY NOT PRIVACY BY DEFAULT? 61
Lauren E. Willis

WHY THEY CAN WATCH YOU: ASSESSING THE CONSTITUTIONALITY OF WARRANTLESS UNMANNED AERIAL SURVEILLANCE BY LAW ENFORCEMENT 135
Brandon Nagy

REGULATING DISRUPTIVE INNOVATION 175
Nathan Cortez

BEYOND EFFICIENCY: CONSEQUENCE-SENSITIVE THEORIES OF COPYRIGHT 229
Oren Bracha & Talha Syed

FREE TRADE IN PATENTED GOODS: INTERNATIONAL EXHAUSTION FOR PATENTS 317
Sarah R. Wasserman Rajec

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 in print three times each year (March, September, December), with a fourth issue published online only (July), 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, University of California, Berkeley Law—Library, LL123 Boalt Hall—South Addition, Berkeley, CA 94720-7210.

Correspondence. Address all correspondence regarding subscriptions, address changes, claims for non-receipt, single copies, advertising, and permission to reprint to Journal Publications, University of California, Berkeley Law—Library, LL123 Boalt Hall—South Addition, Berkeley, CA 94705-7210; (510) 643-6600; JournalPublications@law.berkeley.edu. *Authors*: see section titled Information for Authors.

Subscriptions. Annual subscriptions are $65.00 for individuals and $85.00 for organizations. Single issues are $30.00. 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 ($30.00) 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 29 BERKELEY TECH. L.J. ___ (2014).

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*, the *Bolt*, a collection of short comments and updates about new developments in law and technology written by BTLJ members, and *BTLJ Commentaries*, an exclusively online publication for pieces that are especially time-sensitive and shorter than typical law review articles.

.

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, including his or her full name, credentials, degrees earned, academic or professional affiliations, and citations to all previously published legal 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).

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:

Partners

COOLEY LLP

COVINGTON & BURLING LLP

FENWICK & WEST LLP

ORRICK, HERRINGTON & SUTCLIFFE LLP

Benefactors

FISH & RICHARDSON P.C.

KASOWITZ BENSON TORRES & FRIEDMAN LLP

KIRKLAND & ELLIS LLP

LATHAM & WATKINS LLP

MCDERMOTT WILL & EMERY

MORRISON & FOERSTER LLP

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

WEIL, GOTSHAL & MANGES LLP

WHITE & CASE LLP

WILMER CUTLER PICKERING HALE AND DORR LLP

WILSON SONSINI GOODRICH & ROSATI

WINSTON & STRAWN LLP

Members

ALSTON & BIRD LLP

BAKER BOTTS LLP

BAKER & MCKENZIE LLP

BINGHAM MCCUTCHEN LLP

DURIE TANGRI LLP

GILLIN JACOBSON ELLIS LARSEN & LUCEY

GTC LAW GROUP LLP & AFFILIATES

GUNDERSON DETTMER STOUGH VILLENEUVE FRANKLIN & HACHIGIAN, LLP

HAYNES AND BOONE, LLP

HICKMAN PALERMO TRUONG BECKER BINGHAM WONG, LLP

HOGAN LOVELLS LLP

IRELL & MANELLA LLP

KEKER & VAN NEST LLP

KILPATRICK TOWNSEND & STOCKTON LLP

KNOBBE MARTENS OLSON & BEAR LLP

MUNGER, TOLLES & OLSON LLP

O'MELVENY & MYERS LLP

PAUL HASTINGS LLP

ROPES & GRAY LLP

SIDLEY AUSTIN LLP

SIMPSON THACHER & BARTLETT LLP

TURNER BOYD LLP

VAN PELT, YI & JAMES LLP

WEAVER AUSTIN VILLENEUVE & SAMPSON, LLP

BOARD OF EDITORS 2013–2014

Executive Committee

Editor-in-Chief
EMILY CHEN

Managing Editor
SARAH ORRICK

Senior Articles Editors
E. ROSS COHEN
MARK LANGER
ETHAN WEINER

Senior Executive Editor
LALI MADDURI

Senior Annual Review Editors
JULIE KENT
ROBIN KUNTZ

Senior Scholarship Editor
NICHOLAS LAMPROS

Editorial Board

Submissions Editors
BRADY BLASCO
ANAGHA KRISHNAN

Production Editors
DOMINIQUE CAAMANO
JENNA STOKES

Bluebook Editors
ANTHONY CHANRASMI
DANIELA SPENCER
MARGARET YI

Annual Review Editors
LAUREN BLAKELY
EMILY LY

Notes & Comments Editors
SIENA CARUSO
STEVEN S. MCCARTY-SNEAD

Symposium Editors
CHRISTINA FARMER
ALEX S. LI

Web Content Editor
DINA ROUMIANTSEVA

Web & Technology Editors
EDUARD MELESHINSKY
STEVEN SWANSON

External Relations Editor
GAVIN LIU

Member Relations Editor
ANDREW YOCOPIS

Articles Editors

ALI AMIRMAHANI
DALLAS BULLARD
MAISSA CHOURAKI
JONATHAN FRANCIS
S. ZUBIN GAUTAM
JEFFREY HIRSCHEY
BAILEY LANGNER
KASSANDRA MALDONADO
JOE MORNIN
LEXI RUBOW
ALLISON SCHMITT
ROSALIND SCHONWALD
CHRISTOPHER SUNG
AMY TU
BONNIE WATSON
MARK ZAMBARDA

MEMBERSHIP

Vol. 29 No. 1

Associate Editors

SERGIO ALVES, JR.
NATHAN ANDERSON
RAVI ANTANI
AHMAD ASIR
NORRIS BOOTHE
ELIZABETH CAPEL
DUANE CARVER
YIN CHOI
CHAN HEE CHU
SHELI DENOLA
PHIL DJAMBAZOFF
MATTHEW DONOHUE
PRATUSHA ERRABALLI
SIMONE FRIEDLANDER
JOSHUA FURMAN
CAROLINA GARCIA
RYAN GATZEMEYER
ROXANA GUIDERO
MEREDITH HOUSE
ENEDA HOXHA
VANESSA KATZ
JACKIE KIMBLE
JINGYUAN LUO
DIANA OBRADOVICH
SMITA RAJMOHAN
THOMAS STEINER
IRENE SULAIMAN
MATTHEW THANABALAN
MISHA TSUKERMAN
KELLY VARGAS
WENDY LILY WANG

Members

OM ALLADI
AIDA ASHOURI
MARION BARBEZIEUX
STEPHEN BOONE
NICOLE CABRERA
GOLDA CALONGE
STEPHEN CHANG
YOON HYUN AL CHO
SOPHIE COOPER
DANIELA CRONEMBOLD
RIDDHI DAGUPTA
AUDE DE CRAYENCOUR
LUKE DIAMOND
LINDSAY DOCTO
JAMES DUNN
MISA EIRITZ
MOJDEH FARSI
DANIELE FERRONE
ARYEH FINK
JOHN FLYNN
JULY GALINDO
JESUS RAUL GARCIA
ALVARO GARCIA-DELGADO
CAROLYN GILBERT
MARK GRAY
JIEYAO HE
CHRISTOPHER HEALY
GWEN HINZE
YAO-JEN HO
THADDEUS HOUSTON
ZIYAN HUANG
TUP INGRAM
ALICIA INTRIAGO
JENNIFER IVERS
HANNAH JIAM
FRANÇOIS JOLI-COEUR
DAMION JURRENS
MEHTAB KHAN
SAMUEL KLAUS
BINXIN LI
ASHLEY LIN
CHRISTIAN J. MARTINEZ
NICK MONACELLI
ANA CAROLINA MONTEIRO
MATT MORETTI
CHRISTOPHER MORTWEET
WADE ORBELIAN
DANIEL OSBORN
JORDYN OSTROFF
LISA PATEL
KATHIA PAUCHARD
TABETHA PEAVEY
MARCO PENSATO
MALIKE POLAT
STUYVIE PYNE
VANESSA RIBEIRO
JAMES RICE
CRISTINA RIVA
SYDNEY RYAN
ORNNATTHA SABYEROOP
GINNY SCHOLTES
JESUS M. SEPULVEDA
DONGBIAO SHEN
MICHAEL LIU SU
JONATHAN TAMIMI
YUEN TIN
EMMIE TRAN
NOAH WALD
YUE (JOY) WANG
TARA WATERMAN
CURTIS WHATLEY
EVAN WHITTAL-WILLIAMS
ANDREW WU
ZELIN YANG
SORIN ZAHARIA
OGNJEN ZIVOJNOVIC

BTLJ ADVISORY BOARD

ROBERT BARR
Executive Director of the
Berkeley Center for Law & Technology
U.C. Berkeley School of Law
Berkeley, California

ROBERT C. BERRING, JR.
Walter Perry Johnson Professor of Law
U.C. Berkeley School of Law
Berkeley, California

JESSE H. CHOPER
Earl Warren Professor of Public Law
U.C. Berkeley School of Law
Berkeley, California

PETER S. MENELL
Professor of Law and Faculty Director of the
Berkeley Center for Law & Technology
U.C. Berkeley School of Law
Berkeley, California

ROBERT P. MERGES
Wilson Sonsini Goodrich & Rosati Professor
of Law and Faculty Director of the Berkeley
Center for Law & Technology
U.C. Berkeley School of Law
Berkeley, California

REGIS MCKENNA
Chairman and CEO
Regis McKenna, Inc.
Palo Alto, California

DEIRDRE K. MULLIGAN
Assistant Professor and Faculty Director of the
Berkeley Center for Law and Technology
U.C. Berkeley School of Information
Berkeley, California

JAMES POOLEY
Deputy Director General of the
World Intellectual Property Organization
Washington, D.C.

MATTHEW D. POWERS
Tensegrity Law Group, LLP
Redwood Shores, California

PAMELA SAMUELSON
Professor of Law & Information
and Faculty Director of the
Berkeley Center for Law & Technology
U.C. Berkeley School of Law
Berkeley, California

LIONEL S. SOBEL
Visiting Professor of Law
U.C.L.A. School of Law
Los Angeles, California

LARRY W. SONSINI
Wilson Sonsini Goodrich & Rosati
Palo Alto, California

MICHAEL STERN
Cooley LLP
Palo Alto, California

MICHAEL TRAYNOR
Cobalt LLP
Berkeley, California

THOMAS F. VILLENEUVE
Gunderson Dettmer Stough Villeneuve
Franklin & Hachigian LLP
Redwood City, California

BERKELEY CENTER FOR LAW & TECHNOLOGY 2013–2014

Executive Director
ROBERT BARR

Faculty Directors
KENNETH A. BAMBERGER
PETER S. MENELL
ROBERT P. MERGES
DEIRDRE MULLIGAN
PAMELA SAMUELSON
PAUL SCHWARTZ
MOLLY S. VAN HOUWELING

Associate Director
LOUISE LEE

Affiliated Faculty
CHRIS JAY HOOFNAGLE
TALHA SYED
JENNIFER M. URBAN

In Memoriam
SUZANNE SCOTCHMER
1950–2014

GAME OVER FOR FIRST SALE

Stephen McIntyre[†]

ABSTRACT

Video game companies have long considered secondhand game retailers a threat to their bottom lines. Some companies are now experimenting with technological tools to discourage and even prevent gamers from buying and selling used games. For example, "tethering" technology suppresses secondhand sales by permanently identifying particular media items (such as video game discs) with a single user's device. This technology flies in the face of copyright law's "first sale" doctrine, which gives lawful purchasers the right to sell, lease, and lend DVDs, CDs, and other media.

This Article answers a question posed by many commentators: whether it is legal to employ technology that restricts first sale rights. The answer hinges on two statutes: the Digital Millennium Copyright Act ("DMCA") and the Sherman Antitrust Act. The DMCA broadly protects technological measures that restrict access to copyrighted material, which would likely include technologies aimed at suppressing secondhand video game sales. If end users came up with a method for getting around these devices, they could potentially incur liability under the DMCA. However, some courts have interpreted the DMCA narrowly so as to allow circumvention that does not clearly lead to copyright infringement. Since accessing the content on a video game disc or other similar media does not constitute infringement, this construction of the statute would most likely permit users to lawfully circumvent restrictive technology.

Moreover, technology that abridges first sale rights may violate the Sherman Act, which prohibits monopolists from acquiring, maintaining, or extending their market power through predatory or anticompetitive means. Antitrust would be a fitting remedy; the first sale doctrine is intrinsically linked to antitrust jurisprudence. Courts recognize the importance of secondhand markets, and have held that monopolists may not use technology to suppress competition. However, the difficulty of demonstrating that any company possesses monopoly power in the relevant market could be fatal to a Sherman Act challenge.

Given these doctrinal complexities, it may ultimately fall to consumers to vote with their wallets and choose not to patronize companies that engage in business practices that abridge first sale rights. Otherwise, it really could be "Game Over" for the first sale doctrine.

© 2014 Stephen McIntyre.

† J.D., Duke University School of Law; M.A., East Asian Studies, Duke University; B.A., Chinese, Brigham Young University.

TABLE OF CONTENTS

I. **INTRODUCTION** 2

II. **A LINK TO THE PAST: OLD DOCTRINE, NEW TECHNOLOGY** 7

A. THE COPYRIGHT FIRST SALE DOCTRINE 7
B. LICENSING ARRANGEMENTS 11
C. APPLICATION TO COMPUTER SOFTWARE AND VIDEO GAMES 12

III. **ALL YOUR BASE ARE BELONG TO US: FIRST SALE UNDER FIRE** 14

A. THE RISE OF THE END USER LICENSE AGREEMENT 15
B. DIGITAL RIGHTS MANAGEMENT ("DRM") AND THE DIGITAL MILLENNIUM COPYRIGHT ACT ("DMCA") 20

IV. **PLAYING WITH POWER: CIRCUMVENTION TO PRESERVE FIRST SALE** 28

A. EARLY SHOTS ACROSS THE BOW 28
B. TETHERING: THE LATEST ATTACK ON FIRST SALE 30
C. LAWFUL CIRCUMVENTION? 31

V. **ONE-UP: THE COMING COLLISION—AND POSSIBLE SOLUTION** 43

A. THE FIRST SALE DOCTRINE'S ANTITRUST ORIGINS 44
B. ANTITRUST, AFTERMARKETS, AND THE DMCA 46
C. MONOPOLIZATION IN THE VIDEO GAME AFTERMARKET 47
D. THE RELEVANT MARKET PROBLEM 52

VI. **CONCLUSION: GAME OVER?** 59

I. INTRODUCTION

In the early days of 2013, just as the holiday shopping season was coming to a close, an obscure legal filing took the video game world by storm. An Internet user known only as "gofreak" posted a brief message to an online discussion forum describing a patent application that Sony Computer Entertainment, maker of the popular line of PlayStation gaming consoles, had filed with the U.S. Patent & Trademark Office a few months earlier.[1] The application described a technology that would permanently affiliate

1. Gofreak, *Sony Has Researched New Tech for Suppressing Second-Hand Game Sales*, NEOGAF (Jan. 3, 2013, 8:36 AM), http://www.neogaf.com/forum/showthread.php?t=506560.

video game discs with a particular video game console, such that a game would not run if the disc were inserted into another gamer's device.[2] This "Disc ID" system would effectively preclude gamers from lending, trading, or selling video games, since each disc could only function on a single console. According to the patent application, this was Disc ID's main purpose. The filing stated that Disc ID would "suppress" secondhand video game sales by forcing all interested gamers to purchase new discs.[3] The intended outcome was enhanced "redistribution" of profits to game makers and developers.[4]

As the news spread, bloggers and tech publications registered their outrage. The website ITworld called the Disc ID plans "nefarious" and predicted that gamers would "grab their pitchforks and torches" to protest the technology.[5] TechDirt sardonically described the news as coming from the "Sony-obviously-feels-it's-not-hated-enough-already" department.[6] Some commentators were more restrained; Ars Technica, for example, correctly pointed out that "the fact that Sony has applied to patent this idea is a far cry from confirmation that this kind of protection system is in the works for [Sony's next game console,] the PlayStation 4."[7] Indeed, once the Disc ID controversy attracted mainstream media attention,[8] Sony began to equivocate[9] and ultimately decided against employing the technology in the PlayStation 4.[10]

---

2. *See generally* U.S. Patent Application No. 13/611,243, Publication No. 2013/0007892 (published Jan. 3, 2013) [hereinafter Sony Patent App.].

3. *Id.* at 2.

4. *Id.*

5. Peter Smith, *Sony Patent Suggests PS4 May Not Run Used Games*, ITWORLD (Jan. 4, 2013, 8:30 AM), http://www.itworld.com/personal-tech/333199/sony-patent-suggests-ps4-may-not-run-used-video-games.

6. Tim Cushing, *Sony Patent Application Takes on Used Game Sales, Piracy with Embedded RFID Chips in Game Discs*, TECHDIRT (Jan. 4, 2013, 4:28 AM), http://www.techdirt.com/articles/20130103/11044521571/sony-patent-application-takes-used-game-sales-piracy-with-embedded-rfid-chips-game-discs.shtml.

7. Kyle Orland, *Examining Sony's Internet-Free Method for Blocking Used Game Sales*, ARS TECHNICA (Jan. 3, 2013, 9:55 AM), http://arstechnica.com/gaming/2013/01/examining-sonys-internet-free-method-for-blocking-used-game-sales.

8. *See, e.g.*, Jared Newman, *The Death of Used Games Is a Rumor That Won't Die*, TIME (Feb. 6, 2013), http://techland.time.com/2013/02/06/the-death-of-used-games-is-a-rumor-that-wont-die; Chad Sapieha, *Sony's Proposed Patent Has Power to Kill Used Game Sales*, FINANCIAL POST (Jan. 13, 2013, 11:26 AM), http://business.financialpost.com/2013/01/15/sonys-proposed-patent-has-power-to-kill-used-game-sales.

9. *See* Chris Kohler, *What's Up with PlayStation 4 and Used Games?*, WIRED (Feb. 26, 2013, 2:55 PM), http://www.wired.com/gamelife/2013/02/playstation-4-used-games; Kyle Orland, *Sony UK Exec: PS4 Used Game Question "Isn't Clarified Just Yet,"* ARS TECHNICA (Feb. 26, 2013, 1:30 PM), http://arstechnica.com/gaming/2013/02/sony-uk-exec-ps4-used-game-

Video game companies have been cool to secondhand game sales for some time now. The chain retailer GameStop makes billions of dollars each year in used game sales[11] and has attracted harsh criticism from the industry.[12] Because GameStop offers pre-owned copies of popular games at discounted prices within months or even weeks of release, game companies argue that secondhand sales "deprive publishers of revenue, with the proceeds of the sale lining only the retailers' pockets."[13] Others in the industry have directed their ire at the gamers themselves. When game maker THQ introduced a system that allowed only original purchasers to access certain content in popular games, an executive told the press that he did not sympathize with used game buyers because "when the game's bought used we get cheated."[14] Other developers have pursued similar strategies to discourage secondhand game purchases.[15] GameStop insists that its used games business is in fact good for the video game makers' bottom lines,[16] a claim that enjoys some empirical support.[17] Nonetheless, many game developers remain unconvinced. Predictably, GameStop's stock took a hit when Sony's patent application became public[18] and dropped again when rumors surfaced that

question-isnt-clarified-just-yet/; Steven Ruygrok, *PlayStation 4 May Be Banning Used Games After All, Like Xbox 720*, EXAMINER.COM (Feb. 28, 2013), http://www.examiner.com/article/playstation-4-may-be-banning-used-games-after-all-like-xbox-720.

10. *See* Richard George, *E3 2013: PlayStation 4 Supports Used Games, Doesn't Require Online Connection*, IGN (June 10, 2013), http://www.ign.com/articles/2013/06/11/e3-2013-playstation-4-supports-used-games-dont-require-online-connection.

11. *See* Yukari Iwatani Kane & Miguel Bustillo, *Used Games Score Big for GameStop*, WALL ST. J., Jan. 21, 2009, http://online.wsj.com/article/SB123249378212700025.html. GameStop also sells new games.

12. *See Developers Respond to GameStop's Used Games Market Comments*, GAMEPOLITICS.COM (Aug. 23, 2011), http://gamepolitics.com/2011/08/23/developers-respond-gamestop039s-used-games-market-comments (compiling criticisms from developers at several companies).

13. Nathan Brown, *GameStop: Pre-Owned Benefits Publishers*, EDGE (Aug. 22, 2011, 1:22 PM), http://www.edge-online.com/news/gamestop-pre-owned-benefits-publishers.

14. Tom Pakinkis, *Pre-owned 'Cheats Developers'—THQ*, CVG (Aug. 23, 2010, 10:51 AM), http://www.computerandvideogames.com/261330/pre-owned-cheats-developers-thq.

15. *See id.*

16. Dan Pearson, *GameStop: Used Sales Benefit the Whole Industry*, GAMESINDUSTRY (Aug. 23, 2011, 9:24 AM), http://www.gamesindustry.biz/articles/2011-08-23-gamestop-used-sales-benefit-the-whole-industry.

17. *See, e.g.*, Nicole F. Velasquez et al., *The Impact of a Secondary Market on Video Game Purchase Intentions*, 16 REV. BUS. INFO. SYS. 103, 109 (2012) (finding that "the accessibility . . . of a secondary market is important to consumers' purchase behaviors in the primary market," and suggesting that "game developers and publishers would be wise to reassess their push to criminalize the resale of video games").

18. *GameStop Shares Fall on Sony Patent Application*, YAHOO! NEWS (Jan. 3, 2013, 3:10 PM), http://news.yahoo.com/gamestop-shares-fall-sony-patent-201046439.html.

Microsoft would also employ anti-used game technology in its next-generation Xbox One console.[19]

The Disc ID patent application is the content industry's latest volley in a decades-long fight over copyright's first sale doctrine. The first sale doctrine provides that the lawful owner of a copy of a copyrighted work has the right to sell, lease, lend, or otherwise transfer that particular copy.[20] With the rise of digital technology and personal computing in the 1980s, copyright-dependent industries grew increasingly anxious over the security of their products. New technology made it dramatically easier and cheaper to copy and distribute movies, computer programs, and other media. The first sale doctrine proved especially troublesome, since it propped up secondhand markets and a rental industry that companies in Hollywood and Silicon Valley alike considered threatening. Software makers were rightly concerned that consumers would rent computer programs for a matter of dollars, only to create their own illicit copies at home before returning the originals to the rental store. Not only did these industries successfully lobby Congress for statutory first sale exceptions, but software companies also adopted widespread licensing practices that precluded first sales from taking place to begin with.[21]

Technologies like Disc ID, which "tether" particular copies of video game discs, DVDs, or other media to a single user's device,[22] are simply new tools for accomplishing an old goal: cabining the first sale doctrine. The question many industry analysts have posed is whether these technologies, if adopted, would be legal.[23] This Article seeks to answer that question. The legality of tethering technologies hinges on two statutes: the Digital Millennium Copyright Act ("DMCA") and the Sherman Antitrust Act

19. Ben Lamoreux, *GameStop Stock Plunges 10% Following Recent Xbox Rumors*, GENGAME (Feb. 6, 2013), http://gengame.net/2013/02/gamestop-stock-plunges-10-following-recent-xbox-rumors. Microsoft did not end up incorporating this technology in Xbox One. Ben Gilbert, *Microsoft Reverses Xbox One DRM Policy, Kills Required Online Check-In and Used Game Complications (Updated)*, ENGADGET (June 19, 2013, 4:33 PM), http://www.engadget.com/2013/06/19/xbox-one-drm-used-games-reversal/.

20. 17 U.S.C. § 109(a) (2012).

21. *See infra* Part III.

22. Although Disc ID's immediate application is to video games, the patent application states that the technology is "similarly applicable to various kinds of electronic content such as an office suite, images, and music content." Sony Patent App., *supra* note 2, at 2.

23. *See, e.g.*, Cushing, *supra* note 6 ("[C]reating discs that are 'locked' to a certain system would seem to violate the right of first sale."); Orland, *supra* note 7 ("While this kind of resale-blocking technology would seemingly run afoul of the first sale doctrine codified into US law, legal experts seem unsure about whether that doctrine would be enough to overcome the end-user license agreements common to video game sales.").

("Sherman Act"). The DMCA protects "technological protection measures," including devices like Disc ID, that control consumers' access to copyrighted material.[24] Circumventing these protection measures to gain access to music, video, or other content is ordinarily prohibited.[25] While some courts would likely uphold tethering technologies against circumvention, others have interpreted the DMCA narrowly, allowing circumvention where it does not clearly lead to copyright infringement.[26] Since accessing the content stored on a video game disc or similar media would not, in itself, constitute infringement, consumers might be able to circumvent tethering technologies without violating the DMCA. Given the split in legal authority, however, it is impossible to say with any certainty how a court might rule on this issue without knowing the jurisdiction in which the court sits.

Moreover, technology that affects first sale rights may run afoul of the Sherman Act, which prohibits monopolization through predatory or anticompetitive means.[27] Courts have long recognized the first sale doctrine's antitrust significance and the importance of secondhand markets.[28] While judges are rightfully cautious about standing in the way of technological innovation, it is well established that monopolists may not use technology to suppress competition.[29] An antitrust challenge to a tethering technology could therefore be plausible. As a prerequisite to monopolization liability, however, an antitrust defendant must possess monopoly power in the relevant market.[30] Depending on industry conditions, it could be difficult if not impossible to show that any relevant company enjoys a monopoly. In these circumstances, the preservation of first sale rights may ultimately depend on consumers choosing not to patronize companies that engage in business practices to which they object.

This Article proceeds as follows. Part II introduces the first sale doctrine and explains its application to software and video games. Part III then describes the content industry's past and ongoing efforts to restrict first sale through lobbying, licensing, and litigation. Part IV analyzes tethering

24. *See infra* Section III.B.

25. 17 U.S.C. § 1201(a)(1)(A) (2012).

26. *See infra* Section IV.C.

27. 15 U.S.C. § 2 (2012).

28. *See* United States v. Aluminum Co. of Am. (*Alcoa*), 148 F.2d 416, 425 (2d Cir. 1945); United States v. United Shoe Mach. Corp., 110 F. Supp. 295, 350 (D. Mass. 1953), *aff'd per curiam*, 347 U.S. 521 (1954).

29. *See, e.g.*, United States v. Microsoft Corp., 253 F.3d 34, 64–66 (D.C. Cir. 2001) (en banc).

30. Verizon Commc'ns Inc. v. Law Offices of Curtis V. Trinko, LLP, 540 U.S. 398, 407 (2004).

technology under the DMCA's anticircumvention provisions and concludes that, at least in some jurisdictions, the law might permit gamers to circumvent tethering systems without incurring liability. Part V then discusses whether technologically restricting first sale rights would support a monopolization claim under the Sherman Act, using the video game industry as an example. While a claim would potentially be plausible, the difficulty of defining the relevant market so as to establish that any video game maker possesses monopoly power could be prohibitive. Part VI then offers concluding remarks.

II. A LINK TO THE PAST: OLD DOCTRINE, NEW TECHNOLOGY

A. THE COPYRIGHT FIRST SALE DOCTRINE

The first sale doctrine permits the lawful owner of a copy of a copyrighted work to freely resell that copy without restriction.[31] The Supreme Court first recognized the doctrine over a century ago in *Bobbs-Merrill Co. v. Straus*.[32] The plaintiffs in *Bobbs-Merrill* sued to enforce a notice printed in each copy of a novel called *The Castaway*, which read: "The price of this book at retail is $1 net. No dealer is licensed to sell it at a less price, and a sale at a less price will be treated as an infringement of the copyright."[33] The defendant booksellers had full knowledge of this restriction but nonetheless elected to sell copies of *The Castaway* for eighty-nine cents.[34] The Court held that, although the Copyright Act reserved to the copyright holder "the sole right to vend" copies of his or her work, the Act did not "create the right to impose, by notice, . . . a limitation at which the book shall be sold at retail by future purchasers, with whom there is no privity of contract."[35] By producing and selling copies of *The Castaway* "in quantities and at a price satisfactory to [them]," the plaintiffs had "exercised the right to vend."[36] The Court refused to read into the statute a right to control all *future* sales.[37]

Congress has codified the first sale doctrine in § 109(a) of the Copyright Act, which states that "the owner of a particular copy or phonorecord lawfully made under this title, or any person authorized by such owner, is

31. *See* 17 U.S.C. § 109(a) (2012); *see also* Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351, 1355–56 (2013); Apple Inc. v. Psystar Corp., 658 F.3d 1150, 1155 (9th Cir. 2011).
32. Bobbs-Merrill Co. v. Straus, 210 U.S. 339 (1908).
33. *Id.* at 341.
34. *Id.* at 342.
35. *Id.* at 350.
36. *Id.* at 351.
37. *Id.*

entitled, without the authority of the copyright owner, to sell or otherwise dispose of the possession of that copy or phonorecord."[38] Because the first sale doctrine is only triggered by an actual sale or other transfer of title,[39] only bona fide owners may avail themselves of its protection.[40]

The first sale doctrine lies at the intersection between intellectual property and tangible property.[41] According to Professors Melville and David Nimmer, once a copyright owner has consented to the sale or distribution of copies of a work, "continued control over the distribution of copies" would only serve as "a device for controlling the disposition of the tangible personal property that embodies the work."[42] At this point, "the policy favoring a copyright monopoly . . . gives way to the policy opposing restraints of trade and restraints on alienation."[43] Without the first sale rule, copyright holders could extend the copyright monopoly "so far as to . . . control . . . the disposition of lawfully obtained tangible personal property."[44] In this sense, the first sale doctrine can be traced to the ancient common law rule disfavoring limitations on the transfer of private property.[45]

There is also a sound economic rationale for the first sale doctrine. For one, the doctrine avoids the "prohibitive transaction costs" of granting copyright holders unfettered control over successive sales of copies of their works.[46] In the doctrine's absence, successive purchasers or possessors of a copy would have to "negotiate with the [copyright] owner each time they

38. 17 U.S.C. § 109(a) (2012).

39. Adobe Sys. Inc. v. One Stop Micro, Inc., 84 F. Supp. 2d 1086, 1089 (N.D. Cal. 2000).

40. Adobe Sys. Inc. v. Christenson, 891 F. Supp. 2d 1194, 1203 (D. Nev. 2012).

41. *See* Brilliance Audio, Inc. v. Haights Cross Commc'ns, Inc., 474 F.3d 365, 373 (6th Cir. 2007). The court in *Brilliance Audio* stated:

> This bargain, first developed in the common law, and later codified in the first sale doctrine, provides that once a copyright owner consents to release a copy of a work to an individual (by sale, gift, or otherwise), the copyright owner relinquishes all rights to that particular copy. . . . The first sale doctrine ensures that the copyright monopoly does not intrude on the personal property rights of the individual owner, given that the law generally disfavors restraints of trade and restraints on alienation.

Id. (footnote and citations omitted).

42. 2-8 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 8.12[A] (Matthew Bender Rev. Ed. 2013).

43. *Id.*

44. Allison v. Vintage Sports Plaques, 136 F.3d 1443, 1447–48 (11th Cir. 1998).

45. *See* Sebastian Int'l, Inc. v. Consumer Contacts (PTY) Ltd., 847 F.2d 1093, 1096 (3d Cir. 1988) (citing H.R. REP. NO. 98-987 (1984)).

46. *See* Pearson Educ., Inc. v. Liu, 656 F. Supp. 2d 407, 410 (S.D.N.Y. 2009).

contemplate a further sale or other transfer of the copy."[47] Moreover, where a "first sale" has taken place, "it may fairly be said that the copyright proprietor has received his reward" for the use of his or her work, and no further control over the copy's distribution is warranted.[48] Copyright exists not to enrich authors and artists but to benefit the public.[49] With the consummation of a sale, the copyright monopoly has served its constitutionally mandated purpose—to incentivize the creation and dissemination of new works to the public—and any justification for continued control over the work's distribution evaporates.[50] First sale therefore balances copyright holders' economic incentives to create and publish new works against the public's interest in accessing those works.[51]

A number of public institutions and industries owe their existence to the first sale doctrine. For example, according to the American Library Association, "first sale is what allows libraries to do what we do—lend books and materials to our patrons, the public."[52] American public libraries collectively house nearly a billion media items, the vast majority of which are printed books. Academic library collections account for over 1.3 billion items in total, and public school media centers contain a billion more.[53] Americans collectively borrow these materials 4.4 billion times per year.[54] Not only do

47. *Id.* (quoting 2 PAUL GOLDSTEIN, GOLDSTEIN ON COPYRIGHT § 7.6.1 (2005)).

48. Burke & Van Heusen, Inc. v. Arrow Drug, Inc., 233 F. Supp. 881, 884 (E.D. Pa. 1964).

49. *See* U.S. CONST. art. I, § 8, cl. 8 (empowering Congress "[t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries"); Fogerty v. Fantasy, Inc., 510 U.S. 517, 527 (1994) ("copyright law ultimately serves the purpose of enriching the general public through access to creative works"); United States v. Paramount Pictures, Inc., 334 U.S. 131, 158 (1948) ("The copyright law, like the patent statutes, makes reward to the owner a secondary consideration.").

50. *See* Eldred v. Ashcroft, 537 U.S. 186, 241 (2003) (Stevens, J., dissenting) ("[A]s our cases repeatedly and consistently emphasize, ultimate public access is the overriding purpose of the constitutional provision."); Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 429 (1984) ("[Copyright] is intended to motivate the creative activity of authors and inventors by the provision of a special reward, and to allow the public access to the products of their genius after the limited period of exclusive control has expired.").

51. *See* Twentieth Century Music Corp. v. Aiken, 422 U.S. 151, 156 (1975) ("The limited scope of the copyright holder's statutory monopoly . . . reflects a balance of competing claims upon the public interest: Creative work is to be encouraged and rewarded, but private motivation must ultimately serve the cause of promoting broad public availability of literature, music, and the other arts.").

52. *First Sale Doctrine and* Kirtsaeng v. Wiley & Sons, Inc., AMERICAN LIBRARY ASSOCIATION, http://www.ala.org/advocacy/copyright/firstsale (last visited Nov. 3, 2013).

53. Brief for Am. Library Ass'n et al. as Amici Curiae Supporting Petitioner at 10–11, Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351 (2013) (No. 11-697).

54. *Id.*

libraries make for a more literate, cultured, and informed citizenry, but they also promote local economic development.[55] Post-sale restrictions on the disposition of copyrighted materials would prevent libraries from providing this crucial public service.

Likewise, businesses that profit from the buying, selling, or renting of secondhand media could not exist without the first sale doctrine. Movie and video game rental businesses, such as Redbox, GameFly, and Netflix, purchase inventory and then rent items to customers for a fee.[56] Although technological innovations like on-demand streaming video have challenged the traditional rental business model,[57] the movie and video game rental industries still generate billions of dollars per year.[58] Markets for used goods can also be substantial. GameStop, for example, reaped $2.6 billion in revenues from used game sales in 2011 alone.[59] These secondary markets are a boon for consumers, who benefit from increased price competition.[60] And contrary to conventional wisdom, which maintains that used goods supplant demand for new goods, secondhand markets may often benefit suppliers of new goods as well.[61] These markets all depend upon the first sale doctrine.

---

55. *See generally* ROBERT H. GLASS ET AL., THE ROLE OF PUBLIC LIBRARIES IN LOCAL ECONOMIC DEVELOPMENT (2000), *available at* http://www.ipsr.ku.edu/resrep/pdf/m260.pdf.

56. *See* Julie H. Mortimer, *Vertical Contracts in the Video Rental Industry*, 75 REV. ECON. STUD. 165, 165–69 (2008). The rental market represents a significant revenue stream for content producers (i.e., copyright owners), especially under profit-sharing schemes that have become widespread in the industry. *See id.* at 165–69.

57. *See* Courtney Subramanian, *The End of an Era: Blockbuster Ceases Rentals in Advance of Closing*, TIME, Nov. 9, 2013, http://business.time.com/2013/11/09/the-end-of-an-era-blockbuster-ceases-rentals-in-advance-of-closing ("Parent company Dish Network . . . attributed [Blockbuster's] decline to increased digital streaming . . . .").

58. IBISWORLD, IBISWORLD INDUSTRY REPORT 53223: DVD, GAME & VIDEO RENTAL IN THE US 5 (May 2011), *available at* http://www.aba.com/aba/documents/CommercialInsights/IBISWorld_DVD.pdf.

59. Alexander Sliwinski, *GameStop: 70% of Trade-in Credit Spent on New Games*, JOYSTIQ (Aug. 6, 2012, 10:15 PM), http://www.joystiq.com/2012/08/06/gamestop-70-of-trade-in-credit-spent-on-new-games.

60. *See* Anindya Ghose et al., *Durable Goods Competition in Secondary Electronic Markets*, *in* PROCEEDINGS OF THE TWENTY-FOURTH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS 341, 346 (2003), *available at* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.231.3720. Ghose states:

> The presence of an active used-goods market creates competition for new goods because the option of buying used goods is now incentive compatible for some consumers who would have bought new goods before. This enhanced competition forces suppliers to decrease the new-good prices in order to remain competitive with used goods.

Id.

61. *See id.* at 350 ("[W]e show that, contrary to popular perceptions, the presence of a used-goods market is beneficial for suppliers too, under a relatively wide range of

The significance of the first sale doctrine can hardly be overstated. This "longstanding and fundamental . . . limitation on the public distribution right"[62] promotes commerce, safeguards competition, and props up entire industries.[63] Although first enunciated over a century ago, the doctrine remains central to our understanding of copyright and property.

B. LICENSING ARRANGEMENTS

Notwithstanding the first sale doctrine, copyright owners do have an effective means for controlling copies of their work in users' possession. While "the first sale inquiry examines *ownership* of the tangible property in which the copyrighted work has been embodied,"[64] a copyright owner may lawfully transfer *possession* of a particular copy without forfeiting title to it.[65] The law permits "copyright owners [to] create licensing arrangements so that users acquire only a license to use the particular copy . . . and do not acquire title that permits further transfer or sale of that copy without the permission of the copyright owner."[66] A license functions as a contract,[67] dictating the terms by which the licensee may possess and use the copy of the work in question. If the licensee violates the terms of the license in a manner that implicates copyright, the copyright owner may sue for both breach of contract and infringement[68]—the latter of which carries significant statutory damages.[69] The notion that copyright owners may contract around the first

conditions."); Hal R. Varian, *Reading Between the Lines of Used Book Sales*, N.Y. TIMES, July 28, 2005, http://www.nytimes.com/2005/07/28/technology/28scene.html ("[T]he used [book] market does not have a big impact in terms of lost sales in the new [book] market.").

62. John A. Rothchild, *The Incredible Shrinking First-Sale Rule: Are Software Resale Limits Lawful?*, 57 RUTGERS L. REV. 1, 75 (2004).

63. *See* Andrew Degner, Wise-*ing Up About the First Sale Doctrine: A Look at the Ninth Circuit's Approach in* Vernor v. Autodesk, 35 COLUM. J.L. & ARTS 581, 602 (2012) ("[I]t is important to recall that the first sale doctrine was designed to protect the rights of individuals to their property against unreasonable restraints upon alienation under the guise of the copyright law.").

64. NIMMER & NIMMER, *supra* note 42, § 8.12[B][1][d][i] (emphasis added).

65. *See* United States v. Drebin, 557 F.2d 1316, 1330 (9th Cir. 1977) ("[T]ransfer of title [is] the *sine qua non* of the first sale doctrine."); UMG Recordings, Inc. v. Augusto, 628 F.3d 1175, 1180 (9th Cir. 2011) ("We have recognized . . . that not every transfer of possession of a copy transfers title.").

66. *UMG Recordings*, 628 F.3d at 1180.

67. *See* Flava Works, Inc. v. Gunter, 689 F.3d 754, 755 (7th Cir. 2012) (Posner, J.) ("[A] copyright license is just a type of contract").

68. *See* MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928, 940 (9th Cir. 2010) ("To recover for copyright infringement based on breach of a license agreement, (1) the copying must exceed the scope of the defendant's license and (2) the copyright owner's complaint must be grounded in an exclusive right of copyright . . . .").

69. *See* 17 U.S.C. § 504(c)–(d) (2012).

sale doctrine by way of licensing is as old as the doctrine itself. In *Bobbs-Merrill*, the Supreme Court acknowledged that its first sale holding did not reach license agreements controlling subsequent sales of a book.[70]

C. APPLICATION TO COMPUTER SOFTWARE AND VIDEO GAMES

Computer programs raise special challenges for copyright law. Under the 1980 amendments to the Copyright Act,[71] computer programs became eligible for federal copyright protection.[72] With the advent of personal computing in the early 1980s, software producers became increasingly concerned with unauthorized copying.[73] They were especially worried about software rental, which the Copyright Act permitted (and which, in fact, came to pass in the rental industry).[74] Due to the ease of copying software, consumers could, in lieu of purchasing a program, "simply rent a copy of the program[] and duplicate it."[75] While unauthorized duplication ordinarily constitutes infringement, it would have been impractical and uneconomical for copyright owners to track down and sue individual consumers.[76] Going after the rental stores made more sense, but the first sale doctrine immunized software rental stores from liability.[77] Software producers consequently decided to begin licensing, rather than selling, their computer programs.[78] This allowed software owners to "retain ownership and . . . limit the user's rights to copy, transfer or modify the software, thereby making the first sale doctrine inapplicable."[79] Licensing continues to be the norm in software distribution.[80]

At the same time, the computer industry lobbied Congress for an exception to the first sale doctrine.[81] Recognizing that first sale permitted

---

70. Bobbs-Merrill Co. v. Straus, 210 U.S. 339, 350 (1908). By its own terms, § 109(a) of the Copyright Act only applies to "the *owner* of a particular copy or phonorecord made under this title, or any person authorized by such owner." 17 U.S.C. § 109(a) (emphasis added).

71. *Cf.* Judith Klerman Smith, *The Computer Software Rental Act: Amending the "First Sale Doctrine" to Protect Computer Software Copyright*, 20 LOY. L.A. L. REV. 1613, 1620–21 (1987).

72. *Id.*; *see* 17 U.S.C. § 101 (defining "computer program").

73. Smith, *supra* note 71, at 1626–28.

74. *See id.* at 1626, 1628.

75. Step-Saver Data Sys., Inc. v. Wyse Tech., 939 F.2d 91, 96 n.7 (3d Cir. 1991).

76. *See id.* ("This copying by the individual consumers would presumably infringe the copyright, but usually it would be far too expensive for the copyright holder to identify and sue each individual copier.").

77. *Id.*; Smith, *supra* note 71, at 1624–25.

78. *See* Smith, *supra* note 71, at 1630–31.

79. *Id.*

80. Apple Inc. v. Psystar Corp., 658 F.3d 1150, 1155 (9th Cir. 2011).

81. *See* Jessica Litman, *Revising Copyright Law for the Information Age*, 75 OR. L. REV. 19, 23, 35–36 (1996).

"rental firms [to] legally dispose of legitimately purchased copies of computer programs," and that rented software could be copied with little more than the push of a button,[82] Congress enacted the Computer Software Rental Amendments Act of 1990.[83] While not disturbing "the ability of copyright owners and users to enter into license agreements regarding the use of computer programs,"[84] the Act amended § 109 to prohibit "any person in possession of a particular copy of a computer program" from renting, leasing, or lending the copy for commercial purposes.[85] Between licensing and legislation, the software industry had successfully made an end run around the first sale doctrine by the early 1990s.

Video games are of course computer programs,[86] but they have charted a different course from other types of software covered by the Copyright Act. Unlike personal computing software, video games designed for home consoles are ordinarily difficult to copy.[87] In fact, Nintendo designed its groundbreaking NES console to play game cartridges rather than floppy disks precisely because cartridges were harder to copy.[88] Video games are, in this respect, more akin to videocassettes and DVDs than to conventional software: popping a video game cartridge or disc into the console allows the gamer to access its contents, but it does not create a permanent copy on the device (i.e., the game is not "installed" or otherwise stored on a local hard drive). For these reasons, video game makers have been somewhat less guarded,[89] and they ordinarily sell rather than license their products.[90]

---

82. S. REP. NO. 101-265, at 3 (1990).

83. Computer Software Rental Amendments Act of 1990, Pub. L. No. 101-650, 104 Stat. 5089.

84. S. REP. NO. 101-265, at 5.

85. 17 U.S.C. § 109(b)(1)(A) (2012).

86. *See* Erik Bauman, *The Nexus Analysis: License Enforcement in the Wake of* MDY v. Blizzard, 27 BERKELEY TECH. L.J. 503, 525 (2012).

87. *See* S. REP. NO. 101-265, at 5. In 1990, video games manufactured for home consoles were stored on difficult-to-copy microchips encased in cartridges. *Id.* Subsequent generations of consoles, such as the Sony PlayStation, utilized easier-to-copy compact discs rather than cartridges. However, the consoles themselves were incapable of running pirated discs without modifications to their hardware. *See* Brian Fitzgerald, *The PlayStation Mod Chip: A Technological Guarantee of the Digital Consumer's Liberty or Copyright Menace/Circumvention Device?*, 10 MEDIA & ARTS L. REV. 85, 86 (2005).

88. *See* DAVID SHEFF, GAME OVER: HOW NINTENDO ZAPPED AN AMERICAN INDUSTRY, CAPTURED YOUR DOLLARS, AND ENSLAVED YOUR CHILDREN 33 (1993) ("The system would play games on cartridges, not disks. Floppy disks were threatening to computerphobes and, more important, they were copiable.").

89. This is not at all to say that video game makers have been *un*concerned about piracy. *Pong*, one of the earliest video games, was "copied with abandon." SHEFF, *supra* note 88, at 140. Of the 100,000 *Pong* arcade games sold in 1974, only one-tenth were produced by Atari. *Id.*

Congress determined that there was no need to extend the 1990 Amendments Act to video games[91] and, accordingly, excluded video games manufactured for use on home consoles from the Act's coverage.[92] The first sale doctrine therefore continues to apply with full force to video game sales.

In sum, copyright law has relied upon the first sale doctrine for over a century to mediate the tension between intellectual property and personal property. The doctrine promotes efficiency and competition, and serves as the legal foundation for many public institutions and private enterprises.[93] However, the doctrine is not absolute. Copyright owners may contract around it by way of licensing, which allows them to maintain title in distributed media and dictate the terms of use. While computer programs enjoy additional statutory exemptions from first sale, console-based video games remain fully subject to the first sale doctrine, and licensing has not yet become commonplace in the video game industry. This is not to say, however, that first sale is without threat in the gaming world.

III. ALL YOUR BASE ARE BELONG TO US: FIRST SALE UNDER FIRE[94]

First sale's standing in copyright law has probably never been more precarious than it is today. While many have prematurely reported the doctrine's demise—a law review article published almost thirty years ago

90. *See* Adobe Sys., Inc. v. One Stop Micro, Inc., 84 F. Supp. 2d 1086, 1091 (N.D. Cal. 2000) ("[W]ith the exception of video games, software is licensed rather than sold." (quoting Raymond Nimmer)).

91. *See* S. REP. NO. 101-265, at 5 ("At the present time, the rationale for prohibiting the rental of software available in other formats does not apply to video game cartridges.").

92. *See* 17 U.S.C. § 109(b)(1)(B)(ii) (2012) ("This subsection does not apply to . . . a computer program embodied in or used in conjunction with a limited purpose computer that is designed for playing video games and may be designed for other purposes.").

93. *See supra* notes 52–63 and accompanying text.

94. Explaining a cultural reference may defeat the purpose of making it to begin with, but this one is so obscure that it probably warrants comment. The phrase "All Your Base Are Belong to Us" is an awkward translation of "all your bases are under our control" that originally appeared in the European version of *Zero Wing*, a video game released for Sega's Mega Drive console in 1991. The catchphrase "swept across the internet at the dawn of the 21st century," and was recently referenced in Disney's video game-themed animated film *Wreck-It Ralph*. *See All Your Base Are Belong to Us*, KNOW YOUR MEME, http://knowyourmeme.com/memes/all-your-base-are-belong-to-us (last visited Feb. 27, 2013); Luke McKinney, *Leeroy Jenkins Lives! The 6 Best Video Game References in "Wreck-It Ralph,"* MOVIELINE (Nov. 12, 2012, 5:30 PM), http://movieline.com/2012/11/12/wreck-it-ralph-best-video-game-references-aerith-kano-konami-code.

pronounced it "the defense that never was"[95]—the doctrine has indeed been under siege for decades. Since the advent of consumer copying technologies, such as Xerox machines and VCRs, the content industry[96] has incessantly fought first sale through litigation, lobbying, and licensing.[97] Digital technology has now brought this fight to a head.[98]

This Section focuses on two practices the content industry has widely adopted to limit first sale rights: first, the ubiquitous "shrinkwrap" or "clickthrough" license, which precludes transfer of ownership and affords copyright holders far-reaching control over media in their customers' hands; and second, technologies that restrict consumers' access to and usage of copyrighted media, even following a legitimate first sale.

A. THE RISE OF THE END USER LICENSE AGREEMENT

As Professor Raymond Nimmer writes, contract law plays an "overriding role" in the law of first sale: "Contractual terms determine the applicability of first sale doctrine as a matter of copyright law . . . . The presence or absence of a first sale ownership privilege thus depends on the terms of a contract."[99] While freedom of contract has been heralded as indispensable to a "free enterprise system based on an unheard of division of labor," the "development of large scale enterprise with its mass production and mass distribution made a new type of contract inevitable—the standardized mass contract."[100] The concept of standardized contracts, or contracts of adhesion, entered the American legal lexicon in the second decade of the twentieth century.[101] Far removed from the idealized notion of two parties "meet[ing] each other on a footing of social and approximate economic equality" to

95. Richard Colby, *First Sale Doctrine—The Defense That Never Was?*, 32 J. COPYRIGHT SOC'Y U.S.A. 77, 77 (1984).

96. The "content industry" broadly encompasses companies that own and produce mass media, such as music and movies. *See generally* Mark A. Lemley, *Is the Sky Falling on the Content Industries?*, 9 J. ON TELECOMM. & HIGH TECH. L. 125 (2011).

97. For example, after the movie industry failed to kill home video players through litigation, see Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984), it lobbied Congress to legislatively prohibit the rental, lease, and lending of videocassettes and other audiovisual works. *See* Ralph S. Brown, Jr., *Eligibility for Copyright Protection: A Search for Principled Standards*, 70 MINN. L. REV. 579, 584 & n.30 (1985).

98. *See* R. Anthony Reese, *The First Sale Doctrine in the Era of Digital Networks*, 44 B.C. L. REV. 577, 581 (2003).

99. Raymond T. Nimmer, *Copyright First Sale and the Overriding Role of Contract*, 51 SANTA CLARA L. REV. 1311, 1345–46 (2011).

100. Friedrich Kessler, *Contracts of Adhesion—Some Thoughts About Freedom of Contract*, 43 COLUM. L. REV. 629, 629, 631 (1943).

101. *See* Edwin W. Patterson, *The Delivery of a Life Insurance Policy*, 33 HARV. L. REV. 198, 222 (1919).

achieve a unity of minds through self-interested bargaining,[102] the terms of a contract of adhesion are entirely dictated by one party to another—take it or leave it.[103]

Courts generally uphold contracts of adhesion, although the equitable doctrine of unconscionability may preclude enforcement of particularly egregious contracts.[104] It is of no consequence that these contracts are seldom read, and even less frequently understood; so long as the purchaser "assents" to the terms, the contract is valid.[105] This is an extremely low bar. In *ProCD, Inc. v. Zeidenberg*, the defendant did not even *see* the contractual terms until after he had paid for the plaintiffs' product and received it in the mail.[106] Because the defendant used the product after having an opportunity to read the contract, however, Judge Easterbrook reasoned that he had assented to the contractual terms and had become bound by them.[107] There are limits to the doctrine of assent—a contract may be unenforceable, for instance, in the absence of a reasonably conspicuous notice that a party is binding itself.[108] But the nomenclature now applied to contracts of adhesion (e.g., "shrinkwrap," "clickthrough") suggests just how little courts require in the way of assent.

Copyright licenses are simply contracts,[109] and copyright owners routinely use contracts of adhesion to avoid the first sale doctrine. The practice first arose in the 1980s with the mass marketization of computer software.[110]

---

102. Kessler, *supra* note 100, at 630.

103. *See* Victoria v. Superior Court, 710 P.2d 833, 837 (Cal. 1985). The court in *Victoria* stated:

> The term "adhesion contract" refers to standardized contract forms offered to consumers of goods and services on essentially a "take it or leave it" basis without affording the consumer a realistic opportunity to bargain and under such conditions that the consumer cannot obtain the desired product or services except by acquiescing in the form contract.

Id. (quoting Wheeler v. St. Joseph Hosp., 133 Cal. Rptr. 775, 783 (Cal. Ct. App. 1976)).

104. *See, e.g.*, Bragg v. Linden Research, Inc., 487 F. Supp. 2d 593 (E.D. Pa. 2007); Comb v. PayPal, Inc., 218 F. Supp. 2d 1165 (N.D. Cal. 2002); Brower v. Gateway 2000, Inc., 676 N.Y.S.2d 569 (N.Y. App. Div. 1998).

105. *See* Specht v. Netscape Commc'ns Corp., 306 F.3d 17, 29 (2d Cir. 2002) ("Mutual manifestation of assent, whether by written or spoken word or by conduct, is the touchstone of contract.").

106. ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1450 (7th Cir. 1996) (Easterbrook, J.).

107. *Id.* at 1452–53.

108. *See Specht*, 306 F.3d at 35 ("Reasonably conspicuous notice of the existence of contract terms and unambiguous manifestation of assent to those terms by consumers are essential if electronic bargaining is to have integrity and credibility.").

109. Flava Works, Inc. v. Gunter, 689 F.3d 754, 755 (7th Cir. 2012) (Posner, J.).

110. *See* Richard H. Stern, *Shrink-Wrap Licenses of Mass Marketed Software: Enforceable Contracts or Whistling in the Dark?*, 11 RUTGERS COMPUTER & TECH. L.J. 51, 51 (1985)

Copies of computer programs literally had licensing agreements printed on their shrinkwrap packaging; the licenses were supposedly activated when a purchaser removed the wrapping.[111] This creative practice was understandable given the state of copyright law at the time. The scope of copyright protection in computer programs was uncertain, and as previously discussed, the first sale doctrine prevented copyright owners from controlling the software rental business, which facilitated widespread unauthorized copying.[112] Mass licensing enabled software producers to guard against gaps and uncertainty in copyright law.

Copyright protection in software is no longer a subject of serious debate, and the Copyright Act no longer permits software rental, but the proliferation of nonnegotiable licenses (commonly denominated "End User License Agreements," or "EULAs") has continued unabated. As a federal appellate court put it, "[s]oftware licensing agreements, rather than sales, have become ubiquitous in the software industry because they enable the licensor to control the use of the copyright material."[113] The capacity to make the first sale doctrine irrelevant (and, increasingly, obsolete) explains licensing's popularity.[114] The expanding list of names for EULAs reflects the myriad ways in which these contracts are created in today's world; so-called "clickwrap" and "browsewrap" licenses are now common.[115] While most commonly used in connection with software, standardized licenses can be applied to virtually any type of copyrighted media.

(" 'Shrink-wrap,' 'box-top,' or 'tear-open' licenses are a new form of quasi-consensual agreement spawned by the needs of our computer society, or more precisely by the economic needs of mass marketers of software.").

111. David Bender, *Software Protection: The 1985 Perspective*, 77 W. NEW ENG. L. REV. 405, 438–40 (1985). Software licensing began much earlier, in the 1960s, with the birth of the software industry. Glen O. Robinson, *Personal Property Servitudes*, 71 U. CHI. L. REV. 1449, 1473 (2004). In those early days, "a typical mainframe computer cost hundreds of thousands (and perhaps over a million) dollars. . . . [A]pplication software was all custom developed, and a typical computer program might cost tens of thousands (or even more than a hundred thousand) dollars." Bender, *supra*, at 438. Software transactions were therefore substantial, low in volume, and highly individualized to the parties involved. *Id.* License agreements were negotiated and drafted by lawyers. *Id.* In the 1980s, when consumers began purchasing software "over the counter like expensive jelly beans," nonnegotiable, unindividuated licenses became the norm. *Id.* at 438–39.

112. *Id.* at 439.

113. Apple Inc. v. Psystar Corp., 658 F.3d 1150, 1155 (9th Cir. 2011).

114. *See id.* at 1156 ("It is this distinction between sales and licenses that has caused the use of software licensing agreements to flourish and become the preferred form of software transactions.").

115. *See* Cheryl B. Preston & Eli W. McCann, *Unwrapping Shrinkwraps, Clickwraps, and Browsewraps: How the Law Went Wrong from Horse Traders to the Law of the Horse*, 26 BYU J. PUB. L. 1, 17–19 (2011).

The pervasiveness of EULAs means that copyright defendants asserting a first sale defense often have difficulty proving that a true sale—a transfer of ownership—ever took place to begin with. The case law is not uniform,[116] but the modern trend largely favors copyright owners over consumers, licenses over ownership.[117] The Ninth Circuit has been especially prolific in articulating how copyright owners can avoid the pitfall of actually selling goods to their customers.[118] Without sacrificing much in the way of nuance, one may describe the circuit's recent precedents as establishing the following proposition: by observing certain formalities, a copyright holder can retain ownership in virtually any copyrighted media it "sells," thereby avoiding the first sale doctrine entirely.[119] More specifically, the Ninth Circuit has held that the following factors determine whether a transfer constitutes a sale or a license: (1) whether the copyright owner specifies that a user is granted a license, (2) whether the copyright owner significantly restricts the user's ability to transfer the software, and (3) whether the copyright owner imposes notable use restrictions.[120] In other words, a transfer is a license and not a sale if the copyright owner says it is a license.[121]

---

116. *See* Terence Leong, Note, *When the Software We Buy is Not Actually Ours: An Analysis of* Vernor v. Autodesk *on the First Sale Doctrine and Essential Step Defense*, 10 NW. J. TECH. & INTELL. PROP. 239, 239 (2012).

117. *See* Jenny Lynn Sheridan, *Does the Rise of Property Rights Theory Defeat Copyright's First Sale Doctrine?*, 52 SANTA CLARA L. REV. 297, 371 (2012) (describing courts and Congress as "strong 'property rights' adherents," with "no patience for other values in copyright law other than maximum control over the [intellectual] property by the owner"). *But see* Krause v. Titleserv, Inc., 402 F.3d 119, 122–25 (2d Cir. 2005) (holding that purported "licensee" was in fact an owner, for purposes of 17 U.S.C. § 117(a) "essential step" doctrine, where it "paid substantial sums, to possess and use a copy indefinitely without material restriction, as well as to discard or destroy it at will").

118. *See, e.g.*, Apple Inc. v. Psystar Corp., 658 F.3d 1150 (9th Cir. 2011); UMG Recordings, Inc. v. Augusto, 628 F.3d 1175 (9th Cir. 2011); F.B.T. Prods., LLC v. Aftermath Records, 621 F.3d 958 (9th Cir. 2010); MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928 (9th Cir. 2010); Vernor v. Autodesk, Inc., 621 F.3d 1102 (9th Cir. 2010); Wall Data Inc. v. L.A. Cnty. Sheriff's Dep't, 447 F.3d 769 (9th Cir. 2006); Triad Sys. Corp. v. Se. Express Co., 64 F.3d 1330 (9th Cir. 1995); MAI Sys. Corp. v. Peak Computer, Inc., 991 F.2d 511 (9th Cir. 1993); S.O.S., Inc. v. Payday, Inc., 886 F.2d 1081 (9th Cir. 1989); United States v. Wise, 550 F.2d 1180 (9th Cir. 1977).

119. *See* Corynne McSherry, *"Magic Words" Trump User Rights: Ninth Circuit Ruling in* Vernor v. Autodesk, ELEC. FRONTIER FOUND. (Sept. 13, 2010), https://www.eff.org/deeplinks/2010/09/magic-words-trump-user-rights-ninth-circuit-ruling ("[Under Ninth Circuit precedent,] copyright's first sale doctrine . . . doesn't apply to software (and possibly DVDs, CDs and other 'licensed' content) as long as the vendor saddles the transfer with enough restrictions to transform what the buyer may think is sale into a mere license.").

120. *Vernor*, 621 F.3d at 1110–11.

121. *See* Sheridan, *supra* note 117, at 353 (copyright owners may avoid the first sale doctrine by "unilaterally plac[ing] reservations on title"). That said, the Ninth Circuit

Widespread licensing may have begun with computer software, but court decisions upholding the standard-form licenses contain no such limiting language.[122] There is nothing preventing copyright holders from applying EULAs to music, DVDs, or even books.[123] Given "how easy it is for a vendor to put a 'license' label on a mass-marketed product with copyrighted . . . components," Pamela Samuelson worries what the Ninth Circuit's precedents "mean for flea markets, bookstores, libraries, garage sales, and auction sites."[124] The prospect of having to trace "chain of title" before purchasing a book or record at a secondhand store may sound absurd, but it is no longer beyond the realm of legal possibility.[125] Courts' reverence for

recognizes that "the mere labeling of an arrangement as a license rather than a sale . . . [is] not by itself dispositive of the issue." *UMG Recordings*, 628 F.3d at 1180. In *UMG Recordings*, the Ninth Circuit held that a record company transferred ownership in unsolicited promotional CDs that it distributed freely by mail, despite having labeled the CDs with licensing language. *See id.* at 1182. The court distinguished software licensees "who order and pay to acquire copies" from passive recipients of unrequested CDs. *Id.* at 1183. Unlike software licensees, the CD recipients did not manifest assent to the contractual terms, and their assent could not be inferred from the mere failure to return the CDs (especially since the record company did not request the CDs' return). *Id.* at 1182–83. *UMG Recordings* therefore stands as a reminder that there must be *some* manifestation of assent on the licensee's part, however minimal. The copyright holder must also have a degree of knowledge or control over the disposition of media in order to be considered its owner. *Id.*

122. *Compare* F.B.T. Prods., LLC v. Aftermath Records, 621 F.3d 958, 965 (9th Cir. 2010) (holding that license exists where a copyright holder "transfers a copy of copyrighted material, retains title, limits the uses to which the material may be put, and is compensated periodically based on the transferee's exploitation of the material"), *and Vernor*, 621 F.3d at 1109 (holding that a court may consider whether a license existed, and "whether the copyright owner retained title to the copy, required its return or destruction, forbade its duplication, or required the transferee to maintain possession" to determine whether infringement occurred), *with* ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1149 (7th Cir. 1996) ("Shrinkwrap licenses are enforceable unless their terms are objectionable on grounds applicable to contracts in general."). *See generally* JASON MAZZONE, COPYFRAUD AND OTHER ABUSES OF INTELLECTUAL PROPERTY 132–33 (2011).

123. *See* Corynne McSherry, *2012 in Review: First Sale Under Siege—If You Bought It, You Should Own It*, ELEC. FRONTIER FOUND. (Dec. 23, 2012), https://www.eff.org/deeplinks/2012/12/first-sale-under-siege-if-you-bought-it-you-should-own-it. McSherry wrote:

> Not only does big content deny that first sale doctrine applies to digital goods, but they are also trying to undermine the first sale rights we do have by forcing users to license items they would rather buy. The copyright industry wants you to "license" all your music, your movies, your games—and lose your rights to sell them or modify them as you see fit.

Id.

124. Pamela Samuelson, *Do You Own the Software You Buy?*, COMMC'NS OF THE ACM, Mar. 2011, at 26, 28.

125. *See* Microsoft Corp. v. Harmony Computers & Elecs., Inc., 846 F. Supp. 208, 212 (E.D.N.Y. 1994) ("Although a sale of a copyrighted work by a party authorized by the

contracts of adhesion and deference to copyright owners threaten to eviscerate the first sale doctrine altogether—upending copyright law's balance between users and rights holders, undermining secondary markets and the industries that depend on them, and opening the door to anticompetitive conduct.[126]

B. DIGITAL RIGHTS MANAGEMENT ("DRM") AND THE DIGITAL MILLENNIUM COPYRIGHT ACT ("DMCA")

Copyright and contract afford only so much control over how consumers use the music, movies, video games, and other media in their possession. For one, the effectiveness of these legal strictures depends upon the degree to which they are followed and enforced. Piracy can be difficult (or at least expensive) to root out and consistently punish. Additionally, the law may not extend as far as content producers would like. Copyright protection is subject to limitations and exceptions, including fair use and first sale,[127] which some rights holders see as a threat to their profit margins and reputations. Copyright holders commonly rely on technology to extend and solidify their control over content.[128]

"Technological protection measures" ("TPMs") mediate users' access to and usage of copyrighted media.[129] TPMs can be designed to perform any variety of functions; a simple example is requiring the entry of a password to access content.[130] James Boyle describes these systems as the technological equivalent of barbed wire—just like barbed wire surrounding a farmer's fields, TPMs "fence off" copyrighted material from unwanted trespass, providing "an additional layer of 'physical' protection to the property owner's existing legal protection."[131] Of course, "unlike barbed wire, [TPMs] can also

copyright holder may constitute a 'first sale' for purposes of the first sale doctrine, defendants have the burden of tracing the chain of title to show that their authority to sell Microsoft Products flows from the copyright holder." (citations omitted)). In *Vernor*, the Ninth Circuit brushed aside *amici curiae* eBay and the American Library Association's argument that upholding the license at issue in that case "would force everyone purchasing copyrighted property to trace the chain of title to ensure that a first sale occurred." *Vernor*, 621 F.3d at 1115.

126. *See* Sheridan, *supra* note 117, at 356–57.

127. *See* 17 U.S.C. §§ 107–122 (2012) (enumerating copyright limitations).

128. *See generally* FRED VON LOHMANN, ELEC. FRONTIER FOUND., UNINTENDED CONSEQUENCES: TWELVE YEARS UNDER THE DMCA 1 (2010), *available at* https://www.eff.org/sites/default/files/eff-unintended-consequences-12-years_0.pdf.

129. Dan L. Burk, *Legal and Technical Standards in Digital Rights Management Technology*, 74 FORDHAM L. REV. 537, 546–47 (2005).

130. *Id.* at 547.

131. JAMES BOYLE, THE PUBLIC DOMAIN: ENCLOSING THE COMMONS OF THE MIND 86 (2008).

control what we do once we get access."[132] In addition to regulating access to a music or video file, TPMs may also prevent users from copying it, modifying it, chopping it up, or shifting it from one storage medium to another.[133]

Trepidation over technology-assisted copying is hardly new. Following a setback at the Supreme Court in 1908,[134] music publishers successfully lobbied Congress for protection from the threat of player pianos and the perforated rolls used to play them.[135] In 1982, Jack Valenti of the Motion Picture Association of America told Congress that "the VCR is to the American film producer and the American public as the Boston strangler is to the woman home alone."[136] But the rise of digital technology, particularly the Internet, has elicited unprecedented and seemingly interminable panic in the content industry.[137] Realizing that "technology could . . . safeguard the intellectual property rights threatened by the same technology,"[138] the industry widely embraced digital rights management ("DRM") (essentially, TPMs in the digital sphere), as a means of achieving and maintaining far-reaching control over content.[139] With the coming of personal computing, for instance, software makers employed digital encryption and other technological tools to prevent piracy.[140] The makers of DVDs, e-books, and smartphones have likewise wrapped their products in DRM that restrict both access and usage.[141] Today, society is saturated with DRM;[142] even garage

132. *Id.*

133. *See id.* at 85–86; *see also* Burk, *supra* note 129, at 538–39, 547–48.

134. *See* White-Smith Music Publ'g Co. v. Apollo Co., 209 U.S. 1, 18 (1908) (holding that a "perforated roll" used on a player piano did not constitute an unlawful copy under the Copyright Act).

135. *See* Michael J. Meurer, *Vertical Restraints and Intellectual Property Law: Beyond Antitrust*, 87 MINN. L. REV. 1871, 1908 (2003).

136. *Hearings Before the Subcomm. on Courts, Civil Liberties, and the Admin. of Justice of the Comm. on the Judiciary on H.R. 4783, H.R. 4794, H.R. 4808, H.R. 5250, H.R. 5488, and H.R. 5705*, 97th Cong. 8 (1982).

137. *See* BOYLE, *supra* note 131, at 60–62 (describing industry response to the "Internet threat"). William Patry aptly describes the content industry's response to digital technology as "moral panic." *See* WILLIAM PATRY, MORAL PANICS AND THE COPYRIGHT WARS 28 (2009).

138. Séverine Dusollier, *Electrifying the Fence: The Legal Protection of Technological Measures for Protecting Copyright*, 21 EUR. INTELL. PROP. REV. 285, 285 (1999).

139. BOYLE, *supra* note 131, at 60, 119.

140. *See* Bender, *supra* note 111, at 454–55 (describing "extralegal means" for protecting computer programs).

141. VON LOHMANN, *supra* note 128, at 9–13.

142. *See* Peter Eckersley, *Virtual Markets for Virtual Goods: The Mirror Image of Digital Copyright?*, 18 HARV. J.L. & TECH. 85, 88 (2004) (explaining that DRM seeks to "control the use of information in consumers' homes—and most everywhere else").

door openers and car batteries are controlled by DRM-protected computer programs.[143]

Although DRM is rationalized as necessary for the protection of intellectual property rights,[144] it can, and frequently does, extend content owners' control far beyond the "limited monopoly"[145] bestowed by copyright. By controlling users' ability to copy or manipulate content, copyright holders can now make it impossible for users to exercise fair use rights spelled out in the Copyright Act.[146] Content owners can also employ DRM to tie up public domain material,[147] which is ordinarily "free as the air to common use."[148] Finally, DRM and other protection measures are capable of making an end run around first sale, accomplishing with technology what the software industry has traditionally tried to achieve via restrictive licensing.[149] For example, DRM technology prevents DVDs legally purchased in Japan from playing on most DVD players available in North America.[150]

At first, TPMs were seen as practical "substitutes" for copyright protection.[151] While purportedly bolstering the legal protections of copyright, TPMs "exist[ed] entirely outside of copyright," and functioned independent of it.[152] This was not enough for the content industry; rights holders insisted that TPMs themselves be backed by law.[153] This demand became a reality in

143. *See generally* Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178 (Fed. Cir. 2004) (involving DRM embedded in garage door openers). French car company Renault announced plans to embed DRM in the batteries of its new electric cars, which would allow Renault to rent, rather than sell, the battery to customers. *See* Glyn Moody, *Renault Introduces DRM for Cars*, TECHDIRT (Nov. 12, 2013, 5:01 AM), http://www.techdirt.com/articles/20131108/09350825182/renault-introduces-drm-cars.shtml. The DRM could prevent the battery from charging if, for example, the buyer fell behind on his or her payments. *Id.*

144. *See, e.g.*, Dusollier, *supra* note 138, at 285.

145. 1-1 NIMMER & NIMMER, *supra* note 42, § 1.03[A].

146. *See* BOYLE, *supra* note 131, at 88, 95–96; *see also* 17 U.S.C. § 107 (2012) ("[T]he fair use of a copyrighted work . . . is not an infringement of copyright.").

147. BOYLE, *supra* note 131, at 88.

148. *See* Yochai Benkler, *Free as the Air to Common Use: First Amendment Constraints on Enclosure of the Public Domain*, 74 N.Y.U. L. REV. 354, 354 (1999) (quoting Int'l News Serv. v. Associated Press, 248 U.S. 215, 250 (1918) (Brandeis, J., dissenting)).

149. *See* Burk, *supra* note 129, at 543–45.

150. *See* Stefan Bechtold, *Digital Rights Management in the United States and Europe*, 52 AM. J. COMP. L. 323, 328 & n.19 (2004) ("The 'regional code playback control' used in the DVD standard, for instance, prevents European consumers from playing U.S. DVDs on their domestic DVD players.").

151. Paul Goldstein, *Copyright and Its Substitutes*, 1997 WIS. L. REV. 865, 865.

152. *Id.*

153. *See* Dusollier, *supra* note 138, at 285 ("Building a technical fence around works was not considered as sufficient. Electrifying it by making its circumvention a criminal act was needed. Therefore a due protection of electronic copyright protection and management

1996 with the passage of a series of World Intellectual Property Organization ("WIPO") treaties,[154] which required contracting nations to "provide adequate legal protection and effective legal remedies against the circumvention of effective technological measures."[155] Under these treaties, the mere act of *circumventing* a TPM—cutting the "barbed wire" surrounding a copyrighted work—is itself verboten, regardless of whether the user ever infringes a copyright.[156]

The United States became the first nation to implement the WIPO treaties when President Bill Clinton signed the Digital Millennium Copyright Act ("DMCA") into law in 1998.[157] Even the bill's "proponents acknowledged that the DMCA provisions went beyond what was required to comply with the United States' WIPO treaty obligations."[158] The statute broadly proscribes circumventing TPMs, as well as trafficking in circumvention devices.[159] Specifically, the DMCA protects TPMs that "effectively control[] access to a work" (access controls), and TPMs that "effectively protect[] a right of a copyright holder" (copy controls).[160] The "effectively" qualifier is a very low bar: while the DMCA's provisions do not extend to a TPM that "restricts one form of access but leaves another route wide open,"[161] "the courts are unanimous [in holding] that 'effective' protection does not mean protection that is especially difficult to crack."[162]

---

systems has always been a great concern for rightholders and for the industry developing these systems.").

154. *See* Goldstein, *supra* note 151, at 869.

155. WIPO Copyright Treaty art. 11, Dec. 20, 1996, S. Treaty Doc. 105-17, 36 I.L.M. 65; WIPO Performances and Phonograms Treaty art. 18, Dec. 20, 1996, S. Treaty Doc. 105-17, 36 I.L.M. 76.

156. BOYLE, *supra* note 131, at 87.

157. 2 HANDBOOK OF INFORMATION SECURITY 298 (Hossein Bidgoli ed., 2006).

158. *See* Gwen Hinze, *Brave New World, Ten Years Later: Reviewing the Impact of Policy Choices in the Implementation of the WIPO Internet Treaties' Technological Protection Measure Provisions*, 57 CASE W. RES. L. REV. 779, 794 & n.32 (2007).

159. *See* 17 U.S.C. § 1201(a)–(b) (2012).

160. *Id.*

161. Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 547 (6th Cir. 2004).

162. Jane C. Ginsburg, *Legal Protection for Technological Measures Protecting Works of Authorship: International Obligations and the US Experience*, 29 COLUM. J.L. & ARTS 11, 23 (2005). Courts have explained that requiring more in the way of effectiveness would "gut the statute" by limiting its application to TPMs that successfully thwart circumvention—in other words, precisely those TPMs that have no need for the legal protection. Universal City Studios, Inc. v. Reimerdes, 111 F. Supp. 2d 294, 318 (S.D.N.Y. 2000); *see also* Realnetworks, Inc. v. DVD Copy Control Ass'n, 641 F. Supp. 2d 913, 938 (N.D. Cal. 2009). The court in *Realnetworks* stated:

The mere act of circumventing an access control is illegal, subject only to narrow exceptions—which notably do not include fair use or first sale.[163]

The DMCA forbids trafficking in devices that are designed or marketed for circumventing either access or copy controls.[164] *Universal City Studios v. Reimerdes* was one of the first decisions to apply these provisions.[165] The case revolved around the Content Scramble System ("CSS"), an encryption technology developed by the movie industry to prevent unauthorized access to and copying of DVDs.[166] CSS-protected DVDs can only be played on DVD players that are equipped with the digital "keys" required to decrypt and unscramble CSS encryption.[167] The movie industry licenses CSS technology to manufacturers, who are under strict contractual obligations to maintain the system's security.[168] Among other things, licensed manufacturers may not produce devices that are capable of copying the contents of a DVD.[169]

In September 1999, a Norwegian teenager named Jon Johansen became acquainted with two individuals, known only by their pseudonyms, over the Internet.[170] Together, the three hackers "reverse-engineered a licensed DVD player and discovered the CSS encryption algorithm and keys."[171] With this information in hand, they devised a program called "DeCSS," which was capable of decrypting CSS-protected DVDs.[172] Johansen and his colleagues were purportedly motivated, at least in part, by a desire to create a DVD player that was compatible with the Linux operating system.[173] However, in

> This argument is audacious in its assertion and to countenance it would virtually jettison the law of [§] 1201(b). If this were the law, a copy protection measure would have to protect against every possible current and future means of copying copyrighted content. There could be no liability under this section, for once the copy protection measure had been circumvented, it would purportedly be rendered ineffective. This is circular nonsense.

Id.

163. 17 U.S.C. § 1201(a)(1)(A) (2012); *see id.* § 1201(a)(1)(B)–(E).

164. *Id.* § 1201(a)(2), (b)(1) (2012).

165. Reuven Ashtar, *Licensing as Digital Rights Management, From the Advent of the Web to the iPad*, 13 YALE J.L. & TECH. 141, 160 (2011).

166. *Reimerdes*, 111 F. Supp. 2d at 308.

167. *Id.* at 308, 310.

168. *Id.* at 310.

169. *Id.*

170. *Id.* at 311.

171. *Id.*

172. *Id.*

173. *Id.*

addition to unlocking *access* to movies stored on DVDs, DeCSS allowed users to create *copies* of the digital movie files.[174]

Defendant Eric Corley belonged to a hacker organization named for Emmanuel Goldstein, leader of the underground Brotherhood in George Orwell's *Nineteen Eighty-Four*.[175] Corley published a magazine called *2600: The Hacker Quarterly*, which he founded, appropriately enough, in 1984.[176] In late 1999, Corley posted the source and object code for DeCSS to his website, 2600.com.[177] Corley's website was only one of hundreds from which DeCSS could be downloaded, but for one reason or another, he was one of the few singled out by the movie studios for litigation under the DMCA.[178] The movie studios contended that CSS was an effective protection measure and that Corley had illegally trafficked in a program designed to circumvent CSS.[179] After the district court issued a preliminary injunction, Corley removed DeCSS from 2600.com—but in an act of "electronic civil disobedience," he continued to direct Internet users to as many as 500 other websites offering DeCSS for download.[180] Predictably, this did not help Corley's case.

Judge Lewis A. Kaplan's permanent injunction order in *Reimerdes* reflects an unfamiliarity, and possibly discomfort, with digital technology. The decision began with a lengthy "vocabulary of the case," in which Judge Kaplan methodically defined such terms as *computer*, *Internet*, *CD-ROM*, and *web site*.[181] While venerating coders as "highly skilled human beings [who] can reduce data and instructions to strings of 1's and 0's,"[182] he characterized hackers as degenerates with criminal inclinations.[183] For Judge Kaplan,

174. *Id.*

175. *Id.* at 308; *see* GEORGE ORWELL, NINETEEN EIGHTY-FOUR 11–13 (Signet Classic 1950) (1949).

176. *Reimerdes*, 111 F. Supp. 2d at 308.

177. *Id.* at 309.

178. *Id.* at 311–12.

179. *Id.* at 303, 316.

180. *Id.* at 312–13.

181. *See id.* at 305–07.

182. *Id.* at 306.

183. Judge Kaplan's opinion of hackers was far from veiled:

> The name "2600" was derived from the fact that hackers in the 1960's [sic] found that the transmission of a 2600 hertz tone over a long distance trunk connection gained access to "operator mode" and allowed the user to explore aspects of the telephone system that were not otherwise accessible. Mr. Corley chose the name because he regarded it as a "mystical thing," commemorating something that he evidently admired. Not surprisingly, *2600: The Hacker Quarterly* has included articles on such topics as how to steal an Internet domain name, access other people's e-

hackers like Eric Corley were indistinguishable from burglars, and programs like DeCSS were digital crowbars[184]—except far more dangerous.[185] The judge essentially embraced and codified the content industry's panic mentality: "from Judge Kaplan's language it is evident that he sees [DeCSS] not as an act of expression but as a virus spreading like wildfire. . . . Just as in an epidemic, the harshest measures are called for."[186] The *Reimerdes* decision foresaw—and attempted to avert—a "looming digital Black Death" heralded by DeCSS and other tools for circumventing DRM.[187]

While the specter of unmitigated digital piracy certainly animated Judge Kaplan's opinion,[188] he analyzed the movie studios' DMCA claims within the framework of § 1201(a)(2), which prohibits trafficking in technologies that circumvent *access* controls, rather than § 1201(b), which deals with *copy* controls.[189] The analysis was straightforward: CSS effectively controls access to copyrighted works (i.e., movies stored on DVDs); DeCSS "clearly is a means of circumventing a technological access control measure"; and Corley illegally trafficked in DeCSS.[190] Judge Kaplan acknowledged that access controls may preclude lawful uses of copyrighted material such as fair use,[191] and may even restrict access to works that are not protected by copyright,[192] but concluded that the DMCA "fundamentally altered the landscape," such that some traditional copyright norms do not apply to DRM.[193] In a newly

> mail, intercept cellular phone calls, and break into the computer systems at Costco stores and Federal Express. One issue contains a guide to the federal criminal justice system for readers charged with computer hacking.

Id. at 308–09 (footnotes omitted).

184. *See id.* at 316 (quoting Congress as stating that circumventing DRM is "the electronic equivalent of breaking into a locked room"); *id.* at 329 (comparing laws "preventing people from circumventing technological access control measures" to "laws prohibiting the possession of burglar tools").

185. *See id.* at 315 ("[Publishing DeCSS] is analogous to the publication of a bank vault combination in a national newspaper."). Judge Kaplan continued with this analogy to further emphasize the harm, explaining that "[d]evelopment and implementation of a new DVD copy protection system, however, is far more difficult and costly than reprogramming a combination lock and may carry with it the added problem of rendering the existing installed base of compliant DVD players obsolete." *Id.*

186. BOYLE, *supra* note 131, at 101.

187. *Id.* at 102.

188. *See Reimerdes*, 111 F. Supp. 2d at 309, 313–15, 329–30, 335 & n.230 (emphasizing threat of piracy).

189. *Id.* at 316. The movie studios also brought claims under § 1201(b). *See id.* at 316 n.133.

190. *See id.* at 317–19.

191. *Id.* at 322.

192. *Id.* at 322 n.159.

193. *Id.* at 323.

digital world, fair use and other copyright limitations must give way in order to stamp out what Judge Kaplan described as "the electronic equivalent of breaking into a locked room in order to obtain a copy of a book."[194]

What the *Reimerdes* opinion fails to acknowledge is that, if the breaking-and-entering analogy is taken at face value, then the burglar is implicitly breaking into the locked room in order to gain access to a book he *lawfully owns*.[195] This perfectly (if inadvertently) illustrates why DRM is problematic for the first sale doctrine. As many commenters have expressed, legally protected DRM "give[s] copyright owners the ability to maintain a running control on access to . . . their works," which "frustrates the goal of the first sale doctrine, by extending the rights of the copyright owner beyond the first sale of a particular copy."[196] The U.S. Copyright Office collected and summarized these comments in an August 2001 report that it prepared pursuant to Section 104 of the DMCA.[197] The DMCA's implications for first sale elicited a "dramatic range of opinions," many of which argued that the anticircumvention provisions would restrict the doctrine's application.[198] Several commenters specifically discussed CSS protection, as well as the related practice of DVD region coding.[199]

Although the Copyright Office was not terribly swayed by these arguments,[200] it did acknowledge that "legitimate concerns have been raised about what may develop as the market and technology evolve."[201] The Copyright Office singled out "[t]he practice of using technological measures to tether a copy of a work to a particular hardware device" as raising special challenges for first sale rights.[202] It did not endorse congressional action, since tethering was still uncommon at the time the report was published[203]—but it left the door open for recommending amendments to the first sale

194. *Id.* at 316 (quoting H.R. REP. NO. 105-551, at 17 (1998)).

195. In the context of *Reimerdes*, DeCSS permitted users to gain access to the content stored on *legally purchased* DVDs, using computers running the Linux operating system. *See id.* at 319; *see also* BOYLE, *supra* note 131, at 100 ("In other words, even if Mr. Johansen made DeCSS so that he and his friends could watch DVDs they purchased legally on computers running Linux, they could still be liable for breaking the DMCA").

196. U.S. COPYRIGHT OFFICE, DMCA SECTION 104 REPORT 39 (2001).

197. *See id.* at v.

198. *Id.* at 34.

199. *See id.* at 35–37.

200. *See id.* at 73–76.

201. *Id.* at 96.

202. *Id.* at 76.

203. *Id.*

doctrine in the future.[204] Now may be an opportune time to revisit these concerns.

IV. PLAYING WITH POWER: CIRCUMVENTION TO PRESERVE FIRST SALE

The development of anti-used game technology is but the latest content industry tactic for limiting the first sale doctrine. Although video game technology is generally less susceptible to piracy than computer software, video game titans have long employed DRM and long resisted the first sale doctrine. Tethering technology weds these two ideas, using DRM to permanently link video game discs or other media to particular pieces of hardware. Assuming that the content industry adopts some form of tethering technology in the near future, the question necessarily arises: would circumventing this technology in a manner that permits a single copy to be played on multiple devices violate the DMCA? (The predicate question—*whether* the technology will be circumvented—need not be asked, for "DRM is *always* circumvented" sooner or later.[205]) The answer is far from simple. Despite the *Reimerdes* court's wooden interpretation of the DMCA, subsequent appellate court decisions have produced contradictory views of the statute's anticircumvention provisions.[206] The resulting ambiguity only makes it more likely that video game makers and other content owners will shore up their position through restrictive licensing.

A. EARLY SHOTS ACROSS THE BOW

The video game industry's suspicion toward first sale is by no means of recent vintage. Nintendo first targeted the doctrine in the late 1980s and early 1990s, when the company enjoyed a dominance that the video game industry had never before seen and that has not since been replicated.[207] Nintendo spent years fighting the bourgeoning video game rental industry, which depended on the first sale doctrine for its survival.[208] Rental businesses

204. *See id.* at 96–97.

205. Cory Doctorow, *DRM Gives Companies Security—From Competition*, BOINGBOING (Feb. 17, 2012, 9:10 AM), http://boingboing.net/2012/02/17/drm-gives-companies-security.html (emphasis added).

206. *See, e.g.*, MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928 (9th Cir. 2010) (interpreting the DMCA as granting additional anticircumvention rights to copyright holders); Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178 (Fed. Cir. 2004) (articulating that the DMCA anticircumvention provisions only prohibit forms of access that bear a "reasonable relationship" to copyright protection).

207. *See* SHEFF, *supra* note 88, at 402.

208. *Id.* at 276.

insisted that they were good for Nintendo's bottom line, but Nintendo disagreed.[209] Nintendo executive and lawyer Howard Lincoln characterized video game rental as "nothing less than commercial rape," pointing out that a single copy of a game could be rented out to potentially hundreds of customers who might have otherwise bought a new copy for $40 or $50.[210] Blockbuster generated $1.5 billion in revenue in 1990, as much as ten percent of which was derived from Nintendo game rentals.[211] Aside from profits made from sales of rental cartridges, Nintendo did not receive any of that money.

Considering that demand for Nintendo game cartridges consistently outpaced supply during this period,[212] it is difficult to believe that rental stores displaced new cartridge sales in any significant degree. At the time, Nintendo controlled eighty-five to ninety percent of the Japanese and American home video game markets and generated several billion dollars per year in net sales.[213] The company nonetheless went on the attack. Nintendo considered suing rental businesses, but the first sale doctrine stood squarely in its way.[214] The company then sought to change first sale law by lobbying Congress for a prohibition on video game rentals.[215] Nintendo joined the coalition of software companies pushing for a ban on computer software rental, but when the Video Software Dealers Association (a trade association of video-rental dealers) threatened to crush the bill if it included video games, the software companies succumbed.[216] The bill was rewritten to exclude "computer program[s] embodied in or used in conjunction with a limited purpose computer that is designed for playing video games" and ultimately adopted.[217]

Political infighting may have killed the video game rental prohibition, but the official rationale for excluding it from the Computer Software Rental Amendments Act of 1990 was sound: "It was easy to rent computer software and make a permanent copy at home, but it was nearly impossible to make a copy of 'Super Mario Bros.' "[218] Unlike traditional software makers, video game companies simply did not have as much to lose from the rental

209. *Id.* at 283.
210. *Id.* at 276, 283.
211. *Id.* at 283.
212. *See id.* at 203, 244–45.
213. *See id.* at 196, 349, 402.
214. *Id.* at 284.
215. *Id.* at 283–84.
216. *Id.*
217. 17 U.S.C. § 109(b)(1)(B)(ii) (2012).
218. SHEFF, *supra* note 88, at 284; *see supra* notes 78–83 and accompanying text.

industry. Including video games in the 1990 Amendments Act would have served but one purpose: shoring up Nintendo's monopoly profits. Despite losing in Congress, the video game giant did not back down. Nintendo refused to sell cartridges directly to rental chains and forbade retailers from selling multiple copies of any game to a single customer.[219] In at least one instance, Nintendo threatened retaliation against a retailer who sold to renters.[220]

In short, video games and the first sale doctrine got off to a rocky start. Although revenue-sharing agreements (which have overtaken the rental industry) may have softened game makers' feelings toward rental chains,[221] the existence of a robust secondhand market for video games means that the first sale doctrine remains controversial within the industry.

B. TETHERING: THE LATEST ATTACK ON FIRST SALE

Tethering technology represents a promising strategy for video game producers fed up with competition from the secondhand market. DRM is certainly not new to video games,[222] but as Sony acknowledged in its Disc ID patent application, "[t]ypically, DRM is a technology for the prevention of the unlimited copy [sic] of electronic content."[223] By contrast, tethering technology seeks to "prevent . . . electronic content from being *used* unlimitedly" by successive owners.[224] More to the point, the technology "reliably restricts the use of electronic content dealt in the second-hand markets."[225] Were video game makers to widely adopt tethering technology, buying and selling used games could become a thing of the past.

Tethering technology can take many different forms.[226] The vision set forth in the Disc ID patent application would be to package each video game

---

219. SHEFF, *supra* note 88, at 276, 283.

220. *Id.* at 276–77. Nintendo also sued Blockbuster for including photocopied instruction manuals with rented games. *See* Nadia Oxford, *Trials and Tribulations: Video Games' Many Visits to Court*, 1UP.COM (Dec. 14, 2005), http://www.1up.com/features/trials-and-tribulations.

221. *See* James D. Dana & Kathryn E. Spier, *Revenue Sharing and Vertical Control in the Video Rental Industry*, 49 J. INDUS. ECON. 223 (2001); Julie H. Mortimer, *Vertical Contracts in the Video Rental Industry*, 75 REV. ECON. STUD. 165 (2008).

222. Nintendo's NES console included a "lock-out" security system that prevented counterfeit and unlicensed games from playing. *See* Atari Games Corp. v. Nintendo of Am. Inc., 975 F.2d 832, 844 (Fed. Cir. 1992) (describing NES security system); Atari Games Corp. v. Nintendo of Am., Inc., 897 F.2d 1572, 1574–75 (Fed. Cir. 1990) (same).

223. Sony Patent App., *supra* note 2, at [0006].

224. *Id.* (emphasis added).

225. *Id.*

226. For example, Microsoft originally designed its Xbox One system to require a virtually constant Internet connection, which would permit the company to enforce policies

disc with a "use permission tag," which would be capable of wirelessly communicating with a sensor on the game console.[227] Boiled down to essentials, when a game disc is first played, the use permission tag would associate that specific disc with the specific game console in which it is used.[228] From that point on, every time the disc is played, the user would have to swipe the tag across a scanner located on the console, allowing the tag to check the disc ID and console ID.[229] If the disc ID and console ID match the information stored on the tag, the disc would run; if either the disc ID or the console ID does not match up, the disc would not operate.[230] Since each game disc can only be associated with one console, as a practical matter, lending or selling the disc to another gamer would be pointless—the disc would be incapable of playing on another person's console. The first sale doctrine would be rendered obsolete for game consoles employing this technology.

C. LAWFUL CIRCUMVENTION?

Video game players, proprietors of used games, digital activists, and others[231] would have an obvious interest in coming up with a way to circumvent tethering technology. The question is whether doing so would run afoul of the DMCA's anticircumvention provisions.[232] As a threshold matter, it appears that tethering technology like Disc ID would readily qualify for DMCA protection as a technological measure that "effectively controls access to a work."[233] If the Disc ID patent application is to be believed, the technology "in the ordinary course of operation, requires the application of information, or a process or a treatment, with the authority of the copyright

pertaining to used and rented games. *See* Andrew Goldfarb, *Microsoft Details Xbox One Used Games, Always Online*, IGN (June 6, 2013), http://www.ign.com/articles/2013/06/06/microsoft-details-xbox-one-used-games-always-online.

227. Sony Patent App., *supra* note 2, at [0033] ("In other words, the game disk 210 and the use permission tag 220 are commercially released and transacted as a set and are also distributed integrally as an inseparable set, that is, are distributed as a bundle.").

228. *See* Luke Karmali, *Report: New Sony Patent Blocks Second Hand Games*, IGN (Jan. 3, 2013), http://www.ign.com/articles/2013/01/03/report-new-sony-patent-blocks-second-hand-games.

229. Sony Patent App., *supra* note 2, at 4.

230. *Id.* The patent application also describes a "temporary use device," which would allow a disc to be played on a console for a limited number of times without permanently associating that disc with that console. *See* Sony Patent App., *supra* note 2, at [0075], fig.7. This would presumably facilitate rental.

231. For example, intermediaries such as eBay.

232. *See* U.S. COPYRIGHT OFFICE, *supra* note 196, at 75 ("The only way of accessing the content on another device would be to circumvent the tethering technology, which would violate [§] 1201.").

233. 17 U.S.C. § 1201(a)(1)(A) (2012).

owner, to gain access to the work."[234] That is, in order to obtain access to game content stored on a disc, the user would have to swipe the use permission tag across the game console's reader, and if the information supplied by the tag matches both the disc ID and the console ID, only then will the game run.[235] The tag ensures that users access game content solely on the copyright owners' terms. This "authentication sequence" and others like it readily satisfy the "effectiveness" bar to DMCA protection.[236]

With that preliminary showing in place, circumventing a tethering system may seem like an uncontroverted DMCA violation. The DMCA prohibits users from circumventing TPMs that effectively control access to a work,[237] and breaking a tethering system like Disc ID would do just that. *Reimerdes* left very little wiggle room.[238] Just as CSS encryption prevents DVDs from playing on unlicensed DVD players, tethering technology would prevent copies of copyrighted media from playing on unauthorized devices (i.e., any device other than the one on which the copy was first used). On Judge Kaplan's reasoning, circumventing tethering technology would violate the DMCA just as "clearly" as DeCSS did.[239]

The DMCA has already been applied in the video games context—to an earlier generation of Sony DRM, no less. In *Sony Computer Entertainment America v. GameMasters*,[240] Sony asked a California district court to issue a preliminary injunction blocking defendants from selling the "Game Enhancer," a device that allowed American gamers "to play games sold in Japan or Europe and intended by [Sony] for use exclusively on Japanese or European PlayStation consoles."[241] Sony's PlayStation consoles featured classic lock-and-key DRM: "The PlayStation console is designed to operate only when encrypted data is read from a CD-ROM that verifies that the CD is an authorized, legitimate [Sony] product licensed for distribution in the same geographical territory of the console's sale."[242] Each console could only access those games with "data codes" matching the region in which the

234. *Id.* § 1201(a)(3)(B).

235. *See* Sony Patent App., *supra* note 2, at [0072].

236. *See* Lexmark Int'l Inc. v. Static Control Components, Inc., 387 F.3d 522, 547 (6th Cir. 2004) (holding the effectiveness bar precludes DMCA protection where a technological measure leaves a route to access "wide open").

237. 17 U.S.C. § 1201(a)(1)(A).

238. *See* Universal City Studios, Inc. v. Reimerdes, 111 F. Supp. 2d 294 (S.D.N.Y. 2000).

239. *See id.* at 314; *supra* text accompanying notes 188–94.

240. Sony Computer Entm't Am. Inc. v. GameMasters, 87 F. Supp. 2d 976 (N.D. Cal. 1999).

241. *Id.* at 977, 981.

242. *Id.* at 981.

console itself was sold.[243] In practical terms, the DRM system prevented games sold abroad from playing on American PlayStation consoles.[244] The Game Enhancer circumvented this system, allowing gamers to play legally purchased Japanese and European games on their American consoles.[245] Sony alleged that its lock-and-key system qualified for DMCA protection as an access control measure and that defendants violated the DMCA by trafficking in the Game Enhancer device.[246] The court agreed that "the Game Enhancer appear[ed] to be a device whose primary function is to circumvent 'a technological measure . . . that effectively controls access to a system protected by a registered copyright.' "[247] Finding that Sony was likely to prevail on its DMCA claim, the court issued the preliminary injunction.[248]

Perhaps to a greater extent than *Reimerdes*, *Sony v. GameMasters* appears to foreclose any reasonable argument that circumventing tethering technology would not run afoul of the DMCA. Tethering technology is essentially a more sophisticated version of the DRM at issue in *GameMasters*, except that rather than limiting a game's operability to consoles sold within the same geographic region, tethering limits a game's operability to a *single* console. That the gamer may have legally purchased the game is of no matter; in *GameMasters*, the court acknowledged that the games at issue may have been "legally, validly manufactured and sold in Japan," and that they "do not become transformed into illegal, bootleg infringing games merely because they are transported across the ocean."[249] That did not preclude DMCA liability.[250] As Judge Kaplan would remind us, the DMCA forbids "the electronic equivalent of breaking into a locked room"—even, apparently, when the only purpose is to gain access to a video game that the alleged "burglar" lawfully purchased.[251]

But the story does not end with *GameMasters* and *Reimerdes*. In 2004, the Federal Circuit issued a major decision interpreting the DMCA

243. *Id.* at 987.
244. *Id.* at 981.
245. *Id.*
246. *Id.* at 987.
247. *Id.* (quoting 17 U.S.C. § 1201(a)(2)(A)).
248. *GameMasters*, 87 F. Supp. 2d at 988.
249. *Id.* at 986. The Supreme Court recently clarified that copyrighted media legally purchased abroad may be imported and resold in the United States without violating the Copyright Act. *See* Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351 (2013).
250. *GameMasters*, 87 F. Supp. 2d at 986–88.
251. Universal City Studios, Inc. v. Reimerdes, 111 F. Supp. 2d 294, 316 (S.D.N.Y. 2000) (quoting H.R. REP. NO. 105-551(I), at 17 (1998)); *see also supra* notes 194–99 and accompanying text.

anticircumvention rules within the broader context of copyright protection.[252] In *Chamberlain Group v. Skylink*, the circuit court reviewed a summary judgment order finding that defendant Skylink's "universal" garage door opener did not violate Chamberlain's rights in its "Security+" garage door technology.[253] Unlike ordinary garage door systems, which typically rely on a single fixed transmitter signal, Security+ garage door openers incorporated a "rolling code" computer program, which constantly changed the transmitter signal required to open the garage door.[254] When the Security+ garage door receptor detected a signal, it would check it against the previous 1,024 signals received.[255] If the signal were identical to any of the most recent 1,024 signals, the garage door would not open.[256] If, however, the signal were among the next 4,096 binary signals incorporated in the Security+ system, the garage door would open.[257] The idea behind Security+ was to prevent "code grabbing," a practice whereby burglars might theoretically detect a garage door system's fixed transmitter signal and use it to surreptitiously enter the garage.[258]

Skylink's "Model 39" universal transmitter was designed to interoperate with many common garage door systems, including rolling code and fixed code systems.[259] The Model 39 did not use rolling code technology but rather simulated the effects of it.[260] When the button on a Model 39 opener was pressed, the transmitter would emit three codes in rapid succession, as determined by a complex algorithm.[261] This programming allowed the Model 39 to interoperate with Chamberlain's Security+ system, meaning that homeowners who purchased a Model 39 "universal" opener could use the device to open their Security+ automatic garage door.[262]

Chamberlain characterized its Security+ system as a "technological measure . . . control[ling] access" to a copyrighted work—the copyrighted

252. Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178 (Fed. Cir. 2004).
253. *Id.* at 1181, 1183.
254. *Id.* at 1183.
255. *Id.* at 1184.
256. *Id.*
257. *Id.*
258. *Id.* at 1183–84. As the court noted, and as Chamberlain conceded, "code grabbers are more theoretical than practical burgling devices; none of [Chamberlain's] witnesses had either firsthand knowledge of a single code grabbing problem or familiarity with data demonstrating the existence of a problem." *Id.*
259. *Id.* at 1184.
260. *Id.* at 1184–85.
261. *Id.* at 1185.
262. *Id.*

"rolling code" computer programs that governed the garage door.[263] Because "[t]he only way for the Model 39 to interoperate with a Security+ [garage door opener] is by 'accessing' copyrighted software," Chamberlain argued, Skylink had "committed a per se violation of the DMCA."[264] The district court disagreed and granted Skylink's summary judgment motion.[265] The court held that Chamberlain had not demonstrated that circumventing the Security+ system afforded customers *unauthorized* access to Chamberlain's software.[266]

The Federal Circuit affirmed the district court, but for different reasons. The Federal Circuit premised its decision on the proposition that the DMCA anticircumvention provisions "do not establish a new property right."[267] Unlike copyrights and patents, which are considered property, "liability protection from unauthorized circumvention merely creates a new cause of action under which a defendant may be liable."[268] To the Federal Circuit panel, this distinction between property and liability was "critical" to its interpretation of the DMCA.[269] The court rejected Chamberlain's argument that "Congress empowered manufacturers to prohibit consumers from using embedded software products in conjunction with competing products," reasoning that such a broad reading of the DMCA would be plausible only "if the anticircumvention provisions established a new property right capable of conflicting with the copyright owner's other legal responsibilities," rather than merely a "new way" to secure preexisting intellectual property rights.[270] Contrary to Judge Kaplan's assertion in *Reimerdes*, the Federal Circuit held that "the DMCA emphatically *did not* 'fundamentally alter' the legal landscape governing the reasonable expectations of consumers or competitors."[271]

Since the DMCA merely "introduce[d] new grounds for liability *in the context* of the unauthorized access of copyrighted material,"[272] the court reasoned that the circumvention prohibitions "applie[d] only to

263. *Id.* at 1185, 1197.

264. *Id.* at 1197.

265. Chamberlain Grp., Inc. v. Skylink Techs., Inc., 292 F. Supp. 2d 1040, 1046 (N.D. Ill. 2003), *aff'd*, 381 F.3d 1178 (Fed. Cir. 2004).

266. *See id.* at 1040–41, 1045–46.

267. Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178, 1179, 1192 (Fed. Cir. 2004).

268. *Id.* at 1192–93.

269. *Id.* at 1192.

270. *Id.* at 1193–94.

271. *Id.* at 1194; *cf.* Universal City Studios, Inc. v. Reimerdes, 111 F. Supp. 2d 294, 323 (S.D.N.Y. 2000) ("By prohibiting the provision of circumvention technology, the DMCA fundamentally altered the landscape.").

272. *Chamberlain*, 381 F.3d at 1194 (emphasis added).

circumventions reasonably related to protected rights."[273] The court found support for this construction in legislative history as well as the statute itself, which consistently links "access" to copyright "protection."[274] Chamberlain's arguments predictably relied heavily on *Reimerdes*, but the Federal Circuit found that decision inapposite.[275] Whereas DeCSS permitted not only unauthorized access but also unauthorized copying (i.e., infringement), the Model 39 "enable[d] only legitimate uses of copyrighted software."[276] In the decision's key passage, the circuit court concluded that Chamberlain's broad reading of the DMCA would afford the owners of copyrighted works "unlimited rights to hold circumventors liable . . . merely for accessing that work, even if that access enabled only rights that the Copyright Act grants to the public."[277] The court suggested that, even allowing for the substantial deference given Congress in its exercise of the authority bestowed by the Copyright Clause, such a redefinition of copyright owners' rights could potentially violate the Constitution.[278] In affirming the district court, the Federal Circuit articulated the following rule: the DMCA anticircumvention provisions only prohibit forms of access that bear a "reasonable relationship" to copyright protection.[279] Because Chamberlain failed to establish this "critical nexus," the circuit court affirmed the district court's summary judgment in favor of Skylink.[280]

Although the Federal Circuit did not mention the first sale doctrine in its *Chamberlain* opinion, it nevertheless relied heavily upon the classical understanding of personal property rights that first sale embodies. The court

---

273. *Id.* at 1195.

274. *See id.* at 1197 ("[I]t is significant that virtually every clause of § 1201 that mentions 'access' links 'access' to 'protection.' ").

275. *See id.* at 1198 ("Though Chamberlain is correct in considering some of the *Reimerdes* language supportive, it is the differences between the cases, rather than their similarities, that is most instructive in demonstrating precisely what the DMCA permits and what it prohibits.").

276. *Id.*

277. *Id.* at 1200 (emphasis omitted).

278. The Federal Circuit noted:

> [A]s the Supreme Court recently explained, "Congress' exercise of its Copyright Clause authority must be rational." . . . Chamberlain's proposed construction of § 1201(a) implies that in enacting the DMCA, Congress attempted to "give the public appropriate access" to copyrighted works by allowing copyright owners to deny all access to the public. Even under the substantial deference due Congress, such a redefinition borders on the irrational.

Id. (quoting Eldred v. Ashcroft, 537 U.S. 186, 205 n.10 (2003)).

279. *Id.* at 1202.

280. *Id.* at 1204.

began by noting that, in the absence of "explicit restrictions" on their use of the Security+ system, Chamberlain's customers may justifiably assume that they "enjoy all of the rights associated with the use of their [garage door openers] and any software embedded therein" that the laws provide.[281] Neither party disputed that "a homeowner who purchases a Chamberlain [garage door opener] owns it and has a right to use it to access his or her own garage."[282] The court rejected the notion that Chamberlain was "entitled to prohibit legitimate purchasers of its embedded software from 'accessing' the software by using it," since such a construction would "allow copyright owners to prohibit exclusively fair uses even in the absence of any feared foul use."[283] The court declared unequivocally that "[c]onsumers who purchase a product containing a copy of embedded software have the inherent legal right to use that software."[284] While the court described this legal right as "inherent," the right is also explicit: the first sale doctrine grants lawful owners of copyrighted material the right to use the particular copy that they purchased.[285] "What the law authorizes, Chamberlain cannot revoke."[286]

Under *Chamberlain*, the legal argument against circumventing tethering technology is considerably less certain. Assuming that hackers came up with a method for circumventing the technology that did not also facilitate illicit copying—a method that broke through the game console's access controls but not its copy controls—the question would be whether the circumvention bears "a reasonable relationship to the protections that the Copyright Act otherwise affords copyright owners."[287] That is, would circumventing a tethering system like Disc ID permit access to copyrighted video game content "in a manner that . . . infringes or facilitates infringing a right protected by the Copyright Act"?[288] The rights bestowed by the Copyright Act are enumerated in § 106.[289] That section provides that, subject to the limitations and exceptions prescribed in §§ 107 through 122—such as first

281. *Id.* at 1183.
282. *Id.* at 1187.
283. *Id.* at 1202 (emphasis omitted).
284. *Id.*
285. *See* Bell Atl. Bus. Sys. Servs., Inc. v. Hitachi Data Sys. Corp., No. C93-20079 JW, 1995 WL 798935, at *3 (N.D. Cal. Mar. 10, 1995) ("As BABSS correctly points out, 'a "first sale" by Comparex terminates Hitachi Data's right to control the distribution of the service items, and BABSS would have every legal right to use the service items to maintain its customers' equipment.' " (quoting Parfums Givenchy, Inc. v. Drug Emporium, Inc., 38 F.3d 477, 481, 482 n.8 (9th Cir. 1994))).
286. *Chamberlain*, 381 F.3d at 1202.
287. *Id.* at 1202.
288. *Id.* at 1203 (emphasis omitted).
289. 17 U.S.C. § 106 (2012).

sale[290] and fair use[291]—copyright owners in video games and other audiovisual works enjoy exclusive rights of reproduction, preparation of derivative works, distribution, public performance, and public display.[292] Merely *accessing* video game content stored on a disc does not implicate any of those rights.[293]

One might argue that circumventing tethering technology "reasonably relates" to the distribution right because it facilitates the resale of video games.[294] But so long as new game discs are actually *sold* to customers, the first sale doctrine exhausts the copyright owner's distribution right at the moment of transfer.[295] As the *Chamberlain* court pointed out, sanctioning the technological repeal of copyright limitations would contradict DMCA § 1201(c)(1), which states that the provisions against circumvention shall not affect copyright limitations or defenses to infringement, including fair use (and implicitly, first sale).[296] If purchasers of Chamberlain's Security+ system possess an "inherent right" to use the software embedded in their garage door opener, it seems inescapable that lawful purchasers of video games—including secondhand purchasers—must have the same right to use the software embedded in their game discs.[297]

Not all courts have followed the Federal Circuit's lead. Most notably, the Ninth Circuit took issue with *Chamberlain*'s "nexus" requirement in *MDY*

290. *Id.* § 109(a).

291. *Id.* § 107.

292. *Id.* § 106(1)–(5).

293. *See* Storage Tech. Corp. v. Custom Hardware Eng'g & Consulting, Inc., 421 F.3d 1307, 1318 (Fed. Cir. 2005) ("[C]ourts generally have found a violation of the DMCA only when the alleged access was intertwined with a right protected by the Copyright Act.").

294. *Cf. id.* ("A copyright owner alleging a violation of [§] 1201(a) consequently must prove that the circumvention of the technological measure either 'infringes or facilitates infringing a right protected by the Copyright Act.' " (quoting *Chamberlain*, 381 F.3d at 1203)).

295. *See* Sebastian Int'l, Inc. v. Consumer Contacts (PTY) Ltd., 847 F.2d 1093, 1099 (3d Cir. 1988) ("Because that exclusive [distribution] right is specifically limited by the first sale provisions of § 109(a), it necessarily follows that once transfer of ownership has cancelled the distribution right to a copy, the right does not survive . . . .").

296. Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178, 1200 (Fed. Cir. 2004); 17 U.S.C. § 1201(c)(1) (2012); *see* Vernor v. Autodesk, Inc., 621 F.3d 1102, 1107 (9th Cir. 2010) ("The exclusive distribution right is limited by the first sale doctrine, an affirmative defense to copyright infringement that allows owners of copies of copyrighted works to resell those copies."); Allison v. Vintage Sports Plaques, 136 F.3d 1443, 1444 (11th Cir. 1998) (stating that the first sale doctrine is a "well-established limitation on intellectual property rights"); Pearson Educ., Inc. v. Liu, 656 F. Supp. 2d 407, 409 (S.D.N.Y. 2009) ("This limitation on scope of the distribution right, known as the first sale doctrine, is codified in § 109(a) of the Act.").

297. *See Chamberlain*, 381 F.3d at 1202.

Industries, LLC v. Blizzard Entertainment, Inc.[298] That case concerned a "bot" program that a computer programmer had developed for use in Blizzard's popular online computer game, *World of Warcraft*.[299] The bot, called "Glider," essentially controlled the game on auto-pilot. As Glider's website explained, "You tell [Glider] about your character, where you want to kill things, and when you want to kill. Then it kills for you automatically. You can do something else, like eat dinner or go to a movie, and when you return, you'll have a lot more experience and loot."[300] Glider allowed players to quickly amass experience points and in-game currency, rendering their *World of Warcraft* avatars more powerful and allowing them to purchase more expensive items for in-game use.[301] Glider's creator, Michael Donnelly, began selling copies of Glider through his company, MDY Industries, in summer 2005.[302] Shortly after Glider hit the market, Blizzard launched a program called "Warden" to detect and prevent players using bots from connecting to its servers, prompting Donnelly to modify Glider to effectively avoid detection by Blizzard.[303] Donnelly acknowledged in online statements that Glider violated Blizzard's Terms of Use, remarking that "[a]voiding detection is rather exciting, to be sure."[304] After Blizzard sent a cease-and-desist letter to MDY in August 2006 and dispatched an attorney to Donnelly's home in October 2006, MDY brought suit in the District of Arizona seeking a declaration that Glider did not infringe Blizzard's copyrights or any other rights.[305] Blizzard filed counterclaims alleging vicarious and contributory copyright infringement and violations of the DMCA.[306] After a bench trial, the district court held MDY liable for violating the DMCA and Donnelly

298. MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928, 950 (9th Cir. 2010).

299. *Id.* at 934–35.

300. *Id.* at 935. Appropriately, the term "bot" is short for "robot." *Id.*

301. *See id.* at 935–36. The appeal of a "bot" is that it allows players to gain experience points and in-game currency without the monotony of "grinding"—"the action of repeatedly killing mobs inside of a video game in order to mass experience or currency," which may be required in order to progress to higher stages in a role-playing video game. *See* Justin Eldridge, *What is Grinding? Grinding in Video Games*, EXAMINER.COM (Mar. 5, 2012), http://www.examiner.com/article/what-is-grinding-grinding-video-games. The necessity of "grinding" in role-playing video games—particularly subscription-based games like *World of Warcraft*—has sparked a cottage industry in China, where low-paid workers spend hours upon hours grinding for in-game currency—which is then sold to Western gamers online, in exchange for *real* money. *See* Julian Dibbell, *The Life of a Chinese Gold Farmer*, N.Y. TIMES MAG., June 17, 2007, at 36.

302. *MDY Indus.*, 629 F.3d at 936. By September 2008, the company had reaped gross revenues of $3.5 million from Glider sales. *Id.*

303. *Id.*

304. *Id.* (internal quotations omitted).

305. *Id.* at 936–37.

306. *Id.* at 937.

personally liable for copyright infringement, and entered a judgment for $6.5 million.[307]

Blizzard presented several theories of liability for its DMCA claims, but its central argument related to *World of Warcraft*'s "dynamic non-literal elements: that is, the 'real-time experience traveling through different worlds, hearing their sounds, viewing their structures, encountering their inhabitants and monsters, and encountering other players.' "[308] Blizzard argued, and the district court agreed, that Warden effectively controlled access to *World of Warcraft*'s dynamic non-literal elements, and that MDY unlawfully marketed Glider as a tool for circumventing Warden.[309] Since Glider did not infringe or facilitate infringement of Blizzard's copyrights in the game's dynamic non-literal elements, MDY's appeal required the Ninth Circuit to address whether the DMCA prohibits circumvention where infringement is not an issue.[310]

The Ninth Circuit considered, but ultimately rejected, the Federal Circuit's reasoning in *Chamberlain*. The court observed that "[n]either of [§ 1201(a)'s] two subsections explicitly refers to traditional copyright infringement," despite the statute's use of the term "work protected under this title."[311] Further, the DMCA's definition of circumvention describes acts that do not necessarily infringe or facilitate infringement.[312] The court therefore interpreted the DMCA as "granting copyright owners a new anti-circumvention right."[313]

The Ninth Circuit found support for this construction in the DMCA's legislative history.[314] For example, the Senate Judiciary Committee's report explains that the two subsections prohibiting trafficking in circumvention devices—one for access controls, the other copy controls—are "not interchangeable," for the former "protect[s] access to a copyrighted work," while the latter "protect[s] the traditional copyright rights of a copyright owner."[315] The court then cited an example of DMCA liability given in the Committee's report: crafting a tool to bypass a password in order to gain access to a copyrighted work.[316] Notably, in this example, liability would

307. *Id.*
308. *Id.* at 942–43.
309. *Id.* at 943.
310. *Id.*
311. *Id.* at 945 (internal quotations omitted).
312. *See id.* (quoting 17 U.S.C. § 1201(a)(3)(A) (2010)).
313. *Id.*
314. *See id.* at 946–47.
315. *Id.* at 947 (quoting S. REP. NO. 105-109, at 12 (1998)).
316. *Id.*, 629 F.3d at 947.

attach even though the circumvention bears no nexus to infringement.[317] Echoing Judge Kaplan, the Ninth Circuit underscored the moral turpitude of infringement-free circumvention, emphasizing that the DMCA's proscriptions are "roughly analogous to making it illegal to break into a house using a tool, the primary purpose of which is to break into houses."[318] While acknowledging the policy considerations that drove the Federal Circuit's reasoning, the Ninth Circuit concluded that "to follow *Chamberlain* in imposing an infringement nexus requirement" would mean "disregard[ing] the plain language of the statute."[319] The court held that MDY had violated the DMCA with respect to *World of Warcraft*'s dynamic non-literal elements.[320]

With *Chamberlain* and *MDY* both on the books, there is now a circuit split as to whether circumventing an access control gives rise to liability where there is no threat of infringement.[321] Thus far, other circuits have done little to resolve the disagreement. The Fifth Circuit initially embraced the Federal Circuit's rule, but then issued a revised opinion omitting any mention of *Chamberlain*.[322] In affirming Judge Kaplan's decision in *Reimerdes*, the Second Circuit implicitly adopted the stricter rule later embraced by the Ninth Circuit in *MDY*.[323] District courts have gone both ways,[324] as have academics and

317. *Id.*

318. *Id.* (quoting S. REP. NO. 105-109, at 12).

319. *MDY Indus.*, 629 F.3d at 950.

320. *Id.* at 958.

321. *See* Murphy v. Millennium Radio Grp. LLC, 650 F.3d 295, 300 n.4 (3d Cir. 2011) (noting circuit split); Ground Zero Museum Workshop v. Wilson, 813 F. Supp. 2d 678, 691 n.6 (D. Md. 2011). The court in *Ground Zero* stated:

> There is disagreement as to whether liability under § 1201(a) can only exist if the circumvention device facilitated infringement. Neither the Fourth Circuit nor any district court within this circuit has addressed the issue. The Federal Circuit has limited liability under § 1201(a) to devices that facilitate infringement. In contrast, the Ninth Circuit has interpreted the statute to create "a distinct anti-circumvention right under § 1201(a) without an infringement nexus requirement."

Ground Zero, 813 F. Supp. at 691 n.6 (citations omitted).

322. *See* MGE UPS Sys., Inc. v. GE Consumer & Indus. Inc., 612 F.3d 760, 765 (5th Cir. 2010) (original opinion); MGE UPS Sys., Inc. v. GE Consumer & Indus. Inc., 622 F.3d 361 (5th Cir. 2010) (substituted opinion).

323. *See* Universal City Studios, Inc. v. Corley, 273 F.3d 429, 443 (2d Cir. 2001) ("[T]he DMCA targets the *circumvention* of digital walls guarding copyrighted material (and trafficking in circumvention tools), but does not concern itself with the *use* of those materials after circumvention has occurred.").

324. *See, e.g.*, Point 4 Data Corp. v. Tri-State Surgical Supply & Equip., Ltd, No. 11-CV-726, 2012 U.S. Dist. LEXIS 113791, at *12–13 (E.D.N.Y. June 13, 2012) (following *MDY*); Agfa Monotype Corp. v. Adobe Sys., Inc., 404 F. Supp. 2d 1030, 1034–36 (N.D. Ill. 2005) (following *Chamberlain*); *see also* Tracfone Wireless, Inc. v. Zip Wireless Prods., Inc., 716 F.

commentators.[325] Since circumventing tethering technology would likely be lawful under *Chamberlain* but unlawful under *MDY* and *Reimerdes*, it is difficult to make any prediction as to how the controversy might fare in court. As the leading treatise on copyright law acknowledges, "the airy drafting of the Digital Millennium Copyright Act . . . has led to a situation in which no perfect judicial construction presents itself."[326] The outcome of any litigation would therefore largely depend upon the forum in which it is brought and whether the parties' claims and defenses are such that the Federal Circuit would have jurisdiction over any appeals.[327]

The profound uncertainty surrounding the DMCA only makes it more likely that video game companies will resort to restrictive licensing as a legal backstop to DRM. Using technological protection in conjunction with restrictive licensing is not uncommon.[328] The two strategies are mutually reinforcing: the DRM technologically enforces license terms,[329] and if the legal protection afforded the DRM fails, the copyright owner may rely on the license in order to punish and prevent circumvention. This "combination of contractual terms and technological measures" can effectively abridge fair use, first sale, and other rights and protections the end user would normally enjoy.[330] This double-layered protection carries an additional advantage: even if circumventing the DRM is deemed lawful,[331] using the product in violation

Supp. 2d 1275, 1284–85 (N.D. Ga. 2010) (citing *Chamberlain* favorably but not clearly embracing its rule).

325. *See, e.g.*, Alessandra Garbagnati, *The Wrath of the Blizz King: How the Ninth Circuit's Decision in* MDY Industries, Inc. v. Blizzard Entertainment *May Slay the Game Genie*, 34 HASTINGS COMM. & ENT. L.J. 313 (2012) (criticizing *MDY*); Raymond Nimmer, N*inth Circuit Rejects* Chamberlain*, Places DMCA Back on a Proper Track*, 16 CYBERSPACE L. 21 (2011) (embracing *MDY*).

326. 3-12A NIMMER & NIMMER, *supra* note 42, § 12A.06[B][7][c][ii].

327. *See* 28 U.S.C. § 1295 (2012) (prescribing Federal Circuit's jurisdiction).

328. Burk, *supra* note 129, at 547.

329. *See id.* at 547–48.

330. Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178, 1202 (Fed. Cir. 2004).

331. It is unclear whether even the Federal Circuit would apply *Chamberlain*'s review to a licensee:

> It is not clear whether a consumer who circumvents a technological measure controlling access to a copyrighted work in a manner that enables uses permitted under the Copyright Act but prohibited by contract can be subject to liability under the DMCA. Because Chamberlain did not attempt to limit its customers [sic] use of its product by contract, however, we do not reach this issue.

Id. at 1202 n.17.

of the license terms may give rise to infringement liability.[332] The benefits of licensing are not lost on video game companies; Sony has been licensing rather than selling PlayStation 3 games for several years.[333] The EULA for a popular line of PlayStation games provides: "THIS SOFTWARE IS LICENSED, NOT SOLD. . . . You agree not to . . . distribute, lease, license, sell, rent, convert into convertible currency, or otherwise transfer or assign the Software."[334] It appears that the age of video game ownership is already drawing to a close.

V. ONE-UP: THE COMING COLLISION—AND POSSIBLE SOLUTION

In *Chamberlain* and *MDY*, the Federal Circuit and Ninth Circuit both intimated that a conflict between the antitrust laws and copyright owners' abrogations of the first sale doctrine may be on the horizon.[335] Video game DRM has already attracted antitrust scrutiny.[336] As protective technology becomes more sophisticated and more aggressive, we can expect to see further antitrust activity in the years ahead. In fact, antitrust may be a uniquely appropriate remedy for preserving first sale. The first sale doctrine first arose in the context of an elaborate price-fixing scheme that itself was

332. MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928, 940 (9th Cir. 2010). In the case of video games and computer software, end users operating under a license do not enjoy the protections of § 117(a)(1) of the Copyright Act, which provides that there is no infringement when "the owner of a copy of a computer program" runs the software, resulting in the creation of a temporary copy in the computer's random access memory ("RAM"). 17 U.S.C. § 117(a)(1) (2012); *see* Vernor v. Autodesk, Inc., 621 F.3d 1102, 1109–11 (9th Cir. 2010). Licensees may only create RAM copies to the extent provided in their license. *Vernor*, 621 F.3d at 1110–11.

333. *See* Tuan Nguyen, *Sony May Prohibit PS3 Preowned Game Sales*, DAILY TECH (May 24, 2006, 7:22 PM), http://www.dailytech.com/Sony+May+Prohibit+PS3+Preowned+Game+Sales/article2512.htm.

334. *Rockstar Games End User License Agreement*, ROCKSTAR GAMES, http://www.rockstargames.com/eula (last revised Oct. 1, 2013). Rockstar Games is responsible for the popular *Grand Theft Auto* franchise. The fine print on packages for such games directs users to read the EULA online.

335. The Federal Circuit construed the DMCA narrowly so as to avoid a conflict with the antitrust laws. *See* Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178, 1191 n.8, 1201–02 (Fed. Cir. 2004). The Ninth Circuit broadly interpreted the DMCA but left open the possibility that a copyright owner might exercise his or her rights under the DMCA in a manner that violates the antitrust laws. *See MDY Indus.*, 629 F.3d at 951 ("If a § 1201(a)(2) defendant in a future case claims that a plaintiff is attempting to enforce its DMCA anti-circumvention right in a manner that violates antitrust law, we will then consider the interplay between this new anti-circumvention right and antitrust law.").

336. *See, e.g.*, Datel Holdings Ltd. v. Microsoft Corp., 712 F. Supp. 2d 974 (N.D. Cal. 2010).

the subject of antitrust litigation. The courts have long recognized the value of competition from secondhand markets and protected against aftermarket monopolization. Because tethering technology allows a console maker to squelch competition in the aftermarket for video game discs, employing this or a related technology could give rise to antitrust liability. However, it is by no means clear such an argument could succeed in court.

A. THE FIRST SALE DOCTRINE'S ANTITRUST ORIGINS

From its birth, the copyright first sale doctrine has been intertwined with antitrust law. *Bobbs-Merrill Co. v. Straus*,[337] the case in which the U.S. Supreme Court first recognized the first sale doctrine, involved a publisher's resale price maintenance scheme—"[t]hat is, the publisher wanted to control the price of a book after ownership passed from the publisher to the wholesaler and from the wholesaler to the retailer."[338] The wholesaler in *Bobbs-Merrill* was aware of the price restriction printed in copies of the book but did not agree to enforce it and, in fact, sold the book for less than the specified price of one dollar.[339] The Supreme Court's opinion fails to mention that this particular vertical price restraint was "part of a larger scheme by publishers and retailers to engage in horizontal price fixing, which was the subject of [an] earlier New York Court of Appeals case."[340] But as the district court recognized, Bobbs-Merrill placed the restrictive notice in its books "as an attempt . . . , as a member of said American Publishers' Association, to enforce as against this defendant the rules of such associations and combination fixing prices, in an effort to maintain them."[341]

Just three years after deciding *Bobbs-Merrill*, the Supreme Court held in *Dr. Miles Medical Co. v. John D. Parks & Sons Co.* that resale price maintenance constitutes a *per se* violation of the Sherman Act.[342] The Court relied on *Bobbs-*

337. Bobbs-Merrill Co. v. Straus, 210 U.S. 339 (1908).

338. Sharon Billington, *Relief From Online Used Book Sales During New Book Launches*, 29 COLUM. J.L. & ARTS 497, 519 (2006); *see supra* text accompanying notes 32–37.

339. *See* Billington, *supra* note 338, at 519.

340. *Id.* at 519–20 (citing Straus v. Am. Publishers' Ass'n, 69 N.E. 1107 (N.Y. 1904)). In 1913, the U.S. Supreme Court heard an antitrust case involving the New York publishers' price-fixing arrangement. *See* Strauss & Strauss v. Am. Publishers' Ass'n, 231 U.S. 222 (1913). The Court held that, like the Patent Act, the Copyright Act was not "intended to authorize agreements in unlawful restraint of trade and tending to monopoly, in violation of the specific terms of the Sherman Law." *Id.* at 234.

341. Bobbs-Merrill Co. v. Straus, 139 F. 155, 178 (S.D.N.Y. 1905), *aff'd*, 210 U.S. 339 (1908).

342. *See* Dr. Miles Med. Co. v. John D. Park & Sons Co., 220 U.S. 373, 405 (1911). The Supreme Court overturned *Dr. Miles* in 2007. *See* Leegin Creative Leather Prods., Inc. v. PSKS, Inc., 551 U.S. 877, 907 (2007) (holding that vertical price restraints should be analyzed under the rule of reason).

Merrill, reasoning that if "no such privilege [to fix prices for downstream sales] exists under the copyright statutes . . . [i]t will hardly be contended, with respect to such a matter, that the manufacturer of an article of commerce, not protected by any statutory grant, is in any better case."[343] As in *Bobbs-Merrill*, the Court found it legally significant that the complainant had transferred ownership over the goods in question: "The agreements are designed to maintain prices, *after the complainant has parted with the title to the articles*, and to prevent competition among those who trade in them."[344] Repeating in part the language it had used just three years earlier, the Court held that "[t]he complainant having sold its products at prices satisfactory to itself, the public is entitled to whatever advantage may be derived from competition in the subsequent traffic."[345] As subsequent courts recognized, the *Bobbs-Merrill* and *Dr. Miles* decisions go hand in hand—not only because they both concern resale price maintenance, but also because the first sale doctrine and competition are intrinsically linked.[346]

343. *Dr. Miles*, 220 U.S. at 405 (quoting Bobbs-Merrill Co. v. Straus, 210 U.S. 339 (1908)).

344. *Id.*, 220 U.S. at 407 (emphasis added).

345. *Id.* at 409; *see also Bobbs-Merrill*, 210 U.S. at 351. In *Bobbs-Merrill*, the Court stated:

> The owner of the copyright in this case *did sell copies of the book in quantities and at a price satisfactory to it*. It has exercised the right to vend. . . . To add to the right of exclusive sale the authority to control all future retail sales, by a notice that such sales must be made at a fixed sum, would give a right not included in the terms of the statute, and, in our view, extend its operation, by construction, beyond its meaning, when interpreted with a view to ascertaining the legislative intent in its enactment.

Id. (emphasis added).

346. *See, e.g.*, Ethyl Gasoline Corp. v. United States, 309 U.S. 436, 456–58 (1940) (citing both *Bobbs-Merrill* and *Dr. Miles* in holding that "by the authorized sales of the fuel by refiners to jobbers the patent monopoly over it is exhausted," and that "[a]greements for price maintenance of articles moving in interstate commerce are, without more, unreasonable restraints of trade within the meaning of the Sherman Act because they eliminate competition"); Waltham Watch Co. v. Keene, 202 F. 225, 233–34 (S.D.N.Y. 1913) (relying on both *Bobbs-Merrill* and *Dr. Miles* in declaring unlawful "an attempt to monopolize and control prices and destroy competition"); *see also* Herbert Hovenkamp, *The Law of Vertical Integration and the Business Firm: 1880–1960*, 95 IOWA L. REV. 863, 889 (2010) ("However, in its *Dr. Miles* decision three years later the Supreme Court cited both the first-sale doctrine and the Sherman Act for the proposition that even an explicit resale-price-fixing agreement between a manufacturer of a patent medicine and a retailer was contrary to legal policy.").

B. ANTITRUST, AFTERMARKETS, AND THE DMCA

In *Chamberlain*, the Federal Circuit's construction of the DMCA was driven largely by a desire to avoid conflict with antitrust law.[347] The court was concerned that Chamberlain's expansive construction, if adopted, "would allow any manufacturer of any product to add a single copyrighted sentence or software fragment to its product, wrap the copyrighted material in a trivial 'encryption' scheme, and thereby gain the right to restrict consumers' rights to use its products in conjunction with competing products."[348] Such a broad interpretation of the DMCA would "allow virtually any company to attempt to leverage its sales into aftermarket monopolies"[349]—that is, monopolies in markets for parts and accessories to the original product.[350] As the Federal Circuit noted, monopolization of aftermarkets violates the Sherman Act.[351] The Ninth Circuit, for its part, skirted the aftermarket issue in *MDY*,[352] but nonetheless left the door open to antitrust liability, stating that it would "consider the interplay between this new anti-circumvention right and antitrust law" if a future defendant "claims that a plaintiff is attempting to enforce its DMCA anti-circumvention right in a manner that violates antitrust law."[353]

Manufacturers of "durable goods"—products that are intended for use over a period of time after purchase, such as washing machines, cars, and computer printers[354]—are "highly protective of the aftermarkets of their goods because these are the most profitable—a manufacturer may sell the initial product at a low cost but mark up the price of the aftermarket goods or services to make a profit."[355] In fact, "[f]irms may even price equipment *below cost* in order to 'buy' market share that will yield profits from high

347. *See* Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178, 1193 (Fed. Cir. 2004) ("Chamberlain's interpretation of the DMCA would . . . grant manufacturers broad exemptions from both the antitrust laws and the doctrine of copyright misuse.").

348. *Id.* at 1201.

349. *Id.* (citing Eastman Kodak Co. v. Image Tech. Servs., 504 U.S. 451, 455 (1992)).

350. *See* Latimer v. Roaring Toyz, Inc., 601 F.3d 1224, 1228 n.2 (11th Cir. 2010) (defining "aftermarket").

351. *Chamberlain*, 381 F.3d at 1201; *see* Eastman Kodak Co. v. Image Tech. Servs., Inc., 504 U.S. 451, 470–78 (1992).

352. *See* MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928, 951 (9th Cir. 2010) ("[T]here is no clear issue of anti-competitive behavior in this case.").

353. *Id.*

354. ROBERT L. SEXTON, EXPLORING ECONOMICS 563–64 (6th ed. 2011).

355. Jennifer Miller, Note, *The Battle Over "Bots": Anti-Circumvention, the DMCA, and "Cheating" at World of Warcraft*, 80 U. CIN. L. REV. 653, 683 (2011).

aftermarket prices."[356] This is precisely what happens in the video game market. Like washing machines, video game consoles are durable goods that are intended for extended use after purchase.[357] Console manufacturers regularly sell video game consoles at or below cost in order to maximize market share, with the expectation that they will realize profits in aftermarkets—most importantly in video game sales (including licensing royalties from third-party game publishers).[358] Incredibly, when the Xbox 360 and PlayStation 3 consoles were first released, Microsoft and Sony each lost hundreds of dollars for each console sold.[359] Because console makers depend on aftermarkets to recoup these initial losses and generate all of their profits, they have a significant incentive to erect barriers to entry in the aftermarkets. DRM can be, and in fact has been, utilized for the purpose of excluding aftermarket competitors in the video game industry.[360] And because competition in primary markets does not necessarily translate to competition in aftermarkets,[361] DRM can be an effective tool for reducing aftermarket competition and increasing profits.[362] The DMCA's protections for DRM only increase the likelihood of competitive harm in the video game market.[363]

C. MONOPOLIZATION IN THE VIDEO GAME AFTERMARKET

Although *Bobbs-Merrill* and *Dr. Miles* concerned vertical price restraints that implicated § 1 of the Sherman Act, which governs restraints on trade

356. Severin Borenstein et al., *Antitrust Policy in Aftermarkets*, 63 ANTITRUST L.J. 455, 463 (1995) (emphasis added).

357. This Author, for example, still owns a functioning Sega Genesis console that he purchased over two decades ago.

358. Thomas Eisenmann et al., *Strategies for Two-Sided Markets*, 84 HARV. BUS. REV. 92, 98 (2006); *see* Dmitri Williams, *Structure and Competition in the U.S. Home Video Game Industry*, 4 INT'L J. ON MEDIA MGMT. 41, 44 (2002) ("The incentive to sell units below cost is created by the need for a large installed user base; since the systems are proprietary, competition for the hearts and minds of consumers is fierce."). There is also a lucrative aftermarket for accessories and add-ons. *See* Datel Holdings Ltd. v. Microsoft Corp., 712 F. Supp. 2d 974, 979 (N.D. Cal. 2010) ("There is a considerable aftermarket for Xbox 360 accessories and add-ons, apart from Xbox 360 games.").

359. Eric Bangeman, *Sony Taking Big Hit on Each PS3 Sold; Xbox 360 in the Black*, ARS TECHNICA (Nov. 16, 2006, 11:33 AM), http://arstechnica.com/gaming/2006/11/8239.

360. *See, e.g.*, *Datel Holdings*, 712 F. Supp. 2d at 985–86 (analyzing technological barriers preventing competition in aftermarket for Xbox 360 add-ons and accessories); SHEFF, *supra* note 88, at 247–50 (explaining that the NES lock-out chip prevented unlicensed game cartridges from playing).

361. Borenstein et al., *supra* note 356.

362. *See* Anupam Chander, *Exporting DMCA Lockouts*, 54 CLEV. ST. L. REV. 205, 206–07 (2006) (positing that DRM allows "a company with a popular product . . . to attract monopoly rents in follow-on goods").

363. *See id.* at 206–10.

between two or more parties, technological limitations on the exercise of first sale rights would more appropriately be analyzed under § 2, which governs monopolization.[364] Federal courts have at times condemned the accumulation of monopoly power as an evil in itself,[365] but in general, the "mere possession of monopoly power, and the concomitant charging of monopoly prices, is not only not unlawful; it is an important element of the free-market system."[366] Section 2 liability does not attach unless the possession of monopoly power is accompanied by anticompetitive or exclusionary conduct.[367]

By limiting consumers' ability to resell durable goods like video game discs, tethering technologies prevent or inhibit the development of a market for used products. As the Copyright Office recognized in its DMCA Section 104 report, "[t]he practice of using technological measures to tether a copy of a work to a particular hardware device" makes reselling the copy "a useless exercise, since the recipient will always receive nothing more than a useless piece of plastic."[368] This allows the original producer to acquire or maintain a monopoly in the market, potentially in violation of § 2.[369]

---

364. *Id.* "The Sherman Act contains a 'basic distinction between concerted and independent action.' " Copperweld Corp. v. Independence Tube Corp., 467 U.S. 752, 767 (1984) (quoting Monsanto Co. v. Spray-Rite Serv. Corp., 465 U.S. 752, 761 (1984)). Section 2 governs the conduct of a single firm, while § 1 reaches conduct undertaken by multiple entities. *Copperweld Corp.*, 467 U.S. at 767–78. Vertical restraints necessarily involve multiple parties, but technological monopolization does not.

365. *See* Berkey Photo, Inc. v. Eastman Kodak Co., 603 F.2d 263, 273 (2d Cir. 1979) ("Because, like all power, it is laden with the possibility of abuse; because it encourages sloth rather than the active quest for excellence; and because it tends to damage the very fabric of our economy and our society, monopoly power is 'inherently evil.' " (quoting United States v. United Shoe Mach. Corp., 110 F. Supp. 295, 345 (D. Mass. 1953), *aff'd per curiam*, 347 U.S. 521 (1954))); Sandeep Vaheesan, *The Evolving Populisms of Antitrust* 21–22 (Feb. 21, 2013) (unpublished manuscript), *available at* http://ssrn.com/abstract_id=2222368.

366. Verizon Commc'ns Inc. v. Law Offices of Curtis V. Trinko, LLP, 540 U.S. 398, 407 (2004).

367. *Id.*; Aspen Skiing Co. v. Aspen Highlands Skiing Corp., 472 U.S. 585, 595–96 (1985).

368. U.S. COPYRIGHT OFFICE, *supra* note 196, at 75–76.

369. *See* Christy Sports, LLC v. Deer Valley Resort Co., Ltd., 555 F.3d 1188, 1198 (10th Cir. 2009) (finding unilateral conduct unlawful under § 2 if it threatens a monopoly); *see also Aspen Skiing Co.*, 472 U.S. at 595–96 ("[T]he offense of monopolization under § 2 of the Sherman Act has two elements: (1) the possession of monopoly power in a relevant market, and (2) the willful acquisition, maintenance, or use of that power by anticompetitive or exclusionary means or for anticompetitive or exclusionary purposes."). The "willful acquisition or maintenance" of monopoly power that § 2 prohibits is "distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident." *Aspen Skiing Co.*, 472 U.S. at 596 n.19 (quoting United States v. Grinnell Corp., 384 U.S. 563, 570–71 (1966)).

Courts and commentators have long recognized the value of secondhand competition.[370] As the leading treatise on antitrust law explains, "[d]urability bears on market power when the second use 'competes' in some significant sense with the first. . . . To this extent, the durable item competes with the new item and limits the power of its producer."[371] In the well-known *Alcoa* decision, Judge Learned Hand gave the following illustration: even though the owner of a copyright enjoys a "lawful monopoly," he "cannot prevent those to whom he sells from reselling at whatever prices they please."[372] Consequently, "[a]t any moment his control over the market will therefore be limited by that part of what he has formerly sold, which the price he now charges may bring upon the market, as second hand or reclaimed articles."[373] The presence of a secondhand market therefore limits a monopolist's ability to charge monopoly prices.[374]

Attempting to suppress a secondhand market may raise antitrust concerns.[375] In *United States v. United Shoe Machinery Corp.*,[376] defendant United Shoe manufactured a variety of machines used for shoemaking, enjoying an

---

370. *See, e.g.*, United States v. Aluminum Co. of Am. (*Alcoa*), 148 F.2d 416, 424–25 (2d Cir. 1945). The Court in *Alcoa* stated:

> At any given moment therefore "secondary" competes with "virgin" in the ingot market; further, it can, and probably does, set a limit or "ceiling" beyond which the price of "virgin" cannot go, for the cost of its production will in the end depend only upon the expense of scavenging and reconditioning. . . . At any moment his control over the market will therefore be limited by that part of what he has formerly sold, which the price he now charges may bring upon the market, as second hand or reclaimed articles.

Id.

371. PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW: AN ANALYSIS OF ANTITRUST PRINCIPLES AND THEIR APPLICATION § 573a (2012).

372. *Alcoa*, 148 F.2d at 425 (citing United States v. Gen. Elec. Co., 272 U.S. 476, 484 (1926)).

373. *Id.*

374. AREEDA & HOVENKAMP, *supra* note 371, § 573b.

375. *See* Barak Y. Orbach, *The Durapolist Puzzle: Monopoly Power in Durable-Goods Markets*, 21 YALE J. ON REG. 67, 108 (2004) ("Courts therefore usually condemn practices that tend to cripple secondhand markets, such as limiting access to necessary replacement parts and lease-only policies." (footnotes omitted)); Michael Waldman, *Antitrust Perspectives for Durable-Goods Markets* 16 (CESifo Working Paper No. 1306, 2004), *available at* http://www.econstor.eu/dspace/bitstream/10419/18671/1/cesifo1_wp1306.pdf ("The antitrust issue here is whether the antitrust authorities should intervene to eliminate practices that serve to make used units unavailable for consumption. The perspective put forth here is that the answer is frequently yes.").

376. United States v. United Shoe Mach. Corp., 110 F. Supp. 295 (D. Mass. 1953), *aff'd per curiam*, 347 U.S. 521 (1954).

approximately eighty-five percent share of the relevant market.[377] The company engaged in a number of business practices to maintain and extend its monopoly, chief among them being a complex leasing system.[378] United's practice was to lease rather than sell its more important machines.[379] This "lease-only system" was exclusionary in the extreme—for example, the leases imposed a lengthy ten-year term—which allowed United to maintain its market power and engage in price discrimination.[380] By not selling its machines outright, United erected barriers to entry[381] and prevented the development of a secondhand market for shoe machinery.[382] To the extent that used machines did make it to market, United bought them back.[383] "It is a fair inference," the court observed, "that United's purpose in acquiring them was to curtail competition from second-hand shoe machinery."[384] Finding that United had violated § 2,[385] the court fashioned a remedy that, among other things, required United to begin offering its machines for sale as well as lease.[386] This remedy ensured that United machines would eventually reach secondhand markets, from which "United will face a type of substitute competition which will gradually weaken the prohibited market power which it now exercises."[387] The Supreme Court ultimately affirmed the district court's decision.[388]

United Shoe suppressed competition through leasing practices, but a monopolist might accomplish the same end through technological means.[389] Of course, "courts are properly very skeptical about claims that competition

377. *Id.* at 297, 307.

378. *See id.* at 314–29 (describing lease system).

379. *Id.* at 344.

380. *Id.*

381. *See id.* at 340 ("United's leases, in the context of the present shoe machinery market, have created barriers to the entry by competitors into the shoe machinery field.").

382. The court stated:

> United's lease system makes impossible a second-hand market in its own machines. This has two effects. It prevents United from suffering that kind of competition which a second-hand market offers. Also it prevents competitors from acquiring United machines with a view to copying such parts of the machines as are not patented, and with a view to experimenting with improvements without disclosing them to United.

Id. at 325.

383. *Id.* at 333–34.

384. *Id.* at 334.

385. *Id.* at 343.

386. *Id.* at 349, 352.

387. *Id.* at 350.

388. United Shoe Mach. Corp. v. United States, 347 U.S. 521 (1954).

389. Foremost Pro Color, Inc. v. Eastman Kodak Co., 703 F.2d 534, 545 (9th Cir. 1983).

has been harmed by a dominant firm's product design changes."[390] Antitrust law not only permits but also encourages monopolists to "compete aggressively on the merits" and to strive to succeed through "the process of invention and innovation."[391] But a monopolist may be on the hook under the Sherman Act if a plaintiff can prove that a technological "innovation" was in fact implemented for anticompetitive reasons, "rather than for improving the operation of the [product]."[392] In *C.R. Bard v. M3 Systems*, for example, the Federal Circuit upheld an antitrust verdict finding that C.R. Bard had modified a patented medical device (a biopsy "gun") not to improve functionality but to "prevent its competitors' non-infringing, flangeless needles from being used in Bard's guns."[393] Similarly, in *United States v. Microsoft*, the D.C. Circuit held that Microsoft's technological integration of its Windows operating system and Internet Explorer web browser was subject to antitrust scrutiny.[394] The government put forth evidence establishing that the integration did not make Internet Explorer more attractive to consumers, but rather discouraged competitors from distributing rival products.[395]

Under these standards, one can easily envision how a monopolization claim against tethering technology might look. The sole purpose of a tethering technology like Disc ID is to reduce competition in an aftermarket. In fact, according to the patent application, the Disc ID system is specifically designed to "suppress" the secondhand market for video games by preventing game discs from being played on multiple consoles.[396] Eliminating competition from secondhand sellers "supports the redistribution of part of proceeds from sales of the electronic content to the developers."[397] In other words, like the lease-only system in *United Shoe*, tethering technology shields the console maker from secondhand "substitute competition"[398]—thereby enhancing corporate profits and cementing the console maker's monopoly status in the aftermarket for games. The patent application does not suggest that this technological "improvement" would enhance the console's

390. United States v. Microsoft Corp., 253 F.3d 34, 65 (D.C. Cir. 2001) (en banc).

391. *Foremost Pro Color, Inc.*, 703 F.2d at 544–45 (9th Cir. 1983) (quoting Berkey Photo, Inc. v. Eastman Kodak Co., 603 F.2d 263, 281 (2d Cir. 1979)).

392. C.R. Bard, Inc. v. M3 Sys., Inc., 157 F.3d 1340, 1382 (Fed. Cir. 1998).

393. *Id.*

394. *Microsoft Corp.*, 253 F.3d at 64–66.

395. *Id.* at 65.

396. Sony Patent App., *supra* note 2, at 2.

397. *Id.*

398. United States v. United Shoe Mach. Corp., 110 F. Supp. 295, 350 (D. Mass. 1953), *aff'd per curiam*, 347 U.S. 521 (1954).

functionality or make video game products more attractive to consumers. Instead, its sole purpose and effect would be to prevent competitors from selling substitute goods in the form of used games.[399] Because the acquisition of monopoly power through tethering technology does not represent legitimate success "as a consequence of a superior product, business acumen, or historic accident," but rather amounts to aftermarket control through "anticompetitive or exclusionary means,"[400] it provides a plausible basis for antitrust liability under § 2 of the Sherman Act.

D. THE RELEVANT MARKET PROBLEM

A business cannot engage in monopolization in violation of § 2 if it does not possess monopoly power in the relevant market.[401] Defining the relevant market is therefore a crucial issue in § 2 cases. The Supreme Court has explained that the relevant market consists of those "commodities reasonably interchangeable by consumers for the same purposes."[402] Interchangeability requires only rough equivalence—that is, "while there may be some degree of preference for one [product] over the other, either would work effectively."[403] For example, under certain circumstances a market for transportation might include not only cars of varying makes and models, but also bicycles.[404] One of the major indicators of interchangeability is "cross-elasticity of demand," that is, the degree to which a rise in the price of one product precipitates greater demand for another.[405] Products belonging to the same market are typically characterized by higher cross-elasticity of demand.[406]

---

399. *Cf.* C.R. Bard, Inc. v. M3 Sys., Inc., 157 F.3d 1340, 1382 (Fed. Cir. 1998). The court in *C.R. Bard* stated:

> Although Bard contended at trial that it modified its Biopty [sic] gun to make it easier to load and unload, there was substantial evidence that Bard's real reasons for modifying the gun were to raise the cost of entry to potential makers of replacement needles, to make doctors apprehensive about using non-Bard needles, and to preclude the use of "copycat" needles.

Id.

400. Aspen Skiing Co. v. Aspen Highlands Skiing Corp., 472 U.S. 585, 596 & n.19 (1985) (quoting United States v. Grinnell Corp., 384 U.S. 563, 570–71 (1966)).

401. Verizon Commc'ns Inc. v. Law Offices of Curtis V. Trinko, LLP, 540 U.S. 398, 407 (2004).

402. United States v. E.I. du Pont de Nemours & Co., 351 U.S. 377, 395 (1956).

403. Allen-Myland, Inc. v. Int'l Bus. Machs. Corp., 33 F.3d 194, 206 (3d Cir. 1994).

404. *Id.*

405. Queen City Pizza, Inc. v. Domino's Pizza, Inc., 124 F.3d 430, 437–38 (3d Cir. 1997) (quoting Tunis Bros. Co. v. Ford Motor Co., 952 F.2d 715, 722 (3d Cir. 1991)).

406. *Id.*, 124 F.3d at 437–38.

In the case of tethering technologies, defining the relevant market so as to support a monopolization claim could be a considerable (and perhaps insurmountable) challenge. For purposes of this Article, let us consider the video game industry as an example. Perhaps the most logical measure of a video game company's monopoly power is the console market, which "represent[s] the mainstream of the video game industry."[407] Although the console market has historically accommodated no more than two or three major players at any given time, it is marked by fierce competition.[408] With the exception of Nintendo, whose NES console dominated the industry in the late 1980s and early 1990s,[409] the market has not seen a true monopolist in the past few decades.[410] The market for the most recent generation of consoles[411] was split more or less evenly: Nintendo accounted for about forty percent of the market, while Sony and Microsoft held about thirty percent each.[412] Even if Nintendo is taken out of the equation—compared to Sony's PlayStation 3 and Microsoft's Xbox 360, Nintendo's Wii utilized inferior hardware, cost less, and targeted different classes of consumers[413]—the console market would have still been split evenly.[414] No competitor could be considered a monopolist in these conditions,[415] as evidenced by the robust competition between the major console makers. The same is true of other industries in which tethering technology might appear, such as the market for copies of Hollywood movies on DVD or Blu-ray discs.[416]

As discussed above, however, the real market of interest is the aftermarket for video game discs. Since this is where the profits lie, console makers try to "lock in" a critical mass of users at the console level and then

407. Williams, *supra* note 358, at 44.

408. *See id.* at 43.

409. SHEFF, *supra* note 88, at 196, 349, 402.

410. Williams, *supra* note 358, at 43–44.

411. That is, Nintendo's Wii, Sony's PlayStation 3, and Microsoft's Xbox 360. The market structure may yet shift in response to these companies' newly released next-generation consoles: Nintendo's Wii U, Sony's PlayStation 4, and Microsoft's Xbox One. As of this writing, it is too early to predict whether or how the market may change.

412. William D'Angelo, *2013 Year on Year Sales and Market Share Update to February 16th*, VGCHARTZ (Feb. 26, 2013), http://www.vgchartz.com/article/250783/2013-year-on-year-sales-and-market-share-update-to-february-16th.

413. *See* Patricio O'Gorman, *Wii: Creating a Blue Ocean the Nintendo Way*, PALERMO BUS. REV., Aug. 2008, at 97, 100–103.

414. *See* D'Angelo, *supra* note 412.

415. AREEDA & HOVENKAMP, *supra* note 371, § 532c (generally market shares below fifty or sixty percent do not constitute monopoly power).

416. There are six major Hollywood studios (often referred to as the "Big Six"): Disney, Fox, Paramount, Sony, Universal, and Warner Bros. *See* Cynthia Littleton, *Majors Prosper on the Margins*, VARIETY, Apr. 16, 2013, at 38.

extract as much profit from them as possible in the aftermarket.[417] Tethering technology aims to reduce competition and promote console makers' monopoly power in this lucrative aftermarket. It therefore makes sense, at least in theory, to define the "relevant market" as the aftermarket for video games rather than as the primary market for gaming consoles. In fact, interchangeability principles support an even more precise definition: the market for video game discs that are compatible with a particular console (e.g., the market for PlayStation 3 games). Games produced for the Wii or Xbox 360 are not generally interchangeable with those produced for the PlayStation 3. This owes, in part, to the fact that many games are only developed for a single gaming console—the popular *God of War* franchise, for example, is exclusive to Sony's consoles, while games featuring Nintendo's mascot Mario are only available for Nintendo consoles. More fundamentally, however, games manufactured for different consoles are not interchangeable because different consoles are not interoperable.[418] If the price of Xbox games increases, consumers who are "locked in" to the Xbox console cannot opt for less-expensive Wii or PlayStation games without buying a new console. The absence of cross-elasticity of demand favors a market definition that is specific to a single console's games.[419]

Defining the relevant market as the aftermarket for video games compatible with a particular console may be plausible under current law. The Supreme Court has held that the relevant market can be an aftermarket for products belonging to a single brand, at least where consumers are "locked in" to equipment purchased in the primary market.[420] *Eastman Kodak v. Image Technical Services* centered on Kodak's policy of selling photocopier replacement parts only to purchasers of Kodak photocopiers who used Kodak repair services.[421] The Court held that the relevant market could be defined as the aftermarket for Kodak-provided parts and services, even though Kodak did not have market power in the photocopier market.[422] This

417. *See* Williams, *supra* note 358, at 43 ("In seeking to create dominant positions for themselves, each firm release[s] a system incompatible with the others. . . . Meanwhile, the consumer [is] forced into an all-or-nothing decision when purchasing a home machine . . . .").

418. *Id.*

419. *See* Michael L. Katz & Carl Shapiro, *Antitrust in Software Markets*, *in* COMPETITION, INNOVATION AND THE MICROSOFT MONOPOLY: ANTITRUST IN THE DIGITAL MARKETPLACE 30, 39 (Jeffrey A. Eisenach & Thomas M. Lenard eds., 1999) ("Thus, application software markets tend to be defined for a given hardware and operating system configuration, or 'platform.' ").

420. Eastman Kodak Co. v. Image Tech. Servs., 504 U.S. 451, 481–82 (1992).

421. *Id.* at 458.

422. *Id.* at 477–78, 481–82.

holding was premised on a theory of "locking in": because most consumers were incapable of, or simply chose not to undertake, the "sophisticated analysis" required to evaluate the total lifetime cost of Kodak equipment, parts, and services, they could not make informed purchasing decisions in the primary equipment market.[423] (The Court referred to this market imperfection as "information costs."[424]) Once consumers had purchased a Kodak photocopier, they were essentially "locked in" to Kodak's parts and services aftermarket.[425] "If the cost of switching is high," the Court reasoned, "consumers who already have purchased the equipment, and are thus 'locked in,' will tolerate some level of service-price increases before changing equipment brands."[426] The combination of high information and switching costs meant that, from the consumer's perspective, "service and parts for Kodak equipment [were] not interchangeable with other manufacturers' service and parts."[427] The Court concluded that it was appropriate to define the relevant market as "only those companies that service Kodak machines"—i.e., Kodak alone.[428]

The video game market presents some of the same characteristics. As in *Kodak*, video game consumers are apt to being "locked in." Consumers are "forced into an all-or-nothing decision" when purchasing a console: they can only play games manufactured for the console they purchase, and are excluded from playing games developed for other consoles.[429] This means that switching costs are high. Consumers who have committed to a particular console will tolerate some level of price increase in the aftermarket for games before investing in a brand new console. Given the gap between the cost of switching consoles and any marginal increase in the price of games—new consoles typically retail for anywhere from $200 to $600, while games go for $50 or $60—video game companies arguably have considerable latitude to increase game prices before customers start jumping consoles.[430]

---

423. *Id.* at 473–76.
424. *Id.* at 477.
425. *Id.* at 476.
426. *Id.*
427. *Id.* at 477, 481–82.
428. *Id.* at 482.
429. *See* Williams, *supra* note 358, at 43.
430. *See Kodak*, 504 U.S. at 476 ("Under this scenario, a seller profitably could maintain supracompetitive prices in the aftermarket *if the switching costs were high relative to the increase in service prices*, and the number of locked-in customers were high relative to the number of new purchasers." (emphasis added)); *see also* Katz & Shapiro, *supra* note 419, at 39. Katz and Shapiro wrote:

> [C]onsumers use applications software in conjunction with specific configurations of hardware and/or operating systems, and often will not

There is precedent for defining the relevant market to include only software compatible with a particular hardware configuration. In the *Microsoft* litigation, for example, one of the relevant markets identified by the district court was the market for "licensing of all *Intel-compatible* PC operating systems."[431] Microsoft argued that non-Intel-compatible operating systems, chiefly Apple's Mac OS, should be included in the relevant market.[432] The district court found, and the appellate court agreed, that high switching costs militated against including non-Intel-compatible systems in the definition. In the district court's words, there were, at the time, "no products . . . that a significant percentage of computer users worldwide could substitute for [Intel-compatible operating systems] without incurring substantial costs."[433] Mac OS was not interchangeable with Microsoft's Intel-compatible Windows program because most "consumers would not switch from Windows to Mac OS in response to a substantial price increase because of the costs of acquiring the new hardware needed to run Mac OS (an Apple computer and peripherals) and compatible software applications."[434] Under this reasoning, the relevant market in a § 2 challenge to tethering technology might also be defined in hardware-specific terms. Just as Microsoft unquestionably possessed monopoly power in the market for Intel-compatible operating systems (Windows had a ninety-five percent market share),[435] so too does a video game console maker enjoy monopoly status in the market for games that are compatible with its console.[436]

Most courts, however, maintain that high switching costs are insufficient to justify a narrow market definition like that in *Kodak* and *Microsoft*. As one district court explains, aftermarket monopolization is only possible "when there are 'significant information *and* switching costs,' " which sever the link

> purchase new hardware or a new operating system in response to modest shifts in application software prices. In part, this results from the fact that consumers have already purchased the hardware and operating system prior to buying specific pieces of software.

Id.

431. United States v. Microsoft Corp., 253 F.3d 34, 52 (D.C. Cir. 2001) (emphasis added).

432. *Id.*

433. *Id.*

434. *Id.*

435. *Id.* at 54–56.

436. For example, all PlayStation games must be produced by Sony or a third party licensed by Sony. *See* Williams, *supra* note 358, at 44.

between the primary market and the aftermarket.[437] This is intuitive. In *Kodak*, it was consumers' inability to accurately assess lifetime photocopier costs (inclusive of equipment, parts, and service) that made them "locked in."[438] Had information costs been lower, consumers could have accounted for Kodak's aftermarket costs in deciding whether to purchase a Kodak photocopier in the first place.[439] By the same token, had information costs been high and switching costs low, then Kodak's customers could have painlessly switched to a competitor's equipment once they encountered high prices for parts and service. It is the *combination* of product uniqueness, switching costs, and information costs that "generate[s] market power even as [it] delineate[s] the boundaries of the market in which such power is exercised."[440]

Other courts have characterized this relationship in terms of the primary market's interplay with the aftermarket. The First Circuit instructs that "a court may conclude that the aftermarket is the relevant market for antitrust analysis only if evidence supports an inference of monopoly power in the aftermarket *that competition in the primary market appears unable to check*."[441] Where information costs or switching costs (or both) are low, the Third Circuit states, "aftermarket behavior generally is disciplined by competition in the primary product market."[442] As the Supreme Court lucidly explained in *Kodak*, aftermarket monopolization is necessarily tempered by the degree to which consumers can make informed decisions in the primary market or easily switch brands to avoid aftermarket gouging.[443]

This is why an aftermarket monopolization claim against a video game console maker may be doomed to fail, at least under current market conditions. Even though tethering technology would probably cause secondhand dealers to be excluded from the aftermarket for games, and even though switching costs are high, available evidence suggests that information costs are comparatively low. In contrast to *Kodak*, prospective game console

437. Static Control Components, Inc. v. Lexmark Int'l, Inc., 487 F. Supp. 2d 861, 882–83 (E.D. Ky. 2007) (quoting Eastman Kodak Co. v. Image Tech. Servs., 504 U.S. 451, 473 (1992)) (emphasis added).

438. *Kodak*, 504 U.S. at 473–76.

439. *See id.*

440. ID Sec. Sys. Canada, Inc. v. Checkpoint Sys., Inc., 249 F. Supp. 2d 622, 641 (E.D. Pa. 2003).

441. SMS Sys. Maint. Servs., Inc. v. Digital Equip. Corp., 188 F.3d 11, 17 (1st Cir. 1999) (emphasis added).

442. Harrison Aire, Inc. v. Aerostar Int'l, Inc., 423 F.3d 374, 381–82 (3d Cir. 2005).

443. *Kodak*, 504 U.S. at 473–77; *see ID Sec. Sys.*, 249 F. Supp. 2d at 641–43, 647–48 (interpreting *Kodak*).

purchasers have ready access to information about aftermarket costs.[444] Video game pricing is highly visible, easy to comprehend, and not usually subject to wide fluctuation. Moreover, gamers readily understand the implications of technology that precludes the use of secondhand video games. As GameStop CFO Rob Lloyd has claimed, internal market research found that approximately sixty percent of consumers would not buy a new console if it were incapable of playing pre-owned games.[445] In other words, information about a console maker's aftermarket behavior directly impacts consumers' decisions in the primary market. GameStop's research bears out the conclusion that "the high switching costs associated with [video game consoles] are significantly counterbalanced by information costs so low as to be almost nonexistent, a fact that increases a consumer's ability to make an intelligent choice when choosing the company with which it will have a long relationship."[446] A court might therefore be hesitant to conclude that a console-specific aftermarket for video games would be an appropriate market definition.[447]

The difficulty of defining the relevant market so as to establish that any video game company in today's market possesses monopoly power could be fatal to a § 2 claim. Therefore, while antitrust law offers a potential solution to the dilemma posed by tethering technology, consumers' recourse is far from clear. The Sherman Act prizes consumer welfare above all else,[448] but it will not intervene where consumer preference is sufficient to remedy market abuses.[449] In durable goods markets that are dominated by a true monopolist, or in which switching and information costs are both high, the Sherman Act might dictate a different outcome. One can only hope that, where antitrust law falls short, consumers will have the good sense to communicate their interests through purchasing decisions.

---

444. *See Kodak*, 504 U.S. at 473 ("Much of this information is difficult—some of it impossible—to acquire at the time of purchase.").

445. Brian Arnold, *Majority of Gamers Would Not Buy Console That Blocks Used Games*, EXAMINER.COM (Feb. 12, 2013), http://www.examiner.com/article/majority-of-gamers-would-not-buy-console-that-blocks-used-games.

446. ID Sec. Sys. Canada, Inc. v. Checkpoint Sys., Inc., 249 F. Supp. 2d 622, 643 (E.D. Pa. 2003).

447. *See* SMS Sys. Maint. Servs., Inc. v. Digital Equip. Corp., 188 F.3d 11, 15 (1st Cir. 1999) ("Put another way, a court may conclude that the aftermarket is the relevant market for antitrust analysis only if the evidence supports an inference of monopoly power in the aftermarket that competition in the primary market appears unable to check.").

448. Vaheesan, *supra* note 365, at 33–39.

449. *See* Abbott Labs. v. Teva Pharm. USA, Inc., 432 F. Supp. 2d 408, 421 (D. Del. 2006) ("If consumers are free to choose among products, then the success of a new product in the marketplace reflects consumer choice, and 'antitrust should not intervene when an invention pleases customers.' " (quoting AREEDA & HOVENKAMP, *supra* note 371, § 776d)).

VI. CONCLUSION: GAME OVER?

Revisions to the first sale doctrine are not to be taken lightly. Narrowing the doctrine, even modestly, amounts to "a fundamental change in one of the main tenets of copyright law,"[450] and upsets the "traditional copyright bargain" between copyright holders and personal property owners.[451] The first sale doctrine benefits consumers both directly and indirectly: not only does the doctrine facilitate the free alienability of personal property, minimizing transaction costs and obviating the need to negotiate with copyright owners for resale or transfer rights,[452] it also promotes competitive pricing in the marketplace.[453] The Supreme Court has recognized the first sale doctrine's importance to consumers and competition for over a century.[454] In its most recent first sale decision, the Court emphasized that "competition, including freedom to resell, can work to the advantage of the consumer."[455]

This is precisely why the content industry's relationship with first sale is so troubling. For the past three decades, copyright-heavy corporations have chipped away at the doctrine through lobbying, litigation, and licensing. Tethering technology threatens to further undermine first sale by physically preventing users from exercising their first sale rights. More broadly, industry's embrace of tethering technology threatens to stymie competition and enhance copyright owners' monopolies. These technological barriers are themselves backed by the force of law. The marriage of digital technology and strict anticircumvention rules could effect a de facto repeal of the first sale doctrine.

The irony is that even though "the Copyright Act and the Sherman Act [are] both designed ultimately to improve the welfare of consumers in our free market system,"[456] neither body of law provides a clear solution to the problems posed by tethering technology. In the case of video games, consumers could be held liable under the DMCA for circumventing a system like Disc ID, and the Sherman Act might not prevent content producers

450. *Software Rental Amendments of 1990: Hearing Before the Subcomm. on Courts, Intellectual Property and the Admin. of Justice of the H. Comm. on the Judiciary*, 101st Cong. 2 (1990) (statement of Rep. Robert Kastenmeier) (quoting Professor David Lange).

451. Brilliance Audio, Inc. v. Haights Cross Commc'ns, Inc., 474 F.3d 365, 374 (6th Cir. 2007) (construing statutory exception to first sale doctrine narrowly so as to protect the copyright "bargain").

452. Pearson Educ., Inc. v. Liu, 656 F. Supp. 2d 407, 410 (S.D.N.Y. 2009).

453. Ghose et al., *supra* note 60, at 346.

454. *See* Bobbs-Merrill Co. v. Straus, 210 U.S. 339 (1908).

455. Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351, 1363 (2013).

456. Data Gen. Corp. v. Grumman Sys. Support Corp., 36 F.3d 1147, 1187 (1st Cir. 1994).

from technologically decimating the secondhand video game industry. This Article demonstrates that there is a way through the copyright and antitrust thicket, but also acknowledges that doctrinal challenges could preclude a consumer-friendly resolution in court. Ultimately, it may devolve upon consumers to inform themselves and vote with their wallets. Otherwise, it really could be "Game Over" for first sale.

WHY NOT PRIVACY BY DEFAULT?

Lauren E. Willis[†]

ABSTRACT

We live in a Track-Me world, one from which opting out is often not possible. Firms collect reams of data about all of us, quietly tracking our mobile devices, our web surfing, and our email for marketing, pricing, product development, and other purposes. Most consumers both oppose tracking and want many of the benefits tracking can provide. In response, policymakers have proposed to give consumers significant control over when, how, and by whom they are tracked through a system of defaults (i.e., "Track-Me" or "Don't-Track-Me") from which consumers can opt out.

The use of a default scheme is premised on three assumptions. First, that for consumers with weak or conflicted preferences, any default chosen will be "sticky," meaning that more consumers will stay in the default position than would choose it if reaching the position required an affirmative action. Second, that those consumers with a true preference for the opt-out position—and only those consumers—will opt out. Third, that where firms oppose the default position, convincing consumers to opt out will require the firms to explain the default position, resulting in well-informed decisions by consumers.

This Article argues that for tracking defaults, these assumptions may not consistently hold. Past experience with the use of defaults in policymaking teaches that Track-Me defaults are likely to be too sticky, Don't-Track-Me defaults are likely to be too slippery, and neither are likely to result in well-educated consumers.

These conclusions should inform the "Do-Not-Track" policy discussions now taking place in the United States, in the European Union, and at the World Wide Web Consortium. They also cast doubt on the privacy and behavioral economics literatures that advocate the use of "nudges" to improve consumer decisions about privacy.

© 2014 Lauren E. Willis.

† Robert S. Braucher Visiting Professor of Law, Harvard Law School and Professor of Law, Loyola Law School Los Angeles, lauren.willis@lls.edu. My thanks to Omri Ben-Shahar, James Grimmelmann, Chris Hoofnagle, Paul Ohm, Jules Polonetsky, Paul Schwartz, Chris Soghoian, Lior Strahilevitz, Jessamyn West, Jonathan Zittrain, participants at the 2013 Privacy Law Scholars Conference, research assistant Natalie Kim, and the patient editors at the *Berkeley Technology Law Journal*.

TABLE OF CONTENTS

I. **INTRODUCTION** 64

II. **WHY AND WHEN DEFAULTS ARE STICKY** 69

A. MECHANISMS THAT MAKE DEFAULTS STICKY 72
1. *Transaction Barriers* 72
2. *Judgment and Decision Biases* 74
3. *Preference-Formation Effects* 77
B. CONDITIONS UNDER WHICH DEFAULTS ARE OFTEN STICKY 78

III. **DEFAULTS IN THEORY** 80

A. THE THEORY BEHIND THE USE OF DEFAULTS IN POLICYMAKING 80
1. *Policy Defaults* 81
2. *Penalty Defaults* 82
3. *Altering and Framing Rules* 82
4. *An Aside on Forced Choice* 84
B. TRANSLATING DEFAULT THEORY TO TRACKING POLICY 85
1. *Scope of Tracking Defaults* 86
2. *Setting: Track-Me or Don't-Track-Me* 88
3. *Granularity of Available Opt-Out Positions* 90
4. *Altering and Framing Rules* 92

IV. **DEFAULTS IN PRACTICE** 95

A. TWO FAILED DEFAULT SCHEMES 96
1. *The Financial Information Default Scheme* 96
2. *The Checking Account Overdraft Default Scheme* 97
B. HOW FIRMS MAKE THESE DEFAULTS FAIL 99
1. *Transaction Barriers* 99
2. *Judgment and Decision Biases* 102
3. *Preference Formation Effects* 105
C. SUCCESSFUL DEFAULTS 107
1. *Automatic Enrollment in Retirement Savings Plans* 107
2. *Do Not Call* 108
D. CRACKS IN THE THEORY BEHIND THE USE OF DEFAULTS 109
1. *Conditions Where Defaults Do Not Work* 109
2. *Reexamining the Logic Supporting the Use of Defaults in Policymaking* 110

V. **WHY TRACKING DEFAULTS ARE UNLIKELY TO ACHIEVE POLICYMAKERS' GOALS** 111

A. HOW FIRMS COULD MAKE TRACKING DEFAULTS FAIL 112
1. *Erect, Eliminate, or Invert Transaction Barriers* 112

a) Costs 112
b) Confusion 114
c) Futility 114
2. *Harness Judgment and Decision Biases* 115
a) Salience effects 115
b) Omission Bias 116
c) Loss Aversion and the Endowment Effect 116
d) Procrastination and Decision Avoidance 117
e) Excessive Discounting 117
f) Choice Bracketing 118
g) The Illusion of Control 118
h) The Sunk Costs Fallacy 119
3. *Bolster, Undermine, or Reverse Preference Formation Effects* 120
a) The Endorsement Effect 120
b) The Experience Effect 120
B. ALTERING RULES, FRAMING RULES, AND COMPETITION 121
1. *The Limits of Altering and Framing Rules* 122
a) Respect for Heterogeneous Preferences 122
b) Mis-sorting 124
c) Commercial Speech Doctrine Limits 125
d) The First and Last Mover 126
2. *Will Competition Change the Calculus?* 128
VI. **CONSEQUENCES OF PRIVACY POLICY BY DEFAULT** 130

I. INTRODUCTION

We live in a Track-Me[1] world, one from which opting out is, as a practical matter, often not possible. Firms collect reams of data about us for marketing, pricing, product development, and other uses. Sometimes we knowingly participate in the first stage of this process—a firm asks for information, and we provide it, knowing roughly how that firm will use it. But much data collection is passive, invisible, and performed without explicit consent. Most of us know little about downstream users to whom our data may be transferred and none of us can know all the future uses to which our data may be put. Collection of data on the websites we visit, the content of our emails, and the movements of our mobile phones have garnered the most media attention,[2] but as tracking technologies (e.g., facial recognition programs, eye tracking systems, and geolocation sensors) become cheaper and more accurate, the amount of data collected passively and the uses to which it will be put will only increase.[3]

Although preferences for information privacy vary, wide majorities of people both express opposition to the extent of this data collection and have taken some steps to avoid being tracked.[4] They wish to avoid uses of their

1. This Article uses "Track-Me" and "Don't-Track-Me" to avoid the indeterminacy of "Opt-In" and "Opt-Out," and uses "tracking" in a colloquial sense to refer to all forms of personal data collection and use for commercial purposes, online and off. For a discussion of "Do-Not-Track" as a technical protocol, see DO NOT TRACK—UNIVERSAL WEB TRACKING OPT OUT, http://donottrack.us (last visited Aug. 1, 2013). Collection and use of personal data for law enforcement purposes is beyond the scope of this Article.

2. Most websites track users, collecting information such as access time, visit duration, mouse movements, and clicks. *See* Andrew Couts, *Top 100 Websites: How They Track Your Every Move Online*, DIGITAL TRENDS (Aug. 30, 2012), http://www.digitaltrends.com/web/top-100-websites-how-are-they-tracking-you (finding that the top 100 websites all track users in some way). Some providers of "free" email scan users' email text. *See, e.g.*, John Pallatto, *Google Defends Scanning E-Mail for Ad Links*, EWEEK (Apr. 23, 2004), http://www.eweek.com/c/a/Messaging-and-Collaboration/Google-Defends-Scanning-EMail-for-Ad-Links (explaining that Google scans email text of Gmail users for targeted ad purposes). Mobile data is often tracked through applications or by cell carriers themselves, then sold to third parties. *See, e.g.*, Olga Kharif & Scott Moritz, *Cell Carriers Sell Users' Tracking Data in $5.5 B Market*, DELAWAREONLINE (June 13, 2013), http://www.delawareonline.com/article/20130613/BUSINESS08/306130050/Cell-carriers-sell-users-tracking-data-5-5-B-market (reporting on multi-billion dollar market for cell phone tracking data).

3. *See generally What They Know*, WALL ST. J., http://online.wsj.com/public/page/what-they-know-digital-privacy.html (last visited Aug. 1, 2013) (providing an overview of current news and issues regarding Internet tracking).

4. *See, e.g.*, LEE RAINIE ET AL., ANONYMITY, PRIVACY, AND SECURITY ONLINE, PEWRESEARCHCENTER REP. 8 (Sept. 5, 2013), http://pewinternet.org/~/media//Files/Reports/2013/PIP_AnonymityOnline_090513.pdf (finding that eighty-six percent of internet users have "taken at least one step to try to mask their behavior or avoid being

data that they experience as harmful (e.g., identity theft, price discrimination, negative employment consequences) as well as the more amorphous costs of a lack of privacy (e.g., the "creepy" feeling of being watched, personal embarrassment and social stigma, decreased space for individual experimentation and reflection).[5] Yet, people want the benefits that tracking can provide, such as online socializing and access to online content.[6]

In response, policymakers have proposed that firms provide consumers with significant control over when, how, and by whom they are tracked through a system of defaults (i.e., "Track-Me" or "Don't-Track-Me") from which consumers can opt out.[7] Ostensibly, this "notice and choice" regime is motivated by a desire to satisfy heterogeneous privacy preferences. Such an approach has deep normative roots; privacy itself has been conceptualized as a right of individuals "to determine for themselves when, how, and to what extent information about them is communicated to others."[8] Defaults as privacy policy also relieve policymakers of some of the contested judgment

tracked"); Chris Crum, Googler: Nobody Wants to Be Tracked Online, WEBPRONEWS (Apr. 2, 2012), http://www.webpronews.com/googler-nobody-wants-to-be-tracked-online-2012-04 (finding that nearly eighty-five percent of users think a business should not be able to anonymously track consumer activity on the business's website).

5. *See, e.g.*, Mark Memmott, *Orbitz Shows Mac Users Pricier Hotel Options: Big Deal or No Brainer?*, THE TWO-WAY: BREAKING NEWS FROM NPR (June 26, 2012), http://www.npr.org/blogs/thetwo-way/2012/06/26/155756095/orbitz-shows-mac-users-pricier-hotel-options-big-deal-or-no-brainer (showing that 85% of readers said they "[w]ould . . . be upset to learn [they] had been shown pricier options just because [they] used a Mac to search for a hotel"); ANITA ALLEN, UNPOPULAR PRIVACY: WHAT MUST WE HIDE? (2011) (arguing that privacy is often necessary for personal dignity, trust, and reputation, all of which preserve freedom to make one's own social, economic, and political choices).

6. *See, e.g.*, Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63, 64 (2012), http://www.stanfordlawreview.org/sites/default/files/online/topics/64-SLRO-63_1.pdf (listing significant benefits currently produced by tracking).

7. *See, e.g.*, Do-Not-Track Online Act of 2013, S. 418, 113th Cong. (2013) (proposing legislation that would require the FTC to create an enforceable "mechanism by which an individual can simply and easily indicate whether the individual prefers to have personal information collected by providers of online services, including by providers of mobile applications and services"); THE WHITE HOUSE, CONSUMER DATA PRIVACY IN A NETWORKED WORLD 11–22 (2012) (proposing a Consumer Privacy Bill of Rights, which includes: "Consumers have a right to exercise control over what personal data companies collect from them and how they use it.").

8. ALAN F. WESTIN, PRIVACY AND FREEDOM 7 (1967). This focus on individual choice may be misplaced, regardless of whether meaningful choice is possible. Social welfare and individual preferences about privacy may not be well aligned, and if so, policymakers ought to make policy based on the social costs and benefits of tracking rather than based on a quest to satisfy preferences. But that debate is beyond the scope of this Article.

calls presented by the social costs and benefits of tracking, leaving each consumer to make an individual cost-benefit tradeoff for herself.[9]

The use of a default scheme from which consumers can opt out is premised on three key assumptions, taken from behavioral economics literature.[10] First, that any default will be "sticky," meaning that more consumers stay with the default than would explicitly choose to do so if forced to make a choice.[11] Second, that those consumers with a true preference for the opt-out position—and only those consumers—will opt out.[12] Third, that where a firm opposes the default position, the firm will be forced to explain it in the course of trying to convince consumers to opt out, resulting in well-informed decisions by consumers.[13] In behavioral economics

9. *Cf.* JAMES P. NEHF, OPEN BOOK: THE FAILED PROMISE OF INFORMATION PRIVACY IN AMERICA 103–04 (2012) (explaining difficulty policymakers have with resolving the incommensurability of the costs and benefits of privacy).

10. A fourth key assumption is that defaults imposed by law can and will be enforced. These assumptions may be naive, as experiences in the European Union, in Israel, and with the Children's Online Privacy Protection Act ("COPPA") in the United States, suggest. *See, e.g.*, Maurizio Borghi et al., *Online Data Processing Consent Under EU Law: A Theoretical Framework and Empirical Evidence from the UK*, 21 INT'L. J. L. & INFO. TECH. 109 (2013) (finding widespread noncompliance with E.U. and U.K. law requiring user consent prior to data collection); Michael Birnhack & Niva Elkin-Koren, *Does Law Matter Online? Empirical Evidence on Privacy Law Compliance*, 17 MICH. TELECOMM. & TECH. L. REV. 337 (2011) (finding widespread noncompliance with Israeli law requiring notice to users about data collection); EUROPEAN NETWORK AND INFO. SEC. AGENCY, PRIVACY CONSIDERATIONS OF ONLINE BEHAVIOURAL TRACKING 16 (2012), http://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/privacy-considerations-of-online-behavioural-tracking (finding that EU law requiring user consent prior to tracking is not being enforced by the EU or member countries); danah boyd et al., *Why Parents Help Their Children Lie to Facebook About Age: Unintended Consequences of the 'Children's Online Privacy Protection Act*,*'* 16 FIRST MONDAY 11 (2011), http://firstmonday.org/ojs/index.php/fm/article/view/3850/3075 (finding widespread circumvention of COPPA's default that websites cannot collect information about children unless their parents opt out). To present the strongest case for tracking defaults, this Article assumes they would be enforceable and enforced, but nonetheless finds that tracking defaults are unlikely to achieve policymakers' professed aims.

11. *See, e.g.*, Cass R. Sunstein, Impersonal Default Rules vs. Active Choices vs. Personalized Default Rules: A Triptych 9 (May 19, 2013) (unpublished manuscript), *available at* http://ssrn.com/abstract_id=2171343 ("In the domain of privacy on the Internet, a great deal depends on the default rule.").

12. *See, e.g.*, Jay P. Kesan & Rajiv C. Shah, *Setting Software Defaults: Perspectives from Law, Computer Science and Behavioral Economics*, 82 NOTRE DAME L. REV. 583, 633 (2006) (suggesting penalty defaults for privacy settings to protect uninformed users yet allow "well-informed individuals" to opt out).

13. *See, e.g.*, Lilian Edwards & Ian Brown, *Data Control and Social Networking: Irreconcilable Ideas?*, *in* HARBORING DATA: INFORMATION SECURITY, LAW AND THE CORPORATION 202, 225 (A. Matwyshyn ed., 2009) ("A general rule that privacy settings be set at the most privacy-friendly setting when a profile is first set up . . . would inform all users that privacy

parlance, tracking defaults are expected to be sticky "policy defaults" (intended to nudge people with weak or unformed preferences toward the default position)[14] or information-forcing "penalty defaults" (intended to educate people about the default and opt-out positions).[15]

This Article demonstrates one reason why tracking defaults will not necessarily nudge consumers to the right position or help consumers meet their own well-informed preferences: the assumptions underlying the use of defaults in policymaking are unlikely to hold in the personal-data-tracking context. Defaults can be too sticky (meaning that consumers stick with the default even though they would opt out were they well-informed) or too slippery (meaning that consumers opt out even though they would prefer the default position were they well-informed), and defaults are not always information-forcing. Firms surround defaults they favor with a powerful campaign to keep consumers in the default position, but meet defaults set contrary to their interests with an equally powerful campaign to drive consumers to opt out. Rather than giving firms an incentive to facilitate consumer exercise of informed choice, many defaults leave firms with opportunities to play on consumer biases or confuse consumers into sticking with or opting out of the default.

Thus, whether a tracking default is sticky or slippery, informative or uninformative will depend on whether firms' interests are aligned with the default. To counter firm manipulation, the law can impose altering rules[16]—rules governing the process by which consumers can opt out—and framing rules[17]—rules governing the presentation of the default to consumers. But normative, legal, and practical constraints limit altering and framing rules, and even the strongest of such rules can be outmaneuvered by firms with the means and motivation to do so.

Unless robust competition over protecting consumer privacy develops in the marketplace—a doubtful prospect—firms will generally prefer for

settings do exist, and force them to learn how to make use of them before they moved on to networking . . . .").

14. *See, e.g.*, Craig R. M. McKenzie et al., *Recommendations Implicit in Policy Defaults*, 17 PSYCHOL. SCI. 414, 414–15 (2006) (defining "policy default" in the context of organ donation).

15. Ian Ayres & Robert Gertner, *Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules*, 99 YALE L.J. 87, 91 (1989); *see also* Edward J. Janger & Paul M. Schwartz, *The Gramm-Leach-Bliley Act, Information Privacy, and the Limits of Default Rules*, 86 MINN. L. REV. 1219, 1239 (2001) (dubbing penalty defaults "information-forcing defaults").

16. *See* Ian Ayres, *Regulating Opting Out: An Economic Theory of Altering Rules*, 121 YALE L.J. 2032 (2012).

17. *See* Elizabeth F. Emens, *Changing Name Changing: Framing Rules and the Future of Marital Names*, 74 U. CHI. L. REV. 763, 840 (2007).

consumers to be in the Track-Me position. Because firms can influence people's responses to tracking defaults, most Track-Me defaults will be too sticky and many Don't-Track-Me defaults will be too slippery. Further, neither the consumers who remain in Track-Me defaults nor those who opt out of Don't-Track-Me defaults will necessarily be making well-informed decisions. Personal-data-tracking defaults, therefore, are unlikely to facilitate the satisfaction of heterogeneous consumer preferences or produce an informed resolution of the conflict between people's desire for information privacy and their desire for the benefits produced by their data.

Privacy law scholars have been skeptical of the idea that a notice-and-choice regime could produce robust individual decision making about personal data privacy.[18] Yet their critiques have not confronted, and have even equivocated in the face of, the assumptions about defaults made by the standard behavioral economics literature.[19] Some scholars continue to advocate defaults, but on norms-setting grounds rather than as policy defaults or penalty defaults.[20] For defaults to set these norms, however, the signals that defaults provide must be clear. Experiences with other default rules give reason to think that firms will surround tracking defaults with a great deal of noise—noise calculated to confuse the signal.

This Article fills the current intellectual gap with an understanding of how firms' dynamic responses to defaults can undermine policymakers' aims. It thus exposes the limits of defaults as policy tools. Along the way, this Article provides several new insights into defaults: that nearly any bias might lead a consumer to stay in a default position, not just those biases typically

18. The classic article on personal information tracking defaults is Janger & Schwartz, *supra* note 15. *See also, e.g.*, Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1609, 1660–63, 1681–85 (1999) (critiquing the individual choice model of privacy); Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013) (same).

19. *See, e.g.*, Schwartz, *supra* note 18, at 1686–87 (suggesting that a default of minimal data disclosure would allow individuals to "personalize their privacy levels"); Paul M. Schwartz, *Property, Privacy, and Personal Data*, 117 HARV. L. REV. 2055, 2100 (2004) ("This Article prefers an opt-in default because . . . it would place pressure on the better-informed party to disclose material information about how personal data will be used. This default promises to force the disclosure of hidden information about data-processing practices."); Solove, *supra* note 18, at 1900 ("[P]rivacy self-management should not be abandoned. . . .").

20. *See, e.g.*, Solove, *supra* note 18, at 1903 ("The law should develop and codify basic privacy norms . . . . [I]t can be in a form like the Uniform Commercial Code (UCC), where certain default rules can be waived."); Omer Tene & Jules Polonetsky, *To Track or "Do Not Track": Advancing Transparency and Individual Control in Online Behavioral Advertising*, 13 MINN. J. L. SCI. & TECH. 281, 341 (2012) (suggesting that tracking defaults be used to signal and effectively set social norms about what information should and should not be shared, with whom, and under what conditions).

associated with defaults; that the theory underlying the use of defaults in policymaking fails to address how policymakers should set the scope of defaults and the granularity of opt-out choices; and that the faith in the potential for altering and framing rules to fine-tune the stickiness of defaults is misplaced. Finally, this Article casts doubt on the use of "nudges"[21] in policymaking to help people make better choices about information privacy.[22] While Track-Me and Don't-Track-Me defaults might pave the political path to a better system for regulating personal data tracking, they might also defuse the political will to implement better privacy regulation.

This Article proceeds as follows: Part II describes why and when defaults are sticky. Part III explains the theory behind the use of defaults in policymaking and how policymakers appear poised to translate these theories to the personal-data-tracking arena. Part IV examines the use of defaults in other contexts to assess when and why defaults in practice do and do not function as theorized. Part V considers how tracking defaults can be expected to play out, predicting that Track-Me defaults are likely to be overly sticky and Don't-Track-Me defaults are likely to be overly slippery. Part V also explains why altering rules, framing rules, and competition have limited potential to change these dynamics. Part VI concludes with an exploration of the potential political consequences of using defaults in information privacy policy.

II. WHY AND WHEN DEFAULTS ARE STICKY

A default is a setting or position that has been preselected, but can be altered. For example, many email programs by default have an audible

21. *See* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 6 (2008) (defining a "nudge" as a policy tool that "alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives").

22. A burgeoning literature advocates the use of nudges to encourage people to make better privacy decisions. *See, e.g.*, Rebecca Balebako et al., *Nudging Users Towards Privacy on Mobile Devices*, *in* PROCEEDINGS OF THE 2ND INTERNATIONAL WORKSHOP ON PERSUASION, INFLUENCE, NUDGE & COERCION THROUGH MOBILE DEVICES (2011); M. Ryan Calo, *Against Notice Skepticism in Privacy (and Elsewhere)*, 87 NOTRE DAME L. REV. 1027 (2012); Yang Wang et al*., Privacy Nudges for Social Media: An Exploratory Facebook Study*, Second International Workshop on Privacy and Security in Online Social Media (2013), *available at* http://precog.iiitd.edu.in/events/psosm2013/9psosm6-wang.pdf; Alessandro Acquisti, *Nudging Privacy: The Behavioral Economics of Personal Information*, 7 IEEE 82 (2009).

notification for incoming mail, but the user can turn off the notification or change the sound.[23]

Default settings by definition can be changed, yet few people change them. Tracking consumer activity on the Internet today presents a rough example. Many websites, mobile phones, applications ("apps"), and other devices and programs that can facilitate or inhibit tracking are preset to allow tracking today.[24] Many of these settings are not full-fledged defaults because opting out is not always possible; some devices and programs cannot be used without tracking enabled, and some trackers track consumers even when program or device settings are in the "Do-Not-Track" position.[25] But consumers can opt out of some tracking, at least partially, such that "Track-Me" is a quasi-default today. Consumers can delete certain types of cookies from their browsers or install software that blocks some Internet and cellphone tracking.[26] Consumers can also change tracking options on their

23. *See, e.g.*, *Customize Your AOL Sounds*, AOL HELP, http://help.aol.com/help/microsites/microsite.do?cmd=displayKC&docType=kc&externalId=217059 (last visited Dec. 1, 2013).

24. *See, e.g.*, Claire Cain Miller, *How to Opt Out of Google's Plan to Use Your Name and Comments in Ads*, N.Y. TIMES (Oct. 14, 2013), http://bits.blogs.nytimes.com/2013/10/14/how-to-opt-out-of-googles-plan-to-use-your-name-and-comments-in-ads (discussing how some of Google's settings are set by default to Track-Me); Andrew Couts, *Are Apple's iOS 7 Privacy Settings Purposefully Misleading, or Just a Mess?,* DIGITAL TRENDS (Sept. 29, 2013), http://www.digitaltrends.com/mobile/apple-your-ios-7-privacy-settings-are-a-mess/#ixzz2iCafwri3 (discussing iPhone defaults set to enable tracking). Although most tracking settings are set by default to Track-Me, a few are not. *See, e.g.*, *Safari Blocks all Cookies by Default*, APPLE SUPPORT COMMUNITIES, https://discussions.apple.com/thread/4040376 (last visited Aug. 1, 2013).

25. *See, e.g.*, Peter Maass & Megha Rajagopalan, *That's No Phone. That's My Tracker*, PROPUBLICA (July 13, 2012), http://www.propublica.org/article/thats-no-phone.-thats-my-tracker (reporting that cellphone tracking may continue even if phone is powered down); *Google: Gmail Users 'Have No Legitimate Expectation of Privacy,'* RT.COM (Aug. 13, 2013), http://on.rt.com/47y0ku (reporting that Gmail cannot be used without allowing Google to scan email content for various purposes); Katy Bachman, *Yahoo Says No to Microsoft's 'Do Not Track' Browser, Others Expected to Follow Suit*, ADWEEK (Oct. 26, 2012), http://www.adweek.com/news/technology/yahoo-says-no-microsofts-do-not-track-browser-144847 (reporting that Yahoo and others will not honor a "Do-Not-Track" signal received from an Internet Explorer browser). *See generally* Chris Jay Hoofnagle et al., *Behavioral Advertising: The Offer You Cannot Refuse*, 6 HARV. J.L. & PUB. POL'Y 273 (2012) (detailing how trackers have evaded technology consumers use to prevent tracking).

26. *See, e.g.*, Alan Henry, *Everyone's Trying to Track What You Do on the Web: Here's How to Stop Them*, LIFEHACKER (Feb. 22, 2012), http://lifehacker.com/5887140/everyones-trying-to-track-what-you-do-on-the-web-heres-how-to-stop-them; Erica Naone, *Smartphone Apps: How to Spot and Stop Firms Tracking Your Phone*, CHRISTIAN SCIENCE MONITOR (May 5, 2011), http://www.csmonitor.com/Innovation/Latest-News-Wires/2011/0505/Smartphone-apps-How-to-spot-and-stop-firms-tracking-your-phone.

Facebook, Twitter, or Google accounts,[27] and they can set their browsers to ask websites not to track them.[28]

While a majority of consumers claim they have taken at least some steps to opt out of Internet tracking,[29] these claims are surely more aspirational than representational.[30] It is extremely unlikely that many consumers successfully opt out on each device, browser, and app through which they are tracked. Currently, about 17% of Firefox users in the United States have their browsers set to "Do Not Track."[31] Fewer than 10% of consumers even know that common mobile phone apps track them.[32] Firms that claim they can track nearly all Americans across all their devices may be exaggerating, but probably not by much.[33] Despite widespread opposition to tracking, most consumers stick with most preset Track-Me positions.

---

27. *See, e.g.*, Alan Henry, *Facebook Is Tracking Your Every Move on the Web; Here's How to Stop It*, LIFEHACKER (Sept. 26, 2011), http://lifehacker.com/5843969/facebook-is-tracking-your-every-move-on-the-web-heres-how-to-stop-it; Ryan Tate, *How Google Spies on Your Gmail Account (And How To Stop It)*, GAWKER (May 11, 2011), http://gawker.com/5800868/how-google-spies-on-your-gmail-account-and-how-to-stop-it; Alan Henry, *Twitter Wants to Start Tracking You on the Web, Here's How to Opt-Out*, LIFEHACKER (July 3, 2013), http://lifehacker.com/twitter-wants-to-start-tracking-you-on-the-web-heres-661569459.

28. *See, e.g.*, Thorin Klosowski, *Everywhere You Can Enable 'Do Not Track*,*'* LIFEHACKER (Aug. 5, 2013), http://lifehacker.com/everywhere-you-can-enable-do-not-track-1006138985.

29. *See* Rainie et al., *supra* note 4, at 8 (finding that eighty-six percent of consumers claim to have taken at least one step to avoid being tracked).

30. Consumers claim to engage in significantly more privacy-protective online behavior than they truly do. *See* Carlos Jensen et al., *Privacy Practices of Internet Users: Self-reports Versus Observed Behavior*, 63 INT. J. HUMAN-COMPUTER STUD. 203, 226 (2005).

31. Alex Fowler, *Mozilla's New Do Not Track Dashboard: Firefox Users Continue to Seek out and Enable DNT*, MOZILLA PRIVACY BLOG (May 3, 2013), https://blog.mozilla.org/privacy/2013/05/03/mozillas-new-do-not-track-dashboard-firefox-users-continue-to-seek-out-and-enable-dnt.

32. David Talbot, *Using Crowdsourcing to Protect Your Privacy*, MIT TECH. REV. (Apr. 2, 2012), http://www.technologyreview.com/news/427390/using-crowdsourcing-to-protect-your-privacy.

33. While consumers claim to engage in privacy-protective behavior, firms that collect personal data advertise that they have extensive data on most U.S. adults. *See* Natasha Singer, *Mapping, and Sharing, the Consumer Genome*, N.Y. TIMES, June 16, 2012, *available at* http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html. One firm claims that it can use digital fingerprinting technology, a technology that does not depend on cookies and can identify both mobile devices and computers, to identify ninety-eight percent of Internet users. *See* Adam Tanner, *The Web Cookie Is Dying. Here's the Creepier Technology That Comes Next*, FORBES, June 17, 2013, http://www.forbes.com/sites/adamtanner/2013/06/17/the-web-cookie-is-dying-heres-the-creepier-technology-that-comes-next/. Another firm claims its technology can match 1.5 billion devices and that "its pinpointing [is] so accurate that it c[an] show spouses different, personalized ads on a tablet they share." Claire Cain Miller & Somini Sengupta, *Selling Secrets*

Why do people adhere to default settings even when they dislike those settings? The following Section surveys the mechanisms that make defaults sticky and the conditions under which these mechanisms operate.

A. MECHANISMS THAT MAKE DEFAULTS STICKY

Three types of mechanisms can operate to make defaults sticky: (a) transaction barriers, (b) judgment and decision biases, and (c) the preference-forming effects of defaults.[34] Not every mechanism will affect every default, but where a default is sticky, one or more of these mechanisms are at work.

1. Transaction Barriers

Transaction barriers include the costs of opting out, confusion about the default setting or the opt-out process, and a sense of futility where opting out seems to be impossible. Today's Track-Me quasi-default provides an example of all of these factors.

Opting out of tracking today is costly. Consumers must find the opt-out procedure, if one exists, and execute the steps for opting out, such as installing a program, changing settings, or completing an online form. Even when opting out of a single Track-Me setting is not onerous, it must be completed for each device (e.g., cellphone, tablet, desktop) and program through which one can be tracked.[35] If a consumer wants to indicate fine-grained preferences, the number of firms from which the consumer would need to opt out—on each browser on each device the consumer uses—is vast. One reporter counted over 100 companies that tracked him online in a thirty-six-hour period of ordinary web use.[36] Further, some of these steps must be repeated when consumers upgrade devices or software.[37] In addition, if a consumer deletes all cookies, cookies that had been sending "Do-Not-Track" messages must be reinstalled.[38]

---

of Phone Users to Advertisers, N.Y. TIMES, Oct. 5, 2013, *available at* http://www.nytimes.com/2013/10/06/technology/selling-secrets-of-phone-users-to-advertisers.html.

34. For a more detailed explanation of some of the mechanisms that can make defaults sticky, see Lauren E. Willis, *When Nudges Fail: Slippery Defaults*, 80 U. CHI. L. REV. 1155 (2013).

35. *See, e.g.*, *Privacy Policy*, COX DIGITAL SOLUTIONS, http://www.coxdigitalsolutions.com/privacy_policy.html (last visited Apr. 11, 2014) ("If you use more than one type of browser or more than one computer to access the Internet, you will have to opt out in each browser and on each computer that you use.").

36. Alexis Madrigal, *Digital Shadow: How Companies Track You Online*, WEEK (Apr. 13, 2012), http://theweek.com/article/index/226708/digital-shadow-how-companies-track-you-online.

37. *See, e.g.*, *Privacy Policy*, *supra* note 35 ("You may need to opt out repeatedly. If you delete or otherwise alter your browser's cookie file (including upgrading certain browsers) you may need to opt out again.").

38. *Id.*

Further, several misconceptions about tracking deter opting out. Many people do not even know that firms today can track their Internet use, scan their email text, and follow their device movements for commercial purposes.[39] In one survey nearly all respondents were surprised that a popular flashlight app sent the user's unique cellphone ID and precise location to advertisers.[40] Most consumers falsely believe that the law significantly restricts collection of consumer information[41] and that the existence of a "Privacy Policy" means their information is not shared with third parties.[42] Mistaken about what happens if they do nothing, consumers who prefer not to be tracked have no reason to opt out. Other consumers know about tracking but are unaware of the (albeit limited) ways in which they can opt out.[43] Invisibility of the option to opt out inevitably leads to sticking with the default. Finally, many consumers think they know more about technology related to privacy than they do.[44] Popular misconceptions include the belief that changing browser settings or deleting cookies disables Internet tracking.[45] The result is that consumers who attempt to opt out of tracking

39. *See, e.g.*, Joseph Turow et al., *Americans Reject Tailored Advertising and Three Activities That Enable It* 21, tbl.9 (Sept. 29, 2009) (unpublished manuscript), *available at* http://ssrn.com/abstract=1478214 (finding that only thirty-three percent of users in 2009 survey knew that their Internet use could be tracked across multiple websites without their consent); Adario Strange, *Google's Snooping Gmail the Target of Microsoft's Latest 'Scroogle' Attack*, ITPROPORTAL (Feb. 8, 2013), http://www.itproportal.com/2013/02/08/googles-snooping-gmail-the-target-of-microsoft-latest-scroogle-campaign (reporting that fewer than thirty percent of Gmail users surveyed knew that their email text was scanned for behavioral advertising purposes); Stephanie Clifford & Quentin Hardy, *Attention, Shoppers: Store Is Tracking Your Cell*, N.Y. TIMES, July 14, 2013, http://www.nytimes.com/2013/07/15/business/attention-shopper-stores-are-tracking-your-cell.html (reporting that shoppers were surprised that their movements within a store were monitored through their cellphones).

40. Talbot, *supra* note 32.

41. *See* Turow et al., *supra* note 39, at 21, tbl.9 (2009).

42. Jensen et al., *supra* note 30, at 223; *see also* Ilana Westerman, *What Misconceptions Do Consumers Have About Privacy?*, PRIVACY PERSPECTIVES (June 3, 2013), https://www.privacyassociation.org/privacy_perspectives/post/what_misconceptions_do_consumers_have_about_privacy (discussing the misconceptions consumers have about privacy generally).

43. Two-thirds of consumers are unaware of options they have to limit how much information is collected about them. *See* KRISTEN PURCELL ET AL., SEARCH ENGINE USE 2012, PEW INTERNET & AMERICAN LIFE PROJECT 3 (2012); *see also* Annalect Research Group, *Internet Users' Response to Consumer Online Privacy*, ANNALECT.COM (Mar. 14, 2012), http://annalect.com/white-paper-internet-users-response-to-consumer-online-privacy (finding only twenty-two percent of users aware of some ability to opt out of tracking).

44. *See* Jensen et al., *supra* note 30, at 226.

45. Hoofnagle et al., *supra* note 25, at 277 (changing browser settings or deleting cookies does not prevent all tracking).

today often do not manage to do so to the extent they desire or even mistakenly believe they have done.[46]

A sense of futility presents another barrier to opting out.[47] Some consumers understand that the Track-Me position today is a quasi-default, not a true default, in that entirely opting out is not possible.[48] Others, perhaps accurately given the confusion associated with opting out, believe that they lack the expertise necessary to successfully opt out.[49] Rather than engage in futile attempts to opt out of the default, these consumers might not even try.[50]

2. Judgment and Decision Biases

Judgment and decision biases are another type of mechanism through which defaults can attract adherents. One pair of such biases that can pull people towards a default position is loss aversion[51] and the endowment effect.[52] Experiments in the personal-data-tracking arena illustrate the way in

46. Pedro G. Leon et al., *Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising*, *in* CHI 2012, ACM Press 589–98 (2012), *available at* http://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab11017.pdf ("[M]ultiple participants opted out of only one company . . . despite intending to opt out of all. Others mistook the [registration] page . . . as the opt out page."); *see also* Janger & Schwartz, *supra* note 15, at 1241 (noting that confusing and misleading privacy notices are designed to lead to consumer inaction).

47. *Cf.* Chad A. Proell & Stephen Sauer, *"Stock" Options: The Debilitating Effects of Autonomy and Choice on Self-perceptions of Power*, 23 J. OF BUS. & BEHAV. SCI. 82 (2011) (describing how feelings of powerlessness lead to inaction).

48. Rainie, *supra* note 4, at 12.

49. *See, e.g.*, Blasé Ur et al., *Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising* 8–9 (Carnegie Mellon Univ., Working Paper No. CMU-CyLab-12-007, 2012), *available at* http://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab12007.pdf (reporting that in response to questions about opting out of behavioral advertising, many consumers expressed uncertainty about how to opt out); *Yet Another OPT OUT You Should Think About*, BOXLOUR MEDIA (Oct. 20, 2013), http://www.boxlour.com/?p=319 ("Typically the process to opt out of something is not as easy as a 'click here' button. In fact, they are literally banking on most of us getting so confused on the whole 'opt out' process that we eventually give up. And give up we do.").

50. *See, e.g.*, Blasé Ur et al., *supra* note 49, at 7 (quoting consumer: "It makes me want to go home and delete all my cookies, but then I know that's not gonna help much."); Maria Karyda & Spyros Kokolakis, *Privacy Perceptions Among Members of Online Communities*, *in* DIGITAL PRIVACY 253, 263 (Alessandro Acquisti et al. eds., 2010) (discussing consumer sentiments that user attempts to obtain privacy on the Internet are futile).

51. Loss aversion means weighing losses more heavily than gains against some reference point. *See, e.g.*, William Samuelson & Richard Zeckhauser, *Status Quo Bias in Decision Making*, 1 J. RISK & UNCERTAINTY 7, 19, 31 (1988).

52. The endowment effect, placing a higher value on what one already possesses (or perceives oneself as possessing) than on the same thing when one does not possess it, is a manifestation of loss aversion. *See, e.g.*, Russell Korobkin, *Wrestling with the Endowment Effect,*

which these biases can favor a default position. This research demonstrates an almost absurdly strong endowment effect for privacy. On average, people will give up much more to keep data private when it is private by default, compared to what they will pay to obtain privacy when told the default is for that data to be public; subjects in one experiment "were five times more likely to reject cash offers for their data if they believed that their privacy would be, by default, protected, than if they didn't enjoy such belief."[53] Because today's default is for data not to be private, these biases favor the Track-Me position.

In addition to loss aversion and the endowment effect, the biases that have been identified in the existing literature as contributing to the magnetism of defaults include salience effects,[54] omission bias,[55] and procrastination and decision avoidance.[56] But the existing literature fails to recognize that nearly *any* type of bias might be invoked to favor a default, provided the necessary conditions exist to facilitate the operation of biases. In the case of tracking, for example, excessive discounting,[57] choice

or How to Do Law and Economics Without the Coase Theorem 16 (UCLA Sch. of Law, Law-Econ Research Paper, Paper No. 13-10, 2013), *available at* http://ssrn.com/abstract_id=2289574.

53. *See* Alessandro Acquisti et al., *What Is Privacy Worth* 3, *in* 21ST WORKSHOP ON INFORMATION SYSTEMS AND ECONOMICS (2009), *available at* http://www.futureofprivacy.org/wp-content/uploads/2010/07/privacy-worth-acquisti-FPF.pdf. Similar results were obtained in Jens Grossklags & Alessandro Acquisti, *When 25 Cents Is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information*, Workshop on the Economics of Information Security (2007), *available at* http://people.ischool.berkeley.edu/~jensg/research/paper/Grossklags_Acquisti-WEIS07.pdf.

54. The salience effect is the tendency for salient information to disproportionately affect judgments and non-salient information to be ignored. *See, e.g.*, Tanjim Hossain & John Morgan, . . . *Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay*, 62 ADVANCES IN ECON. ANALYSIS & POL'Y 1, 20 (2006).

55. Omission bias is the tendency to favor inaction over action. It is believed to be caused by a desire to avoid future regret and the fact that people more commonly blame themselves about a poor outcome when they take action than when they fail to take action. *See, e.g.*, Jonathan Baron & Ilana Ritov, *Reference Points and Omission Bias*, 59 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 475, 478 (1994).

56. Procrastination and decision avoidance are biases triggered when decisions or actions appear difficult. These biases can cause people to procrastinate indefinitely or affirmatively decide not to make any decision, either of which could lead to sticking with the default. *See, e.g.*, Ted O'Donoghue & Matthew Rabin, *Choice and Procrastination*, 116 Q.J. ECON. 121 *passim* (2001); Christopher J. Anderson, *The Psychology of Doing Nothing: Forms of Decision Avoidance Result from Reason and Emotion*, 129 PSYCHOL. BULL. 139 *passim* (2003).

57. Excessive discounting refers to people's tendency both to prefer a much smaller gain now to a larger gain later and to prefer a sure gain to an uncertain but probabilistically much larger gain. *See, e.g.*, Yaacov Trope & Nira Liberman, *Construal-Level Theory of Psychological Distance*, 117 PSYCHOL. REV. 440 *passim* (2010) (explaining the tendency for people to discount over psychological distance, including over time and over uncertainty).

bracketing, the illusion of control,[58] and the sunk costs fallacy[59] all may draw consumers to today's Track-Me quasi-default.

Choice bracketing's effect on consumers facing a default may be particularly strong. Choice bracketing refers to whether a decision is evaluated in isolation or as part of a larger set of decisions.[60] Health-related decisions present an intuitive example. If the choice to eat a dessert or go for a run is made in isolation, the benefits of the dessert and costs of the run might easily outweigh the trivial effect of each on health. Yet, the cumulative effect of these daily decisions can be enormous. Someone who views each choice in isolation might make less healthy choices than someone who mentally brackets decisions about diet, exercise, or health broadly.[61]

Personal data privacy decisions present a similar structure. Tracking decisions can be conceived narrowly or broadly. One consumer might make each decision in isolation, considering whether to permit a particular entity to track her at a particular moment in time. Another consumer might conceptualize each tracking decision as part of a broader choice about whether to allow herself to be tracked by any entity anytime. Firms are able to connect data gathered from a variety of sources—off-line, online, and from mobile devices—about a single consumer, collected over time.[62] Thus,

58. The illusion of control is a bias that can lead people to take on more risk than they otherwise would, which might encourage consumers to stick with a risky default if surrounding circumstances invoke the illusion. *See, e.g.*, Mark S. Horswill & Frank P. McKenna, *The Effect of Perceived Control on Risk Taking*, 29 J. APPLIED SOC. PSYCHOL. 377 (1999); Ellen J. Langer, *The Illusion of Control*, 32 J. PERSONALITY & SOC. PSYCHOL. 311 (1975). An example of the illusion is the common belief that one is less prone to experience an accident when one is driving than when one is a passenger, regardless of driving skill. *See* Frank P. McKenna, *It Won't Happen to Me: Unrealistic Optimism or Illusion of Control?*, 84 BRIT. J. PSYCHOL. 39, 39–50 (1993).

59. The sunk costs fallacy refers to the common error of weighing irreversibly incurred costs in a decision about whether to continue with a project or switch course. Hal R. Arkes & Catherine Blumer, *The Psychology of Sunk Cost*, 35 ORG. BEHAV. & HUM. DECISION PROCESSES 124 (1985). For example, someone who has learned how to use one type of software might continue to use that software instead of switching to an easier-to-use program because otherwise the effort of learning the first software seems to have been wasted. But, because that time and effort can never be recovered, it should not come into the calculus about what software to use going forward. Where the option to opt out of a default is not presented until after people have invested in the default, the sunk costs fallacy will favor the default.

60. Daniel Read et al., *Choice Bracketing*, 19 J. OF RISK & UNCERTAINTY 171, 172 (2000).

61. *Cf. id.* at 171 (using example of decisions about smoking cigarettes).

62. *See, e.g.*, Somini Sengupta, *What You Didn't Post, Facebook May Still Know*, N.Y. TIMES, Mar. 25, 2013, *available at* http://www.nytimes.com/2013/03/26/technology/facebook-expands-targeted-advertising-through-outside-data-sources.html (describing aggregation of online and off-line data about individual consumers).

the benefits of privacy and costs of a lack of privacy depend on the whole of privacy, not its parts; it is impossible to know whether tracking by any particular party or of any particular piece of information will tip the scales.[63] Because the probable negative impact of tracking by any one tracker is negligible, bracketing the choice narrowly could incline consumers to allow tracking. In contrast, broad choice bracketing is more likely to lead consumers to select a Don't-Track-Me position because the entirety of potential harms from tracking looms larger.

People often accept the bracketing implicit in a decision's presentation,[64] and many tracking decisions today are presented in a narrow form—e.g., Would you like to opt out of "tailor[ed] ads" from Twitter?[65] Rather than presenting the user with a broad opt-out choice, a single device or program often requires opting out of a series of particular types of tracking.[66] Even where broader choices are presented, such as in browser settings, trackers can ask consumers to alter that setting as to a particular website, a narrow choice. Thus, narrow choice bracketing may fortify today's Track-Me position.

3. Preference-Formation Effects

In addition to transaction barriers and judgment and decision biases, the process of forming preferences can encourage people to keep default settings.[67] This happens in two ways, through the endorsement effect and the experience effect.

The endorsement effect is the interpretation of a default as a form of implicit advice by a more knowledgeable party as to what most people prefer or ought to prefer.[68] Often, consumers follow this implicit advice.[69] For example, when consumers believe tracking defaults are selected based on majoritarian preferences, they may accept them on that basis alone because when it comes to privacy, consumers are particularly prone to following the

63. *Cf.* Solove, *supra* note 18, at 1889–90 (dubbing this the "aggregation effect").

64. Read et al., *supra* note 60, at 188 ("[When] choices come to [people] one at a time, they will bracket them narrowly, and if choices come to them collectively, they will bracket more broadly.").

65. *See* Henry, *supra* note 27.

66. *See, e.g.*, Jason D. O'Grady, *Four Privacy Settings You Should Enable in iOS 7 Immediately*, ZDNET (Sept. 19, 2013), http://www.zdnet.com/four-privacy-settings-you-should-enable-in-ios-7-immediately-7000020902.

67. *See, e.g.*, Cass Sunstein, *Switching the Default Rule*, 77 N.Y.U. L. REV. 106 (2002) (discussing preference-formation effects of defaults).

68. *See, e.g.*, Eric J. Johnson & Daniel G. Goldstein, *Decisions by Default*, *in* THE BEHAVIORAL FOUNDATIONS OF PUBLIC POLICY 417, 421 (Eldar Shafir ed., 2013).

69. *See, e.g.*, McKenzie et al., *supra* note 14, at 414.

crowd.[70] Along the same lines, consumers may agree to today's Track-Me quasi-default because they trust the firms with which they do business and assume these firms have set the default with consumers' best interests in mind.[71]

The experience effect can lead people to develop a preference for the default position after spending some time experiencing it. For example, cellphone users who perceive themselves as benefitting from their existing privacy settings might choose to stick with a phone's Track-Me position. Yet, if Don't-Track-Me had been the phone's default setting, they would have become accustomed to that position and developed a preference for it instead.[72]

B. CONDITIONS UNDER WHICH DEFAULTS ARE OFTEN STICKY

The mechanisms that can give defaults traction do not operate consistently across all consumers, but rather vary with consumers' understanding of their options and of their own preferences. When consumers understand their options and their preferences well, they can usually match the two easily, absent transaction barriers. A consumer who knows she prefers the opt-out position will not be swayed by the implicit advice conveyed in the policymaker's selection of the default, for example.[73] Likewise, consumers who already know their options are less affected by how salient each option is at the moment of decision. In these situations, biases and the preference-formation effects of defaults have little influence on outcomes. That is, defaults will not be sticky.[74]

---

70. Alessandro Acquisti et al., *The Impact of Relative Judgments on Concern About Privacy*, 49 J. MARKETING. RES. 3 (2012).

71. Although cellphone companies do not fare as well, Amazon, Google, and Apple are among the ten companies with the highest reputations among consumers, and trust is one of the major drivers of these ratings. HARRIS INTERACTIVE, THE HARRIS POLL 2013 RQ SUMMARY REPORT 6, 9 (2013), http://www.harrisinteractive.com/vault/2013%20RQ%20Summary%20Report%20FINAL.pdf.

72. *Cf.* Lior Jacob Strahilevitz, *Toward a Positive Theory of Privacy Law*, 126 HARV. L. REV. 2010 (2013) (suggesting that reactive nature of U.S. privacy policymaking leads to less privacy protection because social norms adjust to the status quo).

73. *Cf.* Erin Todd Bronchetti et al., *When a Nudge Isn't Enough: Defaults and Saving Among Low-Income Tax Filers* (Nat'l Bureau of Econ. Research, Working Paper No. 16887, 2011), *available at* http://www.nber.org/papers/w16887 (finding that placing part of tax refund in a savings vehicle by default had no effect on whether taxpayer saved the funds, and suggesting that the result was because taxpayers had well-understood preexisting preferences about whether to save or spend the funds).

74. *See, e.g.*, Korobkin, *supra* note 52, at 1622 (presenting evidence that when preference uncertainty is removed, defaults lose their power); Asa Lofgren et al., *Are Experienced People Affected by a Pre-Set Default Option—Results from a Field Experiment*, 63 J. ENVIRON. ECON. &

But when consumers do not understand their options or their preferences, consumers are more susceptible to the mechanisms that make defaults sticky.[75] The personal-data-tracking arena has both of these features, and thus is one in which defaults can be powerful.[76]

Most consumers understand their information privacy options poorly.[77] The variety of data that can be collected through tracking, the host of entities that can collect or obtain that data, the ways that data can be used, and the potential costs and benefits of data collection are bewildering. The intangible effects of tracking are nigh impossible to assess; for example, most people probably cannot forecast whether any particular tracking will make them feel watched or how such a feeling will impact their lives.[78] Even as to concrete effects, people do not know—and cannot know—what effect their data will have on pricing, access to employment, or chances of identity theft.[79] This complexity defies simplification. One set of researchers, who found that simplifying the language and formatting of privacy policies barely improved

MGMT. 66 (2012) (finding that people who know what they want through experience with a decision are less affected by defaults).

75. *See, e.g.*, Sarah Lichtenstein & Paul Slovic, *The Construction of Preference: An Overview*, *in* THE CONSTRUCTION OF PREFERENCE 1 (Sarah Lichtenstein & Paul Slovic eds., 2006) (observing that judgment and decision biases are strongest when preferences are uncertain); Samuelson & Zeckhauser, *supra* note 51, at 29 (finding that choice difficulty increases the pull of defaults); *id.* at 8 (finding subjects more susceptible to biases favoring the status quo when their preferences are weaker). Other academics have described this process a bit differently, claiming that when consumers lack preexisting preferences, they construct preferences in the course of decision making, and that the framing of the decision through the presence of a default shapes how that preference is constructed. But, a position's status as a default does not alone drive consumers towards that position; rather, the mechanisms that can make defaults sticky must come into play. *See* Julie R. Agnew et al., *Who Chooses Annuities? An Experimental Investigation of the Role of Gender, Framing and Defaults*, 98 AM. ECON. REV. 418, 421 (2008) (finding experimentally that defaults have no effect when transaction barriers, biases, and preference formation effects are absent). Further, many of these mechanisms can lead a consumer to remain in a default position without affecting her preferences; a consumer who sticks with a default due to transaction costs or salience effects does not necessarily prefer the default over the opt-out position, *ceterus paribus*.

76. *Cf.* Eric J. Johnson et al., *Defaults, Framing and Privacy: Why Opting In ≠ Opting Out*, 13 MARKETING LETTERS 5 (2002) (finding large default effect in a privacy-related experiment, and suggesting that people's uncertainty about their preferences contributes to the power of privacy defaults).

77. *Cf.* Jensen et al., *supra* note 30 (finding that consumers have little understanding of even well-publicized privacy-related technologies, such as cookies).

78. *See, e.g.*, Ur et al., *supra* note 49; Aleecia M. McDonald & Lorrie Faith Cranor, *An Empirical Study of How People Perceive Online Behavioral Advertising* (Carnegie Mellon Univ., Working Paper CMU-CyLab-09-015, 2009), *available at* http://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab09015.pdf.

79. *See, e.g.*, Alessandro Acquisti & Jens Grossklags, *Uncertainty, Ambiguity and Privacy*, COMM. & STRATEGY 6 (2012).

consumer comprehension, explains, "even the most readable policies are too difficult for most people to understand and even the best policies are confusing."[80]

Most consumers have a similarly poor grasp of their own privacy preferences and what actions are necessary to meet those preferences. Even consumers who describe themselves as highly valuing privacy act in ways that reveal abundant private information; in one experiment, many consumers who said they wanted to keep information about themselves private revealed that very information when asked for it by an interactive computer figure on a commercial website.[81] Consumers also have competing preferences. For example, they find tracking for advertising purposes scary and creepy yet also smart and useful.[82] They cannot make the tradeoffs between the incommensurable costs and benefits of privacy; they want to keep their personal information private *and* they want the benefits made possible by revealing that information. One survey's findings are telling: 84% of consumers say they would rather receive targeted advertising in exchange for online content than to pay for online content with money, but 93% of these same consumers say that they would opt into a Don't-Track-Me position if given the choice.[83]

III. DEFAULTS IN THEORY

Observing that people tend to stick with default settings, particularly when they are uncertain about their options and their preferences, academics have claimed that defaults can further policy objectives. This Part explains the theory behind this claim. It then describes how policymakers appear likely to translate this theory into practice in the data privacy policy arena.

A. THE THEORY BEHIND THE USE OF DEFAULTS IN POLICYMAKING

Academics have suggested that defaults can be used in policymaking in two ways: to increase the number of people in the default position (policy

80. Aleecia M. McDonald et al., *A Comparative Study of Online Privacy Policies and Formats*, 5672 LECTURE NOTES IN COMPUTER SCI. 37, 52 (2009).

81. *See* Sarah Spiekermann et al., *E-Privacy in 2nd Generation E-Commerce: Privacy Preferences Versus Actual Behavior* § 3.3 (2002), http://people.ischool.berkeley.edu/~jensg/research/paper/grossklags_e-Privacy.pdf.

82. Ur et al., *supra* note 49, at 6; *see also* Ronald Brownstein, *Americans Know They've Already Lost Their Privacy*, NAT'L J. (June 13, 2013), *available at* http://www.nationaljournal.com/next-economy/big-questions/americans-know-they-ve-already-lost-their-privacy-20130613 (finding widespread consumer ambivalence about the effects of personal information sharing).

83. Annalect Research Group, *supra* note 43, at 2.

defaults) or to provide firms with incentives to educate people about the default (penalty defaults).[84] To ensure that policy defaults are sticky for consumers with weak preferences but not for those who truly prefer the opt-out position, and to ensure that penalty defaults are information-forcing, policymakers have sought to employ a variety of altering and framing rules.

1. Policy Defaults

Policy defaults are put in place with the explicit goal of increasing the number of people in the default position.[85] The idea is to set the default to a position that is good for most individuals,[86] under the assumptions that the majority will keep the default setting and that only the minority who have contrary preferences will reject it. The default in theory will guide uncertain individuals to the position that is best for most people but will not prevent those who know they have contrary preferences from opting out.

The iconic case of a policy default is automatic enrollment in defined contribution retirement savings plans, which is believed to be the best position for the vast majority of employees.[87] Employers that have made participation in their plans the default have increased their employee participation rates by forty percent or more.[88] Another example is the default for checking account overdraft coverage, which, in effect, prohibits banks from providing a consumer with expensive overdraft coverage for ATM and nonrecurring debit card transactions unless the consumer opts out of a no-overdraft default setting.[89] Regulators enacted this default scheme on the theory that the default position is best for most consumers, and the few consumers who benefit from overdraft coverage can opt out.[90]

84. *See, e.g.*, THALER & SUNSTEIN, *supra* note 21 (advocating the use of policy defaults); Ayres & Gertner, *supra* note 15, at 91 (advocating the use of penalty defaults).

85. *Accord* McKenzie et al., *supra* note 14.

86. *See* Janger & Schwartz, *supra* note 15, at 1236, 1245–46 (using "majoritarian defaults" to refer to default positions set to what the policymaker believes the majority of relevant parties would want and "normative defaults" to refer to default positions set to what the policymaker believes is best for society).

87. *See, e.g.*, THALER & SUNSTEIN, *supra* note 21; John Beshears et al., *Public Policy and Saving for Retirement: The "Autosave" Features of the Pension Protection Act of 2006*, *in* BETTER LIVING THROUGH ECONOMICS 274, 287 (John G. Seigfried ed., 2010) (quoting Peter Orszag as "declar[ing] the autosave features of the Pension Protection Act 'a stunning example of the success of behavioral economics in effecting public policy' ").

88. *See, e.g.*, William E. Nessmith et al., *Measuring the Effectiveness of Automatic Enrollment*, 31 VANGUARD CENTER FOR RETIREMENT RES. RPT. 6 (2007).

89. 12 C.F.R. § 205 (2014).

90. *Id.*

2. Penalty Defaults

Penalty defaults correct information asymmetries between parties, such as those that commonly exist between firms and consumers. Policymakers set the default to a position disliked by firms, on the premise that firms that want consumers to opt out will be forced to reveal the default and negotiate with consumers.[91] Like policy defaults, penalty defaults respond to the evidence that less knowledgeable consumers are most frequently affected by the default. Firm efforts to explain the default and opt-out positions so as to convince consumers to opt out will, in theory, educate precisely those consumers who need to know more about the default to make a good decision.

Two well-known penalty defaults are the warranties of merchantability[92] and fitness for a particular purpose[93] under the Uniform Commercial Code. The penalty default is that the warranties will apply; a seller that does not want the warranties to be part of the contract for sale must explicitly opt out.[94] The checking account overdraft coverage default could also be a penalty default. If the bank wants consumers to pay the affected types of overdraft fees, it will be forced to explain how overdraft coverage works in the process of convincing consumers to opt out of the default.

3. Altering and Framing Rules

Policymakers attempt to manage the stickiness, slipperiness, and informativeness of defaults through altering and framing rules. Altering rules tinker with the process for opting out. For example, an altering rule might require that a consumer be able to opt out with a single click, or might require a consumer to complete many steps to opt out. Framing rules attempt to manage the presentation of the default and opt-out option to the user, whether by architecture or messaging. Architecturally, for example, a software default setting could be made more or less salient through the positioning of opt-out controls, hidden deep inside multiple menus or popping up in a nagging window on the screen. Opting out of a default also might be made more or less attractive through messages to the user. For example, an alert or notice box might suggest that the user change a software

91. *See* Ayres & Gertner, *supra* note 15, at 91; *see also* Janger & Schwartz, *supra* note 15, at 1239 (dubbing penalty defaults "information-forcing defaults").

92. U.C.C. § 2-314 (2013).

93. *Id.* § 2-315.

94. *Id.* § 2-316. Note that in the consumer context, the Magnuson Moss Act restricts the degree to which these default warranties can be disclaimed. *See* 15 U.S.C. § 2301 *et seq.* (2012).

setting or instead warn the user that changing the setting could cause problems.

Where a default might otherwise be too sticky, altering rules might aim to keep the costs of opting out low and the process for opting out simple. Framing rules might aim to keep the option to opt out visible. But where policymakers fear that a default will be too slippery and will not be information-forcing, they might set altering rules that impose transaction costs on opting out and framing rules that require the provision of information as a precondition of opting out. Altering and framing rules reflect some awareness that defaults alone will not achieve policymakers' desired ends. Theoretically, carefully calibrated altering and framing rules make policy and penalty defaults work properly.[95]

Two examples show how these rules are employed. In the case of automatic enrollment in retirement savings plans, policymakers are concerned that the default could be too sticky. Altering and framing rules therefore require employers to give employees who are enrolled by default written notices of the right to opt out at specified intervals. These notices must be written at a level that the average employee can understand so that the opportunity to opt out is not confusing or invisible.[96] In contrast, policymakers are concerned that the default for checking account overdraft coverage on ATM and debit transactions could be too slippery and might not be information-forcing.[97] Therefore, framing rules require banks to give accountholders notices explaining the potential costs of opting out of the default and into overdraft coverage—notices that must be segregated from other account documents.[98] Altering rules require banks to provide the same

95. *Cf.* Ayres, *supra* note 16, at 2044–45 (explaining that altering rules can be used to "enhance contractual efficiency and equity" and/or to "manage and restrain negative externalities and internalities while simultaneously permitting opt-out for a subset of contractors who, at least as a group, pass a social cost benefit test").

96. 26 C.F.R. § 1.401(k)–3(d)(3) (2014) ("The content requirement . . . is satisfied if the notice is—(A) Sufficiently accurate and comprehensive . . . (B) Written in a manner calculated to be understood by the average employee . . . .").

97. *See* Electronic Fund Transfers, 12 C.F.R. § 205.17(b) (2014). An overdraft occurs when an accountholder attempts to withdraw more from her checking account than is available, and the bank covers the withdrawal and then reimburses itself when the next deposit is made into the account. For what amounts to a short-term loan the bank charges the accountholder a fee, one that is frequently larger than the amount borrowed. *See* FED. DEPOSIT INSURANCE CORP., FDIC STUDY OF BANK OVERDRAFT PROGRAMS 20 tbl.IV–9 (2008), http://www.fdic.gov/bank/analytical/overdraft/FDIC138_Report_Final_v508.pdf (finding in a 2007 survey of banks that the median negative balance from a debit transaction on which an overdraft fee was charged was about twenty dollars and the median fee was twenty-seven dollars).

98. 12 C.F.R. § 205.17(d) (2014).

account terms, conditions, and features to accountholders who do and do not keep the default setting.[99] They also prohibit banks from opting consumers out in routinely unread fine print; consumers must affirmatively opt out.[100]

4. An Aside on Forced Choice

In addition to Track-Me or Don't-Track-Me defaults, a third possibility is forced choice, meaning consumers could not continue with the online or off-line activity from which data will be collected without affirmatively making choices about tracking.[101] Under a forced choice regime, a consumer could not enter a store that tracks cellphones, view a website that tracks browsing activity, use a credit card at a retailer that tracks customer purchases, or open a mobile device application that collects personal data without first deciding whether to be tracked. Forced choice could be attractive where policymakers believe consumers have diverse well-informed preferences that ought to be honored.

However, forced choice does not work well when people are uncertain about their preferences and options, and would be terribly burdensome for consumers if applied to every situation in which a firm wanted to track a consumer.[102] Forced choice would mean that every consumer decision about tracking would be narrowly-bracketed, which, as explained previously, favors the Track-Me position. In fact, as explained below, firms that benefit from tracking consumers would doubtless respond to a Don't-Track-Me default by repeatedly asking consumers to opt-out, creating a scenario similar to forced choice. Most consumers would respond with a reflexive "yes" click so as to move along in their daily activities rather than engage in reflective decision making.[103]

---

99. *Id.* § 205.17(b)(3).

100. *Id.* § 205.17(b)(1)(iii).

101. *See* Sunstein, *supra* note 11 at 18 ("Another possibility would be to ask customers—the first time they open the browser or periodically—about the privacy setting that that they prefer, and perhaps to prevent them from proceeding until they answer."). "Forced" or "mandated" choice is also called "active choosing." *See id.*

102. *See id.* at 21 ("Suppose that the situation is unfamiliar and complicated. Suppose that people lack information or experience. If so, active choosing may impose unjustified or excessive costs on people; it might produce frustration and appear to require pointless red-tape.").

103. In the abstract, the occasional forced choice might be somewhat successful in improving the quality of consumer choice. But this presents a coordination problem, given that no firm will know whether another firm has recently asked a consumer to make a decision about tracking.

B. TRANSLATING DEFAULT THEORY TO TRACKING POLICY

Translating the theory behind the use of defaults in policymaking into practice in the personal-data-tracking context quickly reveals that the academic literature is incomplete. That literature addresses what the position of the default setting ought to be and, roughly, how altering and framing rules ought to be set, but does not address a necessary prior question—the scope of the default setting. It also does not address the granularity of opt-out choices consumers ought to be given.

The following describes the tracking default positions that the theory behind the use of defaults in policymaking might support. It also explores the scope of these default positions, the opt-out choices, and the accompanying altering and framing rules that policymakers appear poised to enact. To do this, it draws primarily from two sources. First, the preliminary draft of the default scheme for tracking of information about individuals' use of the Internet and mobile devices[104] developed by the World Wide Web Consortium ("W3C"), an international body that sets voluntary, widely-adopted technological standards intended "to lead the World Wide Web to its full potential."[105] Second, the proposals made by the Federal Trade Commission ("FTC") for the collection or use of consumer data that can reasonably be linked to particular consumers, computers, or devices.[106] The

104. The current draft scheme is described in two documents: WORLD WIDE WEB CONSORTIUM, TRACKING COMPLIANCE AND SCOPE: W3C WORKING DRAFT, http://www.w3.org/2011/tracking-protection/drafts/tracking-compliance.html (last visited Aug. 15, 2013) [hereinafter W3C TCS], and WORLD WIDE WEB CONSORTIUM, TRACKING PREFERENCE EXPRESSION (DNT)—W3C WORKING DRAFT, http://www.w3.org/TR/tracking-dnt (last visited Aug. 15, 2013) [hereinafter W3C TPE].

105. W3C Mission, http://www.w3.org/Consortium/mission (last visited Nov. 25, 2013); *see also* ABOUT W3C, http://www.w3.org/Consortium (last visited Nov. 25, 2013) ("The World Wide Web Consortium (W3C) is an international community where Member organizations, a full-time staff, and the public work together to develop Web standards."). The W3C is not a body of elected policymakers, but rather a consortium of industry, consumer, and government representatives that seek to develop standards by consensus. The body has no regulatory authority, and must depend on the makers of devices, servers, browsers, and applications to voluntarily implement its standards. However, if it were to come to an agreement on tracking standards, policymakers might well adopt these standards into the law. *But see* Blair Reeves, *Online Advertising, BATNAs & the Failure of Do Not Track*, CLICKZ (Oct. 28, 2013), *available at* http://www.clickz.com/clickz/column/2302960/online-advertising-batnas-the-failure-of-do-not-track (expressing skepticism that the W3C will come to consensus on a tracking standard).

106. These two proposals are contained in: FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE vii, 35–59 (2012), *available at* http://www.ftc.gov/os/2012/03/120326privacyreport.pdf [hereinafter FTC PRIVACY REPORT]; FED. TRADE COMM'N, MOBILE PRIVACY DISCLOSURES 21 (2013), *available at*

W3C and FTC proposals are not legally binding, but default-minded legislators will certainly look to these sources when formulating privacy policy.

1. Scope of Tracking Defaults

The theory behind policy and penalty defaults assumes that defaults are transparently binary—one is either in a clear default position or one is not. But the tracking arena presents a particularly acute example of how complex and unintuitive a default and/or opt-out position might be. This is because "tracking" encompasses a very wide variety of personal data collection practices, and these practices have different costs and benefits depending on, for example:

- types of information collected (e.g., geolocation data, clickstream data,[107] medical data, email content),
- uses of information (e.g., medical research, marketing, pricing, website use analytics, employment),
- collection sites (e.g., particular websites, mobile applications),
- types of users of information (e.g., first-parties, meaning the firms with which consumers know they are interacting, such as wireless service providers and the firms that run websites; affiliates of first-parties; or third-parties to which first-parties sell data or access to data[108]), and
- individual users (particular firms or other entities).[109]

Policymakers could set the default broadly to Track-Me (or Don't-Track-Me) for most of these dimensions. Alternatively, they could set a mix of narrower defaults and unalterable positions.

---

http://www.ftc.gov/os/2013/02/130201mobileprivacyreport.pdf [hereinafter FTC MOBILE REPORT].

107. Clickstream data refers to the information generated by users' mouse movements and clicks through the website they visit, as well as the sites visited, duration of the visit, and order of site visits. *Definition: Clickstream Analysis*, SEARCHCRM, http://searchcrm.techtarget.com/definition/clickstream-analysis (last visited Aug. 1, 2013).

108. But given that third parties can become affiliates of first parties and that first parties can act as the handmaidens of third parties, tracking defaults that turn on these distinctions may be subverted. *Cf.* Mitch Weinstein, *Why Blocking Third-Party Cookies Is Good for Google and Facebook*, ADEXCHANGER (June 20, 2013), http://www.adexchanger.com/data-driven-thinking/why-blocking-third-party-cookies-is-good-for-google-and-facebook.

109. Policymakers could regulate the original collection of data, downstream users, or uses. Enforcement concerns would push toward limiting the original collection of data, even where only some potential downstream users or uses are problematic. However, to present the strongest case for the use of tracking defaults, this Article sets aside these enforcement concerns. *See supra* note 10 and accompanying text.

Consider the decision about whether to require firms to give consumers choice, or to set (or allow firms to set) an unalterable position. For some information, policymakers might decide that the costs of tracking outweigh the benefits, regardless of consumer preferences, and therefore set an unalterable Don't-Track-Me position. For example, policymakers might generally prohibit the tracking of sexual orientation or medical data for employment purposes. For other data, policymakers might decide that the benefits of tracking outweigh the costs, and therefore set an unalterable Track-Me position. For example, firms might be permitted to track clickstream data for website analytics purposes, without giving consumers an ability to opt out.

The broader the default scope, the more easily it can be understood, but the more likely that it will not satisfy most consumer preferences. For example, a Don't-Track-Me default that prohibits all passive personal data collection is intuitive, but most consumers want some tracking for some purposes (e.g., order fulfillment, fraud prevention, or clickstream data tracking for the purposes of pre- or re-populating online forms). A broad setting could thus lead most consumers to try to opt out to some extent, defeating the premise of policy defaults that inertia should lead most consumers to the best position. A broad default that leads most consumers to try to partially opt out could also founder if consumers are confused about their opt-out choices. For example, consumers might have the impression they must accept all tracking to receive the benefit of pre-populated online forms.

Existing tracking proposals therefore attempt to discern which position substantial majorities of consumers would want to be in with respect to each type of information, type of user, etc., and make these unalterable; but where consumers are more heterogeneous in their preferences, the proposals give consumers a choice. The current W3C draft proposal allows consumers to opt out of third-party tracking of geolocation data more granular than the zip-code level and of third-party tracking of other data when used for certain purposes, such as behavioral advertising. First parties could continue to collect any consumer data and could customize content and advertising based on the consumer's interaction with that firm.[110] Under the W3C draft proposal, consumers cannot opt out of any tracking by first parties, tracking of gross geolocation data by any party, or tracking for "permissible

110. W3C TCS, *supra* note 104, § 4.

purposes" by third parties.[111] Similarly, the FTC tracking proposal allows first parties to collect and use information about consumers without giving consumers an opportunity to opt out where "the practice [of data collection and use] is consistent with the context of the transaction or the consumer's existing relationship with the business, or is required or specifically authorized by law."[112] Contextually-appropriate uses not requiring consumer choice include the collection and use of data for the purposes of "fulfillment, fraud prevention, internal operations, . . . and most first-party marketing."[113]

While the precise contours of the scope of the defaults policymakers would adopt are unclear, what is clear is that there are some categories of information tracking from which consumers will not be permitted to opt out. As discussed further below, allowing consumers to opt out of some but not all tracking is likely to confuse consumers and lead some to feel that opting out is futile.

2. Setting: Track-Me or Don't-Track-Me

Within the set of potentially-tracked information about which policymakers wish to give consumers control, policymakers then must decide which defaults should be Track-Me and which should be Don't-Track-Me. If policymakers believe tracking produces more privacy benefits than harms (for example, in the form of financial support for Internet content or application development), they might support a Track-Me default. Because in this scenario policymakers believe that the majority ought to be in the Track-Me position,[114] it would be a policy default.

On the other hand, if policymakers believe that tracking produces more potential privacy harms than benefits (for example, in the form of increased risk of identity theft or decreased space for individual experimentation and growth), they might support a Don't-Track-Me default. Because they believe that the untracked position is best for most people, these policymakers would look for a policy-default effect, meaning that the majority would stick with this default position but those with strong contrary preferences would

111. "Permissible purposes" include frequency capping (measuring how many times a consumer has been shown an ad), billing (to ensure third-party advertisers are paying first-party websites correctly), and debugging. *Id.* § 5.

112. FTC PRIVACY REPORT, *supra* note 106, at 38–39.

113. *Id.* at 39.

114. Policymakers might support a Track-Me default on political expediency or other grounds as well, but this discussion focuses on the tracking default setting that the theory underlying the use of defaults in policymaking would support.

opt out. Many supporters of Don't-Track-Me defaults appear to aim for a policy-default effect.[115]

Policymakers who are uncertain about the social welfare effects of tracking or believe that people have heterogeneous preferences about tracking, also might support a Don't-Track-Me default. Personal-information tracking presents a case of information asymmetry, where one party (the firm) is well informed and the other (the consumer) is poorly informed. Because the business model of many Internet and mobile-application firms depends on the revenue firms can obtain through the sale of tracked information, firms have a strong interest in placing consumers in a Track-Me position.[116] Thus, a policymaker might intend for the Don't-Track-Me default to operate as a penalty default so that firms reveal this default position to consumers in the process of urging consumers to opt out. Consumers would then be free to make informed decisions about whether they want to be tracked. Academics in particular have argued for various forms of Don't-Track-Me defaults on this basis.[117]

Defaults could be set narrowly, with a different setting for each potential combination of data attributes listed above (i.e., type of information, information use, collection site, etc.). For example, the default for collection of health information by third parties for purposes other than diagnosis, treatment, or research could be Don't-Track-Me, but the default for

115. *See, e.g.*, Chris Soghoian, *End the Charade: Regulators Must Protect Users' Privacy by Default*, OFFICE OF THE PRIVACY COMMISSIONER OF CANADA (2010), http://www.priv.gc.ca/information/research-recherche/2010/soghoian_201012_e.asp (implying that a Don't-Track-Me default would be so sticky that advertisers would be forced to abandon behavioral advertising); Tene & Polonetsky, *supra* note 20 (suggesting that supporters of Don't-Track-Me defaults aim to keep people in the Don't-Track-Me position rather than to give people a choice).

116. *See, e.g.*, Vindu Goel, *Facebook Eases Privacy Rules for Teenagers*, N.Y. TIMES, Oct. 16, 2013, http://www.nytimes.com/2013/10/17/technology/facebook-changes-privacy-policy-for-teenagers.html ("But fundamentally, Facebook wants to encourage more public sharing, not less. The company, which has about its 1.2 billion users worldwide, is locked in a battle with Twitter and Google to attract consumer advertisers like food, phone and clothing companies."); James Temple, *Rules Against Tracking Don't Go Far Enough*, SAN FRANCISCO CHRON., Mar. 7, 2012, http://www.sfgate.com/business/article/Rules-against-online-tracking-don-t-go-far-enough-3387373.php ("Targeting ads based on search queries, sites visited, stories read and social connections forms the core of the multimillion-dollar business models of many online companies, including Google, Yahoo and Facebook.").

117. *See, e.g.*, Kesan & Shah, *supra* note 12, at 621 (arguing that setting browser defaults to reject cookies "would ensure that people understood the privacy risks of cookies"); Paul M. Schwartz, *Property, Privacy, and Personal Data*, 117 HARV. L. REV. 2055, 2100 (2004) ("This Article prefers an opt-in default because . . . it would place pressure on the better-informed party to disclose material information about how personal data will be used. This default promises to force the disclosure of hidden information about data-processing practices.").

collection by first parties or for diagnosis, treatment, or research purposes could be Track-Me. Such narrow defaults might respond to reasonable cost-benefit assessments of different tracking practices, but narrow defaults increase the likelihood that consumers will be confused about how the default works.

Existing proposals therefore suggest broader settings. The W3C preliminary draft proposal contemplates that Track-Me would be the default for all tracking, a policy default.[118] The FTC suggests a Don't-Track-Me default for all collection of sensitive information—"information about children, financial and health information, Social Security numbers, and precise geolocation data"—requiring express consent from consumers before the information can be collected.[119] For all other tracking over which consumers are given control—i.e., the collection of non-sensitive information by third parties (or affiliates of first parties where the affiliate relationship is not obvious to consumers) or by first parties tracking consumers across third-party websites, or for first parties sharing non-sensitive information with third parties—the FTC's proposed default is Track-Me.[120]

While the particular settings of the defaults policymakers would adopt is again unknown, it seems probable that policymakers will adopt a mix of Track-Me and Don't-Track-Me positions for those data collection practices that are within the control of the consumer. As discussed further below, this mix of settings could be confusing for many consumers. Its precision could also create the appearance that policymakers have carefully determined which collection practices are potentially harmful and which present no danger of harm, and have assigned each to a Don't-Track-Me or Track-Me default accordingly, potentially lulling consumers into a false sense that the default settings will protect them from any harm from tracking.

3. Granularity of Available Opt-Out Positions

Once policymakers decide to allow consumers to opt out of a particular type of information tracking, they then must decide the scope of the opt-out mechanism. Opting out of a default wholesale might be the only alternative given to consumers, or consumers might be permitted to opt out narrowly.

118. The W3C draft default requires general purpose browsers to be initially set to "Track-Me" or "No Preference," either of which would permit tracking. W3C TPE, *supra* note 104, § 3. However, a consumer can opt out by using a special-purpose browser that is marketed as a privacy-enhancing technology. *Id.*

119. FTC PRIVACY REPORT, *supra* note 106, at 59.

120. *Id.* at 40–42.

For example, a consumer might be given the power to decline tracking by third parties while accepting tracking by first parties, or to choose whether to be tracked on a website-by-website basis. To the extent that tracking is performed through cookie technology, common web browser setting choices today give consumers the ability to opt out on both of these bases. For example, in Firefox a consumer can choose to accept all cookies, accept no cookies, accept cookies from first parties but not from third parties, or make exceptions from each of these for particular websites.[121]

The broader the opt-out mechanism, the more easily it is understood by consumers, but the more likely that consumers will be unable to satisfy their textured preferences. Switching positions wholesale from Track-Me to Don't-Track-Me or vice versa each time a consumer encounters a tracking practice for which she wants a different setting is onerous and can lead to mistakes. For example, if a consumer in a Don't-Track-Me position changes tracking settings wholesale to facilitate a particular transaction, she must go back and change settings again to return to a Don't-Track-Me position, a step she may not remember to perform. A granular set of selective opt-out options would allow sophisticated consumers to satisfy particular preferences but presents the danger of overwhelming the average consumer.[122]

Given that the overarching goal of the use of personal data defaults is to facilitate individual choice about whether, when, and by whom to be tracked, policymakers will no doubt create a scheme by which consumers can selectively opt in or out of tracking defaults. The W3C and FTC proposals all permit consumers to, in effect, opt out on a firm-by-firm basis. For example, the FTC mobile device proposal suggests that the default for most information be Track-Me, that consumers be permitted to opt out from this wholesale, but that consumers then be permitted to opt back into tracking on an application-by-application basis:

> A DNT setting placed at the platform level could give consumers . . . a way to control the transmission of information to third parties as consumers are using apps on their mobile devices . . . . Offering this setting or control through the platform will allow consumers to make a one-time selection rather than having to

121. *See Block Websites from Storing Site Preferences or Login Status in Firefox*, MOZILLA SUPPORT, https://support.mozilla.org/en-US/kb/block-websites-storing-site-preferences (last visited Nov. 25, 2013). As previously noted, however, trackers can track through other means than cookies, such as through digital fingerprinting. *See* Tanner, *supra* note 33.

122. For example, when Facebook added more granular privacy controls, users became confused and more users kept the default settings. *See* Fred Stutzman et al., *Silent Listeners: The Evolution of Privacy and Disclosure on Facebook*, 4 J. PRIVACY & CONFIDENTIALITY 2, 23 (2013).

> make decisions on an app-by-app basis. Apps that wish to offer services to consumers that are supported by behavioral advertising would remain free to engage potential customers in a dialogue to explain the value of behavioral tracking and obtain consent to engage in such tracking.[123]

The W3C draft proposal is similar.[124] If the default were Don't-Track-Me, a comparable scheme might give consumers both a wholesale opt-out choice, allowing them to opt into the Track-Me position with respect to all firms, and granular choices, allowing them to agree to tracking on a firm-by-firm basis.

4. Altering and Framing Rules

Policymakers would then need to select altering and framing rules for these defaults. Recall that the goals of such rules are to prevent the default from being too sticky or too slippery and to inform consumers about the default and option to opt out. Given that the tracking defaults and opt-out choices policymakers appear inclined to embrace are complex, ensuring that consumers understand their choices and are able to act on them is a demanding task. While the theoretically possible altering and framing rules are limitless, this section sketches the general contours of rules that are on the horizon.

First, policymakers will put framing rules in place that aim to make the default and opportunity to opt out salient. For example, framing rules might require the default and opportunity to opt out to be "prominent"[125]—disclosed in words that "are of a type, size, and location sufficiently noticeable for an ordinary consumer to read and comprehend them, in print that contrasts highly with the background on which they appear."[126] The FTC proposal suggests that the choice to opt out should be given to consumers "at a time and in a context relevant to the consumer's decision about whether to allow data collection and use," such as "directly adjacent to where the consumer is entering his or her data" online, "immediately upon signing up for a service," or, for an offline transaction, "close to the time of sale"

123. FTC MOBILE REPORT, *supra* note 106, at 21.

124. W3C TPE, *supra* note 104, § 6.

125. *See* FTC PRIVACY REPORT, *supra* note 106, at 50; *see also* Decision and Order, *In re* ScanScout, Inc., File No. 1023185 (Fed. Trade Comm'n Dec. 14, 2011), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2011/12/111221scanscoutdo.pdf [hereinafter ScanScout Consent Order] (requiring "prominent" placement of a notification about a default and opt-out mechanism).

126. ScanScout Consent Order, *supra* note 125.

through notification on a "sales receipt" or "prominent poster at the location where the transaction takes place."[127]

Second, to forestall confusion, policymakers will enact altering rules that require the mechanism for opting out to be "easy to find and use."[128] More particularly, firms might be required to create an opt-out process that involves no more than one or two clicks of a button to opt out.[129] The opt-out process might be standardized, at least to some extent, across browsers, websites, mobile devices, or applications. Or, the process for opting back into the default after having opted out might be regulated so that firms cannot change a consumer's position through unread fine print. For example, while the W3C draft default scheme does not contain rules regarding the process for opting out of the default,[130] it would require that where a consumer has opted out at the browser level, opting back in to a Track-Me position with respect to any particular firm should be done "explicit[ly]."[131]

These rules aimed at salience and eliminating confusion would work simultaneously to inform consumers about the default and the opt-out opportunity. For example, the W3C preliminary draft requires firms that track consumers and all browsers to "clearly and accurately" with a "brief and neutral explanatory text" explain three things: (1) a consumer can opt out of some third-party tracking, (2) a consumer who has opted out may continue to be tracked for "permissible" purposes, and (3) a consumer who has opted out at the browser level can opt back in with respect to a particular

127. FTC PRIVACY REPORT, *supra* note 106, at 50.

128. FTC MOBILE REPORT, *supra* note 106, at 21.

129. *See* ScanScout Consent Order, *supra* note 125.

130. *See* W3C TPE, *supra* note 104, § 3. The draft states:

> We do not specify how tracking preference choices are offered to the user or how the preference is enabled: each implementation is responsible for determining the user experience by which a tracking preference is enabled. For example, a user might select a check-box in their user agent's configuration, install an extension or add-on that is specifically designed to add a tracking preference expression, or make a choice for privacy that then implicitly includes a tracking preference (e.g., Privacy settings: high). The user-agent might ask the user for their preference during startup, perhaps on first use or after an update adds the tracking protection feature. Likewise, a user might install or configure a proxy to add the expression to their own outgoing requests.

Id.

131. W3C TCS, *supra* note 104, § 6.

firm.[132] Where a consumer has opted out of tracking at the browser level, the consumer's opting back in selectively would need to be "informed."[133]

Third, policymakers might put altering rules in place intended to minimize the costs of opting out. For example, rules might require that consumers be given a "universal" opt-out mechanism at the browser level or, if technologically feasible, the device level.[134] (A universal mechanism would also present consumers with a broadly-bracketed choice, perhaps leading to more privacy-minded decisions.) To keep the cost of opting out low, rules might require the opt-out mechanism be "persistent,"[135] meaning that firms could not require consumers to opt out repeatedly the way that cookie-based opt-out systems do today.

In theory, a default scheme might also prohibit firms from giving consumers incentives to agree to tracking. The law might require firms to treat consumers in the Don't-Track-Me position the same as consumers in the Track-Me position, regardless of the particular purposes for which the firm tracks consumers. However, when the content and features of a website or app depend on revenue generated by tracking, as many currently do, policymakers will be reluctant to undermine this arrangement. For this reason, the W3C preliminary draft explicitly permits firms to condition services on consumers agreeing to be tracked.[136] With a narrow exception for "important product[s] with few substitutes, such as a patented medical device," the FTC proposal also asserts that firms should be permitted to condition access to goods, services, website content, etc. on consumers agreeing to tracking and to offer tracked consumers lower prices or other benefits.[137]

Moreover, policing the line between differential treatment intended to force consumers to agree to tracking and differential treatment that is the necessary consequence of refusing to be tracked would be difficult. Apps or websites could be optimized to function when tracking is enabled, as some websites already do today. For example, many websites warn consumers: "If you turn cookies off, you will not have access to many features that make

132. *Id.* § 3.

133. *Id.* § 6.

134. FTC MOBILE REPORT, *supra* note 106, at 21 (suggesting a universal mechanism to opt out of the Track-Me default be made available at the mobile device level).

135. *Id.*

136. W3C TPE, *supra* note 104, § 1 ("Web sites that are unwilling or unable to offer content without such targeted advertising or data collection need a mechanism to indicate those requirements to the user and allow them (or their user agent) to make an individual choice regarding exceptions.").

137. FTC PRIVACY REPORT, *supra* note 106, at 52.

your user experience more efficient and some parts of our website will not function properly."[138] One goes so far as to warn consumers that choosing not to be tracked will "spoil[] your experience of the website."[139] Google's "Privacy Policy" specifically indicates that its search engine might not display websites in the consumer's language automatically if the consumer disables cookies.[140] Alternatively, a website might deliver behaviorally-targeted ads to consenting consumers and a larger quantity of particularly distracting generic ads to untracked consumers. It might not be possible to detect when the larger quantity or greater distraction is necessary to produce revenue equivalent to behavioral ads and when it is harassment to induce consumers to agree to tracking. Thus, even if altering rules prohibited treating consumers who refuse to be tracked any differently than consumers who agree to be tracked, enforcing this prohibition would likely prove impossible.

* * *

In sum, while the precise default scope, settings, opt-out options, and altering and framing rules policymakers would choose if they were to enact a tracking default scheme are not certain, a few key features can be anticipated. These include the following:

1. the scope of the defaults will be limited, in that there will be some tracking from which consumers cannot opt out;
2. the settings will be a mix of Track-Me and Don't-Track-Me defaults;
3. the control consumers are given will be granular; and
4. altering rules will not prohibit trackers from giving consumers incentives to agree to stay in or move to the Track-Me position.

IV. DEFAULTS IN PRACTICE

Defaults in practice do not always live up to the theory behind using defaults in policymaking. This Part describes two failed default schemes,

138. *Consumer Online Privacy Frequently Asked Questions*, U.S. BANK, https://www.usbank.com/privacy/faq.html (last visited July 30, 2013); *see also Disney Registration Cookies Policy*, DISNEY https://registration.disneyinternational.com/cookiepolicy.htm?p=130&fullScreen=true (last visited Nov. 25, 2013) (containing similar warning language); INFORMATION ABOUT COOKIES ON MONSTER, http://inside.monster.com/cookie-info/inside2.aspx (last visited Nov. 25, 2013) (containing a similar warning).

139. *BBC Worldwide—Cookies Policy*, BBCWORLDWIDE.COM, *available at* http://www.bbcworldwide.com/cookies.aspx (last visited Nov. 29, 2013).

140. *Privacy Policy*, GOOGLE, http://www.google.com/policies/privacy (last modified June 24, 2013) ("However, it's important to remember that many of our services may not function properly if your cookies are disabled. For example, we may not remember your language preferences.").

demonstrates how firms have frustrated these schemes by making the defaults too sticky or too slippery, and, through comparison to examples of relatively successful defaults, extracts a set of conditions under which defaults do not perform in accordance with theory.

A. TWO FAILED DEFAULT SCHEMES

Two default schemes that have failed to meet policymaker aims are those for (1) financial institutions' sharing and use of consumer information[141] and (2) banks' charging of overdraft fees for ATM and nonrecurring debit transactions. The former is too sticky and the latter is too slippery. Neither appear to lead to well-informed consumer decisions about whether to stick with the defaults or opt out.[142]

1. The Financial Information Default Scheme

By default, financial institutions collect, use, and share information about their customers for marketing, pricing, and other purposes.[143] Under federal law, consumers can opt out of this in three respects. First, they can opt out to prevent an institution from sharing their personal information with parties that are not affiliated with the institution.[144] Second, they can opt out to prevent an institution from sharing with its affiliates information about the consumer other than information about the institution's own transactions

141. This is not single mandated default but rather collectively represents the requirements of the Gramm-Leach-Bliley Act (GLBA), 15 U.S.C. § 6802 (prohibiting certain sharing of information with nonaffiliated third parties unless consumers are given the right to opt out), and the requirements of the Fair Credit Reporting Act (FCRA) to avoid being treated as, and therefore subject to the extensive duties placed on, a credit reporting agency, 15 U.S.C. §§ 1681a(d), 1681a(f), 1681s-3 (defining an entity that regularly furnishes consumer reports to others as a consumer reporting agency, but excluding from the definition of consumer report the sharing of certain consumer information among affiliated entities if the consumer is given the option to opt out of this sharing and of the use of this and other information by the affiliate for marketing purposes). The GLBA requires institutions to inform consumers of their GLBA and FCRA opt-out rights. 15 U.S.C. §§ 6803, 6804. For ease of reference, these are treated here as a single default scheme.

142. By other metrics, these defaults may have done some good. For example, they may have led more financial institutions to share less consumer information or more banks to stop charging overdraft fees than would otherwise have occurred. *See* Peter P. Swire, *The Surprising Virtues of the New Financial Privacy Law*, 86 MINN. L. REV. 101 (2002) (making this argument as to the financial information defaults). But, by the metrics of well-informed consumer decisions and, for the overdraft default, a significant reduction in overdraft fees for low-income consumers, the ostensible goals of policymakers, these defaults have failed.

143. *See, e.g.*, Fred H. Cate, *Personal Information in Financial Services: The Value of a Balanced Flow*, Financial Services Coordinating Council 3–22 (2000), *available at* http://www.aba.com/aba/PDF/cate.pdf (surveying the many ways in which financial institutions use consumer information).

144. Privacy of Consumer Financial Information, 16 C.F.R. § 313 (2014).

with the consumer.[145] Third, they can opt out to prevent an institution's affiliates from using transaction and "other" information for marketing purposes.[146] Even if a consumer has opted out to the fullest extent, financial institutions can share all gathered information with joint marketing partners, transaction information with affiliates for non-marketing purposes, and "as [further] permitted by law."[147]

The default scheme includes altering and framing rules intended to educate consumers and facilitate opting out. Institutions must allow consumers to opt out at any time and must provide consumers with a "reasonable means" to do so.[148] Institutions must give consumers initial and annual notices explaining their opt-out rights.[149] These notices must be "clear and conspicuous," meaning "reasonably understandable" (in plain language and easy to read) and "designed to call attention to the nature and significance of the information."[150]

But, despite these rules, the financial information defaults are overly sticky and the scheme is not information-forcing. Although consumers generally do not like banks sharing their information with affiliates or third parties,[151] almost no one opts out.[152] Consumers reviewing model explanatory notices in laboratory conditions poorly understand the defaults and opt-out provisions.[153] Comprehension may be even lower under real-world conditions, in which many consumers will not read the notices.

2. The Checking Account Overdraft Default Scheme

As explained above, federal banking regulators effectively require banks to default consumers out of expensive overdraft coverage for ATM and

145. *Id.*
146. *Id.*
147. 16 C.F.R. § 313.15(4) (2014).
148. *Id.* § 313.7.
149. *Id.* § 313.4–.5.
150. *Id.* § 313.3.
151. Cate, *supra* note 143, at 15 (in consumer testing, finding that respondents "do not seem to like their information being shared with nonaffiliates . . . or affiliates").
152. John Martin, *Opting Out—or Not*, ABCNEWS (June 21, 2001), http://more.abcnews.go.com/sections/wnt/dailynews/privacy_notices_010621.html (finding only .5% of people opt out).
153. Alan Levy & Manoj Hastak, *Consumer Comprehension of Financial Privacy Notices*, INTERAGENCY NOTICE PROJECT, 9 tbl.1 (2008), http://ftc.gov/privacy/privacyinitiatives/Levy-Hastak-Report.pdf (showing that less than half the subjects tested were able to select a bank for a cogent and relevant reason based on even the best form notice regulators could develop).

nonrecurring debit card transactions.[154] Altering and framing rules aim to prevent bank "circumvention or evasion" of the default.[155] First, to prevent banks from placing language opting out of the default in routinely unread account disclosures, opting out requires an "affirmative" accountholder action, such as speaking to a bank representative in person or by phone or clicking a box on an online banking form.[156] Second, banks must provide the same account terms, conditions, and features to accountholders who stick with the default as they provide to accountholders who opt out.[157] Framing rules require banks to provide consumers with specific information about the default and the consequences of opting out in a document or webpage segregated from all other documents or webpages.[158]

In promulgating the overdraft default, regulators explicitly stated that they intended for it to operate as a policy default, akin to the auto-enrollment default for retirement savings.[159] But it appears that the majority of the consumers whom regulators intended to assist—low-income frequent users of overdraft[160]—opt out of the default.[161] The rule also does not function as a penalty default; surveys indicate that consumers who opt out hold key misconceptions about the way the default and opt-out positions work.[162]

---

154. Electronic Fund Transfers, 12 C.F.R. § 205.17(b) (2014). Fees on bank-covered overdrafts occasioned by other types of transactions (chiefly checks and recurring payments) are not included in the policy default because these tend to be for necessities and, if not paid, can result in bounced check or late payment fees. *See* Supplement I to Part 205, Official Staff Interpretations of 12 C.F.R. § 205.17(b)(2), cmt. 2 (2014). ATM and nonrecurring debit transactions tend to be discretionary transactions and when these are declined consumers are not charged a fee. Willis, *supra* note 34, at 1180.

155. *See, e.g.*, Electronic Fund Transfers, 74 Fed. Reg. 59,033, 590,44 (Bd. of Governors of the Fed. Reserve Sys. Nov. 17, 2009).

156. *See* One-Time Debit Card Transactions, 12 C.F.R. § 205.17(b)(1)(iii) (2014); 12 C.F.R. pt. 205, supp. I, § 205.17(b)(1)–(4).

157. 12 C.F.R. § 205.17(b)(3).

158. *Id.*

159. Regulators noted that "studies have suggested [that] consumers are likely to adhere to the established default rule, that is, the outcome that would apply if the consumer takes no action" and cited studies of the effectiveness of automatic enrollment in increasing participation in retirement savings plans. Electronic Fund Transfers, 74 Fed. Reg. at 59,038 & n.25.

160. In 2009, one widely-cited industry consultant estimated that ninety percent of overdraft fees were paid by the poorest ten percent of checking accountholders. *See Debit Card Trap*, N.Y. TIMES, Aug. 20, 2009, http://www.nytimes.com/2009/08/20/opinion/20thu1.html (citing Michael Moebs).

161. *See* Willis, *supra* note 34, at 1184.

162. *See* CTR. FOR RESPONSIBLE LENDING, BANKS COLLECT OVERDRAFT OPT-INS THROUGH MISLEADING MARKETING, RESEARCH BRIEF 2, 6 n.8 (2011), *available at* http://www.responsiblelending.org/overdraft-loans/policy-legislation/regulators/CRL-OD-Survey-Brief-final-2-4-25-11.pdf. The brief states:

B. HOW FIRMS MAKE THESE DEFAULTS FAIL

An examination of how financial institutions and banks present these defaults to consumers in practice demonstrates why these defaults have failed to achieve policymaker goals. Institutions work to bolster the mechanisms that can make defaults sticky to ensure that very few consumers opt out of the financial information defaults; banks work to undermine these mechanisms to encourage accountholders to opt out of the overdraft-coverage default.

1. Transaction Barriers

Transaction barriers that can contribute to the stickiness of defaults—costs, confusion, and futility—can be built higher, eliminated, or even inverted.

In the case of the financial information defaults, institutions prefer for consumers to stay in the default information-sharing position, and therefore build high transaction barriers to opting out. For consumers who attempt to opt out, the altering rule requiring "reasonable means" for opting out keeps the transaction costs of doing so for any one institution fairly low.[163] But consumers must opt out with each financial institution with which they have any business. If transaction costs are not enough, institutions warn consumers that opting out will be costly in other ways:

> If you opt out:
>
> We may need you to repeat information that you have already provided and we may not be able to pre-fill applications for you.
>
> We may have to transfer your phone calls more often.

> Sixty percent (60%) of consumers who opted [out of the default] stated that an important reason they did so was to avoid a fee if their debit card was declined. In fact, a declined debit card costs consumers nothing . . . Sixty-four percent (64%) of consumers who opted [out of the default] stated that an important reason they did so was to avoid bouncing paper checks. The truth is that the opt-in rules cover only debit card and ATM transactions.

Id.

163. The only challenge may be that institutions often require consumers to provide account numbers. *See, e.g.*, *Statutory Form of Opt-Out Notice*, FIRST FOUNDATION, https://www.ff-inc.com/privacy/opt-out-notice.aspx (last visited Nov. 25, 2013); *Opting Out of Information Sharing*, METLIFE, https://eforms.metlife.com/wcm8/PDFFiles/730.pdf (last visited Nov. 25, 2013). Locating these might take some time, particularly for closed accounts.

> We may not offer you the products that best meet your needs.[164]

Another transaction barrier is confusion. As noted above, consumers understand the defaults and their opt-out rights poorly, even after reading the required notices. Consumers may not understand that they can opt out at all; even in the flurry of publicity when the notices first went out, fewer than thirty-five percent of consumers surveyed recalled receiving them.[165] Although the law requires that institutions give consumers initial and annual "conspicuous" notices, these arrive with other documents from the institution, documents that may disappear among the reams of disclosures consumers receive, and routinely ignore, in their daily lives.[166]

Finally, financial institutions may capitalize on futility. Even if a consumer has opted out to the fullest extent, financial institutions can continue to share information with joint marketing or other service providers,[167] share information with affiliates for non-marketing purposes,[168] and otherwise share information "as authorized by law."[169] Unable to opt out entirely, consumers might resign themselves to the defaults.

In comparison, banks structure the presentation of the overdraft default and the process for opting out to have the opposite effect. Banks eliminate transaction costs for many consumers and even make it more costly to stick with the default than to opt out. For new customers and for accountholders using online banking, transaction costs do not fortify the default because these costs are the same whether the consumer sticks with the default or opts out; new accountholders in the process of opening an account or existing accountholders attempting to access online banking must check precisely the same number of boxes regardless of whether they check the box for sticking with the default or for opting out.[170] When the default first became law, some banks flooded existing accountholders, particularly those who had

164. *How We Protect You*, UNITED SERVS. AUTO. ASS'N, https://www.usaa.com/inet/ent_references/CpStaticPages?page=ya_efsk_privacyinfo_adult_pub (last visited Apr. 13, 2014).

165. *See* Fred H. Cate, *The Failure of Fair Information Practice Principles*, *in* CONSUMER PROTECTION IN THE AGE OF THE 'INFORMATION ECONOMY' 341, 360 (Jane Winn ed., 2006) (citing Star Systems, *Financial Privacy: Beyond Title V of Gramm-Leach-Bliley* 9 (2002)).

166. *See* Omri Ben-Shahar & Carl E. Schneider, *The Failure of Mandated Disclosure*, 159 U. PA. L. REV. 647, 665 (2011).

167. 16 C.F.R. § 313.13 (2014).

168. 16 C.F.R. § 313.11.

169. 16 C.F.R. § 313.15(a)(7)(iii).

170. *See* Ben Popken, *Banks Luring You Into Signing Back up for High Overdraft Fees*, THE CONSUMERIST (June 18, 2010), http://con.st/10007945; Phil Villarreal, *When it Comes to Overdraft Opt-In, Chase Won't Take No for an Answer*, THE CONSUMERIST (Aug. 6, 2010), http://con.st/10009792.

overdrafted frequently in the past, with print marketing, in-person bank teller appeals, and telemarketing encouraging them to opt out.[171] The barrage would end only when the consumer opted out,[172] making it costlier to stick with the default—and continue to endure the marketing—than to opt out.[173]

Banks have forestalled the confusion that might otherwise make the overdraft default sticky by ensuring that the opt-out process is salient and easy to use. The volume of overdraft marketing means the option to opt out is difficult to miss. Existing accountholders have been able to opt out by simply pushing a button on an ATM,[174] clicking a button online, or saying "yes" to a bank employee who called to suggest to accountholders that they ought to opt out.[175] Any confusion would run toward opting out, for two reasons. First, banks have framed the opt-out position as a voluntary perk, asking accountholders whether they would like to take advantage of the bank's "courtesy pay"[176] or "account protector"[177] service. Second, banks have capitalized on the fact that the default setting is itself confusing, in that it covers nonrecurring debit and ATM transactions but not checks or recurring debits. Some banks even call overdraft coverage "bounce protection";[178] given that "bounce" is a term usually applied to a dishonored check, the nomenclature probably leads consumers to mistakenly believe that they must opt out in order to receive overdraft coverage for bounced checks.[179]

171. *See* Karen Weise, *Reforms Fail to Halt Bank Revenue on Debit-Card Overdraft Fees*, BLOOMBERG (Oct. 20, 2011), http://www.bloomberg.com/news/2011-10-20/reforms-fail-to-halt-bank-revenue-on-debit-card-overdaft-fees.html; Ben Popken, *Chase Just Goes Ahead and Adds Overdraft Protection to Your Account*, THE CONSUMERIST (Sept. 16, 2010), http://con.st/10011137.

172. *See* Villarreal, *supra* note 170.

173. CTR. FOR RESPONSIBLE LENDING, *supra* note 162, at 3–4 (finding that almost half of surveyed consumers who reported that they opted out of the default said they did so at least in part to stop receiving overdraft marketing).

174. *See* Laura Northrup, *Opt In to Overdraft Protection Right at the ATM*, THE CONSUMERIST (July 29, 2011), http://con.st/10021347.

175. *See* Laura Northrup, *Chase Now Has Human ATM Greeter Who Helpfully Sells Overdraft Protection*, THE CONSUMERIST (July 14, 2010), http://consumerist.com/2010/07/14/chase-installs-atm-greeter-who-sells-debit-card-overdraft-protection.

176. *Courtesy Pay*, SAN MATEO CREDIT UNION, http://www.smcu.org/accounts/courtesy.php (last visited Aug. 1, 2013).

177. *Sovereign Account Protector*, SOVEREIGN BANK, http://www.sovereignbank.com/personal/promotions/sovereign-account-protector.asp (last visited Aug. 1, 2013).

178. *Bounce Protection*, CENTRAL NATIONAL BANK, https://www.centralnational.com/personal/bounceprotection.asp (last visited Dec. 1, 2013).

179. *See* CTR. FOR RESPONSIBLE LENDING, *supra* note 162, at 6 n.8 (finding this to be a common misconception).

2. Judgment and Decision Biases

Firms that benefit from defaults can harness judgment and decision biases to keep consumers in those defaults. But firms that oppose defaults can defuse these biases or even flip them to push consumers out of the default position.

Institutions faced with the financial information defaults work to ensure that biases support the defaults. For example, one financial institution's opt-out notice appears designed to trigger loss aversion and the endowment effect. It explains:

> [We are] known for [our] exceptional member service. Sharing member information as we have outlined here enables us to maintain this service excellence . . . .
>
> However, federal law also requires that we allow you to opt out . . . . Limiting our ability to share financial information . . . will make it difficult for us to serve you as you might expect.[180]

This reminds consumers that sticking with the defaults will "maintain" the status quo, suggests that the defaults ought to form consumers' expectations (the reference point from which consumers should measure gains and losses), and warns that opting out will cause consumers to lose benefits they now have.[181]

Next, financial institutions encourage procrastination and decision avoidance. The law requires consumers to be given the opportunity to opt out at any time so as to reduce transaction barriers. But institutions may emphasize this fact to encourage procrastination[182]—if a consumer can opt out at any time, she need not take action immediately. Further, although framing rules require the opt-out notices themselves to be "reasonably understandable," institutions can make the opt-out decision appear complex and overwhelming. Some institutions offer a plethora of "privacy policies" that consumers must wade through to understand their opt-out rights.[183]

180. *How to Enable Your Cookies*, *supra* note 164.

181. *See also* Janger & Schwartz, *supra* note 15, at 1243 (discussing specific ways in which financial institutions frame the choice to opt out of the Gramm-Leach-Bliley default as a loss).

182. *Privacy Notice*, CAPITALONE, http://www.capitalone.com/media/doc/corporate/english-privacy-notice.pdf (last visited Aug. 1, 2013) ("You can contact us at any time to limit our sharing.").

183. *Privacy Overview*, METLIFE, https://www.metlife.com/about/privacy-policy/index.html (last visited Nov. 25, 2013) (listing on its "privacy" webpage: an "Online Privacy Policy," a "Customer Privacy Policy," an "Auto & Home Privacy Policy," a "HIPAA Notice

Others surround the required notice with voluminous "explanations" that are difficult to read and understand.[184]

In the face of the overdraft default, on the other hand, banks have waged a marketing campaign designed to negate or reverse the biases that might otherwise give the default traction. Banks have attempted to reposition loss aversion and the endowment effect to encourage opting out using two strategies. One was to pitch opting out prior to the date on which the new legal default rule became effective, thus framing the choice as between keeping an existing endowed position or accepting a change by agreeing to the new default.[185] In their marketing, banks explicitly invoked loss aversion to encourage opting out with copy such as "Don't lose your ATM and Debit Card Overdraft Protection"[186] and asking accountholders whether they wanted to "keep [their] account working the same" or "change [their] account."[187] The other has been to frame opting out not as losing an endowed position but as gaining the bank's overdraft "service" by "opting in."[188]

---

of Privacy Practices for Personal Health Information," as well as a link for "Opting Out of Information Sharing").

184. For example, one institution's webpage accompanying its opt-out notice uses the following, over 100-word sentence:

> However if you do not want us, your financial advisor or your bank, credit union or other financial institution to disclose your personal information to the New Financial Institution, and if you do not want your financial advisor or your bank, credit union or other financial institution to retain copies of your personal information when your financial advisor or your bank, credit union or other financial institution terminates his, her or its relationship with us, you may request that we, your financial advisor and your bank, credit union or other financial institution limit the information that is shared with the New Financial Institution by filling out the Privacy Choices Notice and mailing it to [address].

LPL Privacy Policy and Opt-Out Information, WEBSTER BANK, https://www.webster online.com/personal/products-services/investment-services/LPL-privacy-policy.html (last visited Aug. 1, 2013).

185. *See, e.g.*, Andrew Martin & Ron Lieber, *Banks Apply Pressure to Keep Fees Rolling in*, N.Y. TIMES, Feb. 23, 2010, *available at* http://www.nytimes.com/2010/02/23/your-money/credit-and-debit-cards/23fee.html.

186. *Don't Lose Your ATM & Debit Card Overdraft Protection*, LAPEER COUNTY BANK & TRUST COMPANY, http://www.lcbt.com/2747/mirror/debitcardandatmoverdraftprotection.htm (last visited Aug. 1, 2013).

187. *Stay Protected with Shareplus ATM and Debit Card Overdraft Coverage*, SHAREPLUS FED. BANK, *available at* https://secureforms.c3vault1.com/forms/shareplus/pdf/opt-in-details.pdf (last visited Aug. 1, 2013).

188. *See, e.g.*, Jim Bruene, *Debit Card Overdraft Protection: 2 Steps Forward, 1.9 Back*, NETBANKER (July 13, 2010), http://www.netbanker.com/2010/07/debit_card_overdraft_protection_2_steps_forward_19_back.html (providing graphic from credit union

Second, banks harness choice bracketing and play on discounting to spur opting out. Consumers are given the choice whether to accept overdraft coverage for any and all ATM and debit transactions that might overdraft the account, rather than on a transaction-by-transaction basis.[189] This broad choice bracketing directs consumers' focus to the question of whether they might *ever* need overdraft coverage (e.g., for an emergency), favoring opting out. If instead, consumers were faced with narrow decisions about whether to accept overdraft coverage and fees for particular transactions, consumers could selectively use overdraft for emergencies and decline it for non-emergencies. Broad choice bracketing also makes the benefits of opting out appear immediate and certain and the costs delayed and uncertain. Bank advertising includes themes along the lines of "Privilege Pay works like a safety net for your checking account . . . so you don't get left stranded at a gas station"[190]—thus offering accountholders immediate "peace of mind"[191] that funds will be available in an emergency if the consumer opts out.[192] In contrast, banks downplay the costs of overdrafting, emphasizing that opting out of the default and into the bank's overdraft program is a "free" perk.[193]

Third, rather than allowing procrastination, decision avoidance, and salience and omission biases to lead to inertia, banks place some consumers in a mandated choice scenario and have given others deadlines or encouragement to act. As explained above, some banks require new customers and accountholders attempting to use online banking to opt in or out of overdraft coverage before they can open an account or continue to use online banking. These accountholders are forced to take action and cannot procrastinate or avoid making the decision. For existing accountholders who did not use online banking, bank marketing framed the

that advertises "Opt **In** for Overdraft Coverage on Debit and ATM Cards. . .; CLICK HERE TO OPT **IN**" (emphasis added)).

189. 12 C.F.R. § 205.17 (2014).

190. *Privilege Pay*, FARMERS INSURANCE GROUP FED. CREDIT UNION, https://www.figfcu.com/print.php?id=610 (last visited Aug. 1, 2013).

191. *See, e.g.*, *Stay Protected with Shareplus ATM and Debit Card Overdraft Coverage*, *supra* note 187 ("STAY PROTECTED . . . MAINTAIN PEACE OF MIND"); *Sovereign Account Protector*, *supra* note 177 ("Enjoy the peace of mind of knowing your checks, debits, and payments are automatically honored by setting up an Overdraft Protection Plan.").

192. *Privilege Pay*, *supra* note 190 ("Privilege Pay works like a safety net for your checking account . . . so you don't get left stranded at a gas station . . . .")

193. *See, e.g.*, *Overdraft Protection*, CAPITALONE, http://www.capitalone.com/bank/overdraft-protection/ (last visited Nov. 25, 2013) ("Opting in is free and easy."); *Check Card Overdraft Protection for Your Wescom Checking Account*, WESCOM CREDIT UNION, https://www.wescom.org/accounts/whyoptinoverdraftprotection.asp (last visited Nov. 25, 2013) ("Why Opt in to Check Card Overdraft Protection? It's Free.").

decision as one that must be made immediately or by a certain deadline.[194] Bank marketing also trumpets "It's your choice!,"[195] implying that sticking with the default is not a blameless omission.

Fourth, banks use explicit messaging so that the illusion of control instigates opting out. Bank marketing emphasizes that consumers are "in control" of their overdraft decisions[196] and implies that opting out gives consumers more control than sticking with the default.[197] Because the feeling of control encourages riskier behavior, to the extent that consumers understand that they are taking a risk of incurring overdraft fees, these marketing messages could encourage them to opt out.

3. Preference Formation Effects

As with transaction barriers and biases, the preference-forming effects of defaults can be made stronger or weaker.

Institutions bolster any advice implicit in the financial information defaults with explicit advice to consumers that their privacy is already protected and they need not opt out.[198] Documents and webpages accompanying the financial information default notices commonly emphasize foremost that the institution cares about the consumer's privacy.[199] For example, although a close read of one institution's required notice reveals that the institution shares consumer information to the fullest extent

194. *Urgent Notice Regarding Your Public Service Credit Union Debit/ATM Card*, PUBLIC SERVICE CREDIT UNION, https://www.mypscu.com/docs/odpletteronline.pdf (last visited Aug. 1, 2013) ("Urgent notice regarding your . . . Debit/ATM Card. Your immediate response is needed!"); Bruene, *supra* note 188 (providing graphic from credit union that advertises "August 15 is the deadline to apply if you choose to keep coverage").

195. *See, e.g.*, *Overdraft Protection*, *supra* note 193.

196. *Overdraft Privilege: Stay Protected*, FIRST FIN. BANK, https://www.bankatfirst.com/personal/spending-account-options/overdraft-privilege-opt-in.aspx (last visited Aug. 1, 2013) ("You now control whether or not you want to continue overdraft privilege coverage for ATM withdrawals and everyday check card transactions.").

197. *Overdraft Privilege*, FIRST CMTY. CREDIT UNION, https://www.fccu.org/Resources/PDFs/Overdraft%20Privilege_3_13.pdf (last visited Aug. 1, 2013) ("Keep Your Oversights Under Control"); *Debit Card Overdraft Services*, WEBSTER BANK, https://www.websteronline.com/personal/products-services/checking-services/debit_card_overdraft_services.html (last visited Aug. 1, 2013) ("Giving you control for your everyday debit card purchases.")

198. *See, e.g.*, *Our Privacy Promise to You*, UNITED SERVS. AUTO. ASS'N, https://www.usaa.com/inet/ent_references/CpStaticPages?PAGEID=cp_netprivacy_pub (last visited Aug. 1, 2013) ("If you decide that USAA's rigorous practices meet your privacy expectations, **no further action is required**.").

199. *See, e.g.*, *Wells Fargo Privacy and Security*, WELLS FARGO, https://www.wellsfargo.com/privacy_security/privacy (last visited Aug. 1, 2013) (beginning with "[w]e're committed to protecting your privacy").

permitted by law (i.e., with joint marketing partners, affiliates, and non-affiliates, and for both marketing and non-marketing purposes), the webpage from which this notice can be accessed begins boldly: "SAFEGUARDING YOUR PRIVACY" and continues "[THIS INSTITUTION] TAKES OUR COMMITMENT TO PROTECTING YOUR PRIVACY SERIOUSLY."[200]

Institutions explicitly advise consumers that they will benefit by not opting out. For example, one institution explains that sharing information with affiliates provides customers with the following benefits:

1. Prevention of unauthorized transactions or fraud.
2. Account upgrades with additional benefits.
3. Offers for products and services specifically suited to your individual situation. . . .
4. Increased convenience, making it faster and easier for you to do business with us. . . .
5. Enhanced customer service and responsiveness.[201]

The implication is that consumers who opt out of information sharing will not receive these benefits.

In contrast, banks present the overdraft default in a way that negates the preference-forming effects of defaults. Banks typically present accountholders with two checkboxes, one for "opting in" to the bank's "overdraft protection" and another for "opting out," thus concealing which position is the default.[202] Further, banks advise accountholders that the opt-out position, rather than the default, is in accountholders' best interests, with copy such as "overdraft privilege is designed with you in mind."[203]

Finally, before the overdraft policy default came into effect, banks pressured existing accountholders to opt out of the overdraft policy before they had a chance to experience the new default. Similarly, banks force new customers to make a decision when they open an account. Experience, therefore, does not induce accountholders to choose the default, as they have not had the opportunity to discover that they can manage without purchases

200. *Privacy Protection*, CAPITALONE, http://www.capitalone.com/identity-protection/privacy (last visited Nov. 29, 2013).

201. *Regions Privacy FAQs*, REGIONS, http://www.regions.com/about_regions/all_facts.rf (last visited Aug. 1, 2013).

202. *See, e.g.*, *First Commerce Credit Union Opt-In/Opt-Out Form*, FIRST COMMERCE CREDIT UNION, *available at* http://www.firstcommercecu.org/accounts_resources/consumer_education/new_regulation_over_draft_protection_options_your_action_is_required/opt_in_opt_out_form#form (stating "Please complete this form to Opt-in or Opt-out of overdraft protection" and providing two check boxes, one labeled "OPT-IN" and the other labeled "OPT-OUT").

203. *Overdraft Privilege: Stay Protected*, *supra* note 196.

that would overdraft their accounts or that alternative, cheaper sources of overdraft coverage are available.

C. SUCCESSFUL DEFAULTS

Of course, many default schemes do work as intended. Two well-known examples are the above-mentioned retirement savings auto-enrollment default and the Do Not Call registry.

1. Automatic Enrollment in Retirement Savings Plans

Auto-enrollment allows employers to default employees into participation in defined contribution pension plans.[204] This policy default scheme includes framing rules designed to ensure that the defaults are not too sticky, including a requirement that every employee who is enrolled by default be given regular notices of the right to opt out.[205] These notices must be written at a level that the average employee can understand, so that the opportunity to opt out is not confusing.[206]

Auto-enrollment has been extremely successful in its goal of increasing the number of employees in the default position.[207] As noted above, employers that have made participation in their plans the default have increased their employee participation rates dramatically.[208] The increase has been largest for lower-income consumers, which has been interpreted as evidence that defaults are most helpful for those who need the most help.[209]

---

204. *Retirement Topics–Automatic Enrollment*, INTERNAL REVENUE SERV., http://www.irs.gov/Retirement-Plans/Plan-Participant,-Employee/Retirement-Topics---Automatic-Enrollment (last visited Aug. 1, 2013).

205. Safe Harbor Requirements, 26 C.F.R. § 1.401(k)-3(d)(2) (2014).

206. *Id.* § 1.401(k)-3(d)(2)(i)(B).

207. However, the default scheme has decreased the average amount individual employees save, contrary to policymaker aims. *See* Ryan Bubb & Richard H. Pildes, *How Behavioral Economics Trims Its Sails and Why*, 127 HARV. L. REV. (forthcoming 2014).

208. *See, e.g.*, Nessmith et al., *supra* note 88, at 6 (comparing enrollment rates for automatic enrollment plans (86%) with voluntary enrollment plans (45%)).

209. *See* John Beshears et al., *Public Policy and Saving for Retirement: The "Autosave" Features of the Pension Protection Act of 2006*, *in* BETTER LIVING THROUGH ECONOMICS 274, 287 (John J. Siegfried ed., 2010) (noting "clear and compelling evidence that automatic enrollment was an effective means of increasing savings and improving economic well-being, particularly of minorities and of the poor"). This interpretation may be erroneous, as there are some low-income employees who participate because of automatic enrollment but for whom participation is not optimal. The details of this argument are beyond the scope of this Article.

2. Do Not Call

Another popular default scheme is the Do Not Call Registry. By default, telemarketers can call people to try to sell to them, but consumers can stop most of these calls by signing up for the Do Not Call Registry. This process is well known, low cost, and easy—consumers can call a toll-free number or register online.[210] Consumers can opt back into the default wholesale by removing their number from the list,[211] or they can provide written explicit consent to selectively opt back into the default and allow a particular firm to telemarket to them.[212] Because this consent must be in writing, the effect is that telemarketers cannot call consumers and orally attempt to convince them to opt back in.

From its inception, the Do Not Call Registry was heavily publicized, and the public responded. Despite having to take some action to opt into the list, consumers placed ten million phone numbers on it in the first four days it was operative,[213] and today over seventy percent of Americans have placed their numbers on the list.[214] Further, it appears that the default sorts consumers reasonably well. On average, those consumers who do not sign up have the least to gain by doing so, because they do not receive many telemarketing calls; those who have much to gain sign up.[215]

---

210. *FTC Approves Two Reports to Congress on the National Do Not Call Registry*, FED. TRADE COMM'N (Jan. 4, 2010), http://www.ftc.gov/opa/2010/01/donotcall.shtm ("[R]esearch has consistently shown widespread public awareness of the program and a steady increase in the number of phone numbers registered."); *see also Register Your Home or Mobile Phone Number*, NAT'L DO NOT CALL REGISTRY, https://www.donotcall.gov/register/reg.aspx (last visited Aug. 5, 2013)(National Do Not Call online registration form).

211. *National Do Not Call Registry*, FED. TRADE COMM'N, http://www.consumer.ftc.gov/articles/0108-national-do-not-call-registry (last visited Aug. 5, 2013).

212. Abusive Telemarketing Acts or Practices, 16 C.F.R. § 310.4 (b)(1)(iii)(B)(*i*) (2014).

213. *Ten Years of Do Not Call*, FED. TRADE COMM'N, http://www.consumer.ftc.gov/articles/0108-national-do-not-call-registry (last visited Aug. 5, 2013).

214. As of 2007, seventy-two percent of Americans had placed their phone numbers on the list. COUNCIL OF ECON. ADVISORS, 2009 ECONOMIC REPORT OF THE PRESIDENT 244 (2009), *available at* http://georgewbush-whitehouse.archives.gov/cea/ERP_2009_Ch9.pdf.

215. *See* Goh Khim-Yong et al., *Consumer Heterogeneity, Privacy, and Personalization: Evidence from the Do-Not-Call Registry* 4–6, 21 (2009), *available at* http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.216.44 (providing evidence that consumers in more heterogeneous communities more frequently sign up for the Do Not Call list and attributing this to the fact that these consumers receive more telemarketing calls, including more calls marketing goods and services in which they have no interest, because telemarketers cannot target marketing to these consumers as narrowly as consumers living in more homogeneous communities); Goh Khim-Yong et al., *Social Interaction, Observational Learning, and Privacy: The "Do Not Call" Registry* 23–24, 28 (MPRA Working Paper 8225, 2008), *available at* http://mpra.ub.uni-muenchen.de/8225/1/MPRA_paper_8225.pdf (finding that

D. CRACKS IN THE THEORY BEHIND THE USE OF DEFAULTS

The forgoing examples demonstrate that defaults are not always sticky or information-forcing. What drives the rift between defaults in theory and defaults in practice? This Section explains the conditions that drive this rift and draws upon the forgoing examples to reexamine the theories behind the use of defaults in policymaking.

1. Conditions Where Defaults Do Not Work

The key difference between the automatic enrollment retirement saving and Do Not Call default schemes, on the one hand, and the financial information and overdraft defaults, on the other, is the presence of parties that have (1) a strong interest in whether the consumer sticks with or opts out of the default and (2) the ability to shape the presentation of the default and the process for opting out. No party with access to affected employees at the point of the auto-enrollment opt-out decision has a strong interest in pushing consumers in or out of the default. The employers that administer the default do not oppose it, but also do not have a strong reason to try to make it too sticky.[216] The Do Not Call Registry has enemies; telemarketers want consumers in the default position.[217] But telemarketers do not shape the presentation of the Do Not Call list or the process for signing up; that is run entirely by the Federal Trade Commission.[218] Nor do telemarketers have an effective way to reach consumers whose numbers are on the list to lobby or confuse them into selectively opting back in, since consumers must give written, signed consent before a telemarketer can call them.[219]

In contrast, the financial information and overdraft defaults are implemented by the firms that want them to fail. Not all financial institutions

as telemarketing call volume increased in an area, more households signed up for the do-not-call list).

216. Higher participation levels benefit employers because participation increases employee productivity and retention, particularly for those employees whom employers value more. *See* William E. Even & David A. Macpherson, *Benefits and Productivity*, *in* BENEFITS FOR THE WORKPLACE OF THE FUTURE 43, 48–49 (Olivia S. Mitchell et al. eds., 2003). But where the employer provides a matching contribution, higher enrollment can also be costly for the employer.

217. Telemarketers fought the Do Not Call list for years. *See* Mainstream Mktg. Servs., Inc. v. Fed. Trade Comm'n, 358 F.3d 1228, 1233 (10th Cir. 2004) (rejecting telemarketers' First Amendment challenge to the Do Not Call Registry), *cert. denied*, 125 S. Ct. 47 (2004).

218. *See* National Do Not Call Registry, FED. TRADE COMM'N, *available at* https://www.donotcall.gov/faq/faqdefault.aspx (last visited Apr. 13, 2014).

219. *See* Q&A for Telemarketers & Sellers About DNC Provisions in TSR, BUREAU OF CONSUMER PROT., *available at* http://business.ftc.gov/documents/alt129-qa-telemarketers-sellers-about-dnc-provisions-tsr (last visited Apr. 13, 2014).

share customer information with third parties or their affiliates, and not all of those that share with their affiliates allow those affiliates to use the information for marketing purposes. But the institutions that share customer information profit from it and so want their customers to retain their default settings.[220] Not all banks charge overdraft fees, but many that do profit enormously from them and thus have every reason to convince accountholders, and frequent overdrafters in particular, to opt out of the overdraft default.[221] Financial institutions and banks also have access to consumers and use that access to shape the opt-out process and to frame the default at the point of consumer decision. The party opposed to the default is thus able to powerfully influence the consumer's ultimate position.

2. Reexamining the Logic Supporting the Use of Defaults in Policymaking

The failure of the financial information and overdraft defaults to facilitate well-informed consumer choices suggests that the logic underlying the use of policy defaults, penalty defaults, and even altering and framing rules, is flawed. The irony is that while the use of defaults in policymaking is premised on the understanding that irrational biases will make defaults sticky, particularly for poorly-informed consumers, these same biases and this same lack of information may lead defaults to be too sticky, too slippery, or uninformative in many situations.

The theory fails to recognize the degree to which the mechanisms that can make defaults sticky can be manipulated by firms to make those defaults too sticky or even slippery. The very lack of well-formed preferences and a good understanding of available options that facilitates the mechanisms that can make defaults sticky also leaves consumers vulnerable to firm manipulation of those mechanisms. The theory also fails to recognize the depth of the information asymmetry between firms and consumers. A firm facing a penalty default may find it easier to alter the decision environment so as to push consumer biases to favor opting out than to inform and negotiate

220. *See, e.g.*, Rupert Jones, *Barclays to Sell Consumer Data*, GUARDIAN (June 24, 2013), http://www.theguardian.com/business/2013/jun/24/barclays-bank-sell-customer-data ("Bank tells 13 million customers it is to start selling information on their spending habits to other companies."); Catherine New, *Beyond Card Fees: Banks Look to Sell Your Data*, DAILYFINANCE (Oct. 25, 2011), http://www.dailyfinance.com/2011/10/25/beyond-card-fees-banks-look-to-sell-your-data ("[Visa and Mastercard] have plans to sell marketers an analysis of anonymous data that . . . could be used to create targeted online advertising.").

221. *See* PATRICIA CASHMAN ET AL., FED. DEPOSIT INS. CORP., FDIC STUDY OF BANK OVERDRAFT PROGRAMS 56 (2008), *available at* http://www.fdic.gov/bank/analytical/overdraft/FDIC138_Report_Final_v508.pdf (finding that in 2007, overdraft fees amounted to about seventy-five percent of bank deposit account service charges revenue and twenty-five percent of total bank noninterest income).

with the consumer. Most altering and framing rules aim to fine-tune transaction barriers or deliver information. These rules make no attempt to alter the biases and preference formation effects that can make defaults sticky or that can be used by firms to nudge consumers into opting out. Some altering and framing rules may not even alter transaction barriers or be information-forcing. Not all affirmative actions will inhibit opting out—an affirmative mouse click can become reflexive—and not all legally-required notices are read and understood.[222]

Could the financial information and overdraft defaults and their accompanying altering and framing rules be improved? Undoubtedly they could be. But as detailed below, there are limits to altering and framing rules. In the face of strong firm preferences about which position consumers ought to be in, it is possible that even the most carefully crafted default and altering and framing rules would not produce abundant well-informed individual consumer choice.

V. WHY TRACKING DEFAULTS ARE UNLIKELY TO ACHIEVE POLICYMAKERS' GOALS

Currently, transaction barriers, biases, and preference formation effects make the Track-Me position too sticky. Would that change if Track-Me were to become a true policy default, with surrounding altering and framing rules to help those who prefer to opt out to do so? To a degree yes, because some existing transaction barriers would be removed. But biases and preference formation effects would still give a Track-Me default considerable traction, and firms would find ways to erect some transaction barriers without running afoul of altering and framing rules. Moreover, if firms were required to respect a consumer's choice not to be tracked, they would have a stronger incentive to convince consumers not to opt out than they do today. The greater effort firms would expend on keeping consumers in the default could lead fewer consumers to opt out.

What if Don't-Track-Me were the default? Would that lead to consumers sorting themselves into positions that reflect their well-informed preferences? While a Don't-Track-Me setting would require firms to spend significant resources on maneuvering consumers out of the default, firms determined to do so could be successful without necessarily educating many consumers along the way.

This Part suggests strategies that firms could use to keep or put consumers in a Track-Me position, based on strategies firms have used to

222. *See* Ben-Shahar & Schneider, *supra* note 166, at 704–29.

respond to existing defaults. These suggestions are not definitive or exhaustive predictions; firm strategies would necessarily depend on subtle contextual details that cannot be known in advance. But the examples here give the flavor of strategies firms would likely use. This Part then turns to an explanation of why altering rules, framing rules, and competition among firms cannot ensure that tracking defaults will result in informed consumer decisions.

A. HOW FIRMS COULD MAKE TRACKING DEFAULTS FAIL

Even with altering and framing rules intended to encourage consumers to make well-informed decisions that reflect their preferences, firms would have ample opportunities to make these defaults fail. This Section outlines a firm's three main avenues for ensuring the failure of a default: (1) utilizing transaction barriers, (2) harnessing judgment and decision biases, and (3) influencing preference formation effects.

1. Erect, Eliminate, or Invert Transaction Barriers

a) Costs

First, firms would make it easy for consumers to reach the Track-Me position, regardless of the default. For example, when faced with a consumer who has not opted out of a Don't-Track-Me default (or who has opted out of a Track-Me default), a website, program, or device could minimize the costs of opting out by offering the consumer a one-click-opt-out method.

Second, firms would make it costly for consumers to opt into or stay in a Don't-Track-Me position. Although altering and framing rules should lower the costs to consumers who select the Don't-Track-Me position as compared to those costs today, the opt-out process would still be time-consuming because, given current technology, it would need to be performed for every browser on every device and again when new browsers or devices are used. If legally permitted, firms might give coupons and discounts to those who agree to tracking, making it immediately and visibly costly to be in a Don't-Track-Me position.[223] Firms might condition access to content, apps, and other services on consumers being in the Track-Me position, as is the case for many apps and some email services today.[224] When the Netherlands made

223. *See* A. Michael Froomkin, *The Death of Privacy?*, 52 STAN. L. REV. 1461, 1534–35 (2000) (explaining that consumers sell their personal information too cheaply).

224. *See, e.g.*, Adrienne Porter Felt et al., *Android Permissions: User Attention, Comprehension, and Behavior*, *in* SYMPOSIUM ON USABLE PRIVACY AND SECURITY (2012) (discussing conditioning app downloads on permission to track); RT.COM, *supra* note 25 (discussing conditioning Gmail use on permission to scan email content).

Don't-Track-Me the default for websites, most Dutch websites placed "cookie walls" between consumers and website content, pop-up dialog boxes requiring users to accept all cookies to access the websites.[225] As explained above, short of conditioning access on tracking, firms could make refusing to be tracked costly by designing their apps or websites to function poorly when a consumer is in a Don't-Track-Me position.

If the law were to prohibit explicit differential treatment of consumers who did not agree to be tracked, firms might impose more subtle costs instead. For example, firms might inundate consumers with marketing that would only stop if the consumer agreed to be tracked. They might place a complex login process between the website's content and users who are not in a Track-Me position, just as banks require accountholders who have not opted out of the overdraft default to click through a pop-up dialog box to access online banking. Consumers would realize that they can save time by agreeing to be tracked and would soon click "I agree to be tracked" buttons reflexively.[226] The Dutch Parliament found that the cookie walls referenced above led to "mindless clicking of 'I accept' buttons."[227]

Even more subtle costs and perks are possible. For example, rather than contextual advertising (showing consumers ads based on the content the consumer is accessing), firms could show a steady stream of particularly annoying ads (e.g., ads that obscure content, load and play slowly, or contain distracting movement and noise) to anyone who was not in the Track-Me position. In the Netherlands, a few websites now permit access to some content without consumer consent to tracking, but a large pop-up box asking for permission to track remains in the foreground, making it difficult to view the entire site.[228]

---

225. *See, e.g.*, Natali Helberger, *Freedom of Expression and the Dutch Cookie-Wall* 2 (Univ. of Amsterdam – Inst. for Info. Law, 2013), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2197251; *Possible Changes to the Privacy Law in the Netherlands,* BIZCOMMUNITY.COM (Feb. 28, 2013), http://www.bizcommunity.com/Article/152/16/89967.html.

226. *Cf.* FTC PRIVACY REPORT, *supra* note 106, at 49 (noting "choice 'fatigue' " problem when web users are repeatedly asked to opt out); Ayres, *supra* note 16, at 2069 (noting the problem that some altering rules are ineffective because people become "habituated to the speed bumps").

227. Helberger, *supra* note 225, at 4. Analysts found that over 90%, and in most cases closer to 99%, of consumers agreed to be tracked when faced with cookie walls on Dutch websites. Bart Schutz, *Can Defaults Save Online Privacy (or "How to Cure a Self-Destructive Law")?*, ONLINEPERSUASION.COM (May 23, 2013), http://www.online-persuasion.com/can-defaults-save-online-privacy-or-how-to-cure-a-self-destructive-law.

228. *See, e.g.*, DE BRAUW, BLACKSTONE, WESTBROEK, http://www.werkenbijdebrauw.nl/ (Dutch website for Dutch law firm) (last visited Nov. 29, 2013); DE

b) Confusion

Firms would have no trouble using confusion to their advantage. Altering and framing rules should combat some of the confusion that exists today, but the default scheme would almost certainly be confusing to many. It likely would consist of a complex mix of Track-Me defaults, Don't-Track-Me defaults, and unalterable Track-Me positions, comparable to financial information and overdraft defaults, which cover some types of information and overdrafts, but not others. Even if altering rules required firms to provide consumers with a simple, easy-to-use opt-out process at the browser or device level, firms would ask consumers to make granular choices about tracking, following the example set by Facebook's notoriously confusing fifty privacy buttons and 170 privacy-setting options.[229]

Firms might also ask consumers to "click here to opt in to our Privacy Policy" to take advantage of consumers who mistake a "privacy policy" for a promise not to share consumer information. If regulators were to prohibit calling a policy that permits tracking a "privacy" policy, firms would find other confusing labels, such as a "Know Your Customer Policy" or a "Personalized Settings Policy." The former sounds like a duty placed on firms and the latter sounds like a benefit designed for consumers, both of which avoid "tracking," a "dirty word" among firms that track consumers today.[230]

c) Futility

Under a true default regime, choosing not to be tracked would be significantly less futile than it is today, because firms would have a stronger legal obligation to respect consumer choices. Yet, consumers likely would not be permitted to avoid all tracking for all purposes; firms would continue to track for "permissible purposes" even those consumers in the Don't-Track-Me position. As with the financial information defaults, consumers might feel that resisting the position preferred by firms is futile.

BRAUW, BLACKSTONE, WESTBROEK, http://www.debrauw.com/ (English website for Dutch law firm) (last visited Nov. 29, 2013).

229. *See* Nick Bilton, *Price of Facebook Privacy? Start Clicking*, N.Y. TIMES, May 13, 2010, *available at* http://www.nytimes.com/2010/05/13/technology/personaltech/13basics.html.

230. *See* Miller & Sengupta, *supra* note 33 (quoting COO of a tracking technology company as saying " 'Tracking is a dirty word' ").

2. Harness Judgment and Decision Biases

a) Salience Effects

The salience of tracking strongly influences most people's privacy-related actions. One experiment found that consumers who are reminded of the privacy implications of disclosure disclose little, but those who are not reminded appear to forget about privacy—many will reveal socially stigmatized and even illegal behavior.[231]

For a Track-Me default, framing rules intended to make the option to opt out "clear and conspicuous" should increase the salience of the default and opt-out choice somewhat, especially if media coverage of a new law about tracking raises consumer awareness. But at the moment consumers could potentially opt out, firms could minimize the salience of this choice. Websites are frequently cluttered with material the user is not interested in, and part of the skill of using the web is learning how to mentally screen out this content. Website design is so flexible that even with reasonably detailed framing rules about font size, positioning, and the like, firms could probably leverage users' skill in screening out information to hide an opt-out option in plain sight.[232] COPPA, for example, requires that privacy policies for children's websites be positioned prominently, yet they are frequently surrounded by other materials that draw users' attention away.[233] One set of lab experiments found that while reminding people about privacy increases privacy-protective behavior, a mere fifteen-second delay between the reminder and the privacy-related decision was enough to negate the effects of the reminder.[234] More blatantly, firms could place content relevant to opting out such that it falls outside the viewing area of small-device screens.[235]

231. Leslie John et al., *Strangers on a Plane*, 37 J. CONSUMER RES. 858, 859–60 (2011).

232. For examples, see *Dark Patterns Library*, Darkpatterns.org, http://darkpatterns.org/ (last visited Nov. 29, 2013) ("A Dark Pattern is a type of user interface that appears to have been carefully crafted to trick users into doing things . . . . Dark Patterns . . . are carefully crafted with a solid understanding of human psychology, and they do not have the user's interests in mind.").

233. *See, e.g.*, JOSEPH TUROW, PRIVACY POLICIES ON CHILDREN'S WEBSITES: DO THEY PLAY BY THE RULES? (Annenberg Public Policy Center Report 2001) (showing how children's websites follow the regulations about the placement of privacy policy links yet are able to surround the links with distracters or otherwise reduce the links' visibility).

234. Idris Adjerid et al., *Sleights of Privacy: Framing, Disclosures, and the Limits of Transparency* 8 (unpublished manuscript), *available at* http://www.heinz.cmu.edu/~acquisti/papers/acquisti-sleights-privacy.pdf.

235. *See, e.g.*, FTC MOBILE REPORT, *supra* note 106, at 3 ("[W]ith many devices possessing screens of just a few inches, there are practical challenges in terms of how critical information—such as data collection, sharing of information, and use of geolocation data—is conveyed to consumers.").

Even without a planted distraction, at the concrete moment when consumers use an app, website, or device, they are focused on something else, just as a consumer engaging in a financial transaction is attending to the transaction and not the information-sharing implications. Further, any salience brought to a Track-Me default by framing rules would dissipate over time due to "warning fatigue." Even today, few consumers who are given disclosures explaining that apps collect personal information read—and perhaps as few as three percent understand—the app permissions lists.[236]

For a Don't-Track-Me default, firms might interrupt consumer use of apps, websites, or devices with pop-up screens or similar barriers so that consumer focus is diverted to the tracking decision. Alternatively, the same repeated presentation of an "I agree to be tracked" button that would lower the transaction costs of opting out by inducing reflexive clicks could also reduce the salience of the option *not* to opt out, in that consumers clicking mindlessly will not give this option consideration.

b) Omission Bias

Omission bias that favors tracking today would continue to do so under a full-fledged Track-Me default scheme. Given that most consumers are currently in a Track-Me position, firms facing a Track-Me default could encourage the operation of omission bias by emphasizing to consumers that nothing has changed and they need not take any action.

In contrast, firms facing a Don't-Track-Me default would work to overcome the omission bias, perhaps by placing consumers in a forced-choice scenario. Without the option to do nothing, omission bias would not favor the default. Even if framing rules prevented firms from forcing consumers to choose between the default and opting out, firms might borrow from bank marketing materials that state or imply that consumers "must" take action, so that inaction no longer appears to be a blameless omission.

c) Loss Aversion and the Endowment Effect

With respect to Track-Me defaults, firms could work to ensure that loss aversion and the endowment effect continue to favor the Track-Me position. Firms could encourage consumers to treat the default as the reference point against which gains and losses are measured and as the position with which consumers are currently endowed. For example, firms might characterize Track-Me as the position in which websites, devices, or apps work

236. *See* Felt et al., *supra* note 224, § 5.1.

"properly" or "as you have come to expect" and warn consumers that opting out could impair the user experience.

If a Don't-Track-Me default were imposed, firms could frame opting out of the default not as losing an endowed position but as gaining a "personalized" service. Given that a Don't-Track-Me default would be a change from today's Track-Me world, marketing materials might explicitly invoke loss aversion, asking "Would you like to keep" this service "personalized for you?", or "Would you like to change your settings?" Just as banks asked consumers to opt out of the overdraft default before it became operative, firms might ask consumers to agree to tracking while still in the Track-Me position,[237] thus encouraging consumers to view opting out as keeping the status quo. Firms currently tracking consumers might even tailor their campaigns to specific consumers by determining which feature a consumer has used in the past and warning the consumer that she "could" lose that feature if she does not opt out.

d) Procrastination and Decision Avoidance

Firms facing a Track-Me default could nurture procrastination and decision avoidance by reminding consumers that they can "opt out at any time" and suggesting that opting out requires multiple, time-consuming steps. Although altering rules should make the opt-out process easier than it is today, firms would ensure that the process still appears difficult. Even if framing rules required that firms provide consumers with brief, easily-understandable descriptions of the default and opt-out positions, firms might surround these with voluminous impenetrable "explanations."

In contrast, firms facing a Don't-Track-Me default would probably place consumers in a forced-choice scenario so that procrastination and decision avoidance are not options. Firms might also give consumers false deadlines for opting out just as banks did for the overdraft default, to reduce procrastination.

e) Excessive Discounting

As is true today, the time and effort costs of opting out of a Track-Me default, even if small, would be immediate and certain, whereas any benefits would be uncertain and in the future, particularly given that future uses of

237. The W3C preliminary draft explicitly permits this. W3C TPE, *supra* note 104, § 6.3.1 ("Sites MAY ask for an exception, and have it stored, even when the user's general preference is not enabled.").

information are unknowable.[238] To the extent altering rules would permit, firms might impose additional immediate and tangible costs on consumers who opt out of a Track-Me default. Excessive discounting would thus favor the default.

If Don't-Track-Me were the default, firms could make the potential costs of keeping the default seem probable and clear, and promise immediate peace of mind as a benefit of opting out. Imagine a marketing vignette in which a man tries to impress a woman by showing her something on his computer screen, but the woman's attention is drawn to advertising that pops up and implies something embarrassing about him. She reacts negatively and he tries in vain to claim innocence. Advertising copy might then ask "Tired of ads that weren't meant for you? Opt into personalized advertising today." Firms might also downplay the privacy costs of opting out by offering "free" benefits in exchange for opting out. Just as many apps today are "free" but require the consumer to provide their data in exchange, so too a consumer might be asked to opt out of a Don't-Track-Me default in exchange for "free" apps or services.

f) Choice Bracketing

Altering rules accompanying a Track-Me likely would permit consumers to opt out at the browser or device level—a broadly bracketed decision that could lead individuals to make a decision based on the cumulative effect of loss of information privacy, and thus opt out. But firms might ask consumers to opt back into the default "just for this one" firm, website, or app, triggering a narrowly bracketed decision that would favor immediate benefits over privacy concerns.

For a Don't-Track-Me default, firms would not ask consumers to opt out at the browser or device level—that would benefit the firm's competitors as well as the firm—but instead would seek consent to tracking by that particular firm. Each firm would highlight the narrow nature of that opt-out decision.

g) The Illusion of Control

Perceptions of control strongly affect privacy decision making. Consumers who feel more in control of the exchange of their information

238. *See, e.g.*, Acquisti & Grossklags, *supra* note 79, at 6 ("[A]n individual who is facing privacy sensitive scenarios may be uncertain about the values of possible outcomes and their probability of occurrence, and . . . sometimes she may not even be able to form any beliefs about those values and those probabilities."); NEHF, *supra* note 9, at 126–29 (explaining that consumers cannot know how their information will be used or how those uses will affect them in the future).

with firms are more willing to allow those firms to collect more of their personal information.[239] Giving consumers the illusion of more control leads them to both reveal more sensitive information and allow more publication of that information.[240]

Regardless of the tracking default setting, firms could stress to consumers that they are "in control" of their privacy and thus encourage consumers to share more information. For example, Google today suggests that consumers connect their accounts (enabling tracking across accounts) because "[c]onnecting your accounts puts you in control"; it then reminds consumers, "[r]emember, Google won't share your searches or other private information with third-party services without your consent."[241] Google uses information it gathers about consumers without explicit consent, but the "you are in control" pitch deflects attention from this.[242] A legally enforceable default might give consumers a feeling of greater control than they have today, which could lead to less privacy-protective behavior.

h) The Sunk Costs Fallacy

Under either a Track-Me or a Don't-Track-Me default, firms could exploit the sunk costs fallacy to increase the magnetism of the Track-Me position. Rather than placing tracking walls at the start of a consumer's interaction with a device or program, firms might allow consumers to use the device or program, in a limited or temporary fashion, regardless of the consumer's position. Then, once the consumer has sunk costs into learning to use the device or program, the firm could present a tracking wall to prevent further use. This is similar to what happens today with cellphone apps; consumers must select an app and go through most of the download process before they can learn how much data the app will gather from them

239. Nadia Olivero & Peter Lunt, *Privacy Versus Willingness to Disclose in E-Commerce Exchanges: The Effect of Risk Awareness on the Relative Role of Trust and Control*, 25 J. ECON. PSYCHOL. 243, 259 (2004); *see also* Stutzman et al., *supra* note 122, at 29 (finding that as Facebook gave users more control over settings that determine which other Facebook users can view their pages, users made more content "private" vis-a-vis other users, but also posted more confidential information, such that Facebook and parties to which it provides data obtained more information about users).

240. Laura Brandimarte et al., *Misplaced Confidences: Privacy and the Control Paradox*, SOC. PSYCHOL. & PERSONALITY SCI. 3 (2012).

241. *See* Tate, *supra* note 27 (reporting on "an obscure checkbox on a buried Google account preferences pane, which reads, 'use my Google contact information to suggest accounts from other sites' " and explaining "[b]y default, this box is checked, which means Google has been scanning your Gmail contacts, unless by some miracle you found this option, buried several clicks beyond your Gmail inbox, and disabled it").

242. *Id.*

if they complete the installation process.[243] At that point, the fallacy might incline consumers to agree to tracking, whether that means opting out of a Don't-Track-Me position or back into a Track-Me position.

3. Bolster, Undermine, or Reverse Preference Formation Effects

a) The Endorsement Effect

If a Track-Me default were adopted, firms could channel the implicit advice mechanism to their advantage and reinforce it with explicit advice, along the following lines:

> Most people don't like receiving a lot of advertising for products they don't want and will never buy. That's why Congress decided to make "Know Your Audience" the default for advertising. If you would like to change this setting, you can. But most people prefer to keep the default.

Such scripts would emphasize that the default reflects what most people want and what policymakers have decided that consumers ought to prefer. Firms might also tackle the privacy issue directly and surround any legally mandated disclosure of the right to opt out with assurances that consumers do not need to opt out because the firm "cares" about the consumer's privacy.

In response to a Don't-Track-Me default, firms would likely defuse the implicit advice effect by obscuring which position is the default. Firms might ask consumers to select between "opting in" to the firm's "privacy policy" (where that policy effectively opts the consumer out of the Don't-Track-Me position) or "opting out" (where "opting out" means sticking with the Don't-Track-Me default). Firms could also counter any implicit advice conveyed by a Don't-Track-Me default with explicit advice to opt out.

b) The Experience Effect

For Track-Me defaults, firms would likely emphasize that nothing has changed and the consumer can continue to use the program or device as she has always done. In contrast, firms would likely push consumers to opt out of any Don't-Track-Me position before the position became effective, to avoid the risk that customers might experience the default and develop a preference for it.

* * *

243. *See* Felt et al., *supra* note 224.

Some of the strategies discussed above are at cross-purposes. For example, on the one hand, reminding people that they *control* their own position might encourage risk taking and thus sticking with a Track-Me default. On the other hand, "you are in control" marketing might make salient that consumers are *responsible* for their own positions and thus discourage omission bias that would otherwise favor sticking with a Track-Me default. If a full-fledged tracking default scheme were adopted, firms would gradually refine their marketing by testing a variety of campaigns, using data gleaned from tracking to target their approaches.[244]

B. ALTERING RULES, FRAMING RULES, AND COMPETITION

The analogies between Track-Me and Don't-Track-Me defaults and the existing financial information and overdraft defaults may appear limited in two respects. First, the existing defaults arguably were political compromises, intended to satisfy public clamor for policymaker action without affecting financial institutions' or banks' bottom lines. Perhaps with more political will, altering and framing rules that would facilitate well-informed, unbiased consumer decision making would accompany these defaults. Second, the privacy landscape contains firms—specifically, some web browser and plug-in makers—that are not dependent on revenue generated from tracking. These firms potentially could implement technologies that protect consumer privacy to woo privacy-concerned consumers from doing business with competitors that allow or engage in tracking.

Yet, a legislatively enacted tracking-default scheme is no less likely to be the product of political compromise than the financial-information or overdraft defaults. More fundamentally, as explained further in this Section, no altering and framing rules consonant with the use of defaults in policymaking will be strong enough to neutralize the influence of self-interested firms on consumer behavior. It is likely inherent in the structure of a consumer-choice regime that where consumers have a weak understanding of their preferences and options, firms that have access to consumers and a strong interest in consumers' ultimate positions will shape those positions.

Similarly, financial institutions and banks might compete on the basis that they do not collect consumer financial information and do not provide expensive overdraft coverage, but few do.[245] Competition over privacy

244. *See* Ryan Calo, *Taking Data Seriously: Market Manipulation in the Digital Age*, YALE L. SCH. INFO. SOC'Y PROJECT (Mar. 28, 2013), http://www.yaleisp.org/event/thomson-reuters-speaker-series-ryan-calo.

245. The main exception is the recently launched Bluebird debit card, which is promoted as a way to avoid both overdraft and monthly checking account fees. The card is

appears to be more common, but the effectiveness of this competition is undermined by the limited ability of consumers to understand and assess firms' privacy promises. More importantly, as this Section elaborates, a tracking default scheme might make things worse, because consumers might perceive the law to provide sufficient protection without the help of privacy intermediaries.[246]

1. The Limits of Altering and Framing Rules

Altering and framing rules might seem to easily counter firm ploys to increase or decrease the stickiness or slipperiness of defaults. To make defaults stickier, policymakers might impose costly opt-out procedures. To prevent firms from making defaults too sticky, policymakers might prohibit conditioning transactions on, or giving perks for, sticking with a default. Policymakers might require specified disclosures crafted to frame the default and opt-out positions in ways that harness or defuse biases. Or, policymakers might require disclosures that convey explicit advice about whether consumers ought to stick with the default or opt out. But normative, legal, and practical constraints limit these rules.

a) Respect for Heterogeneous Preferences

There are two assumptions that premise a default regime rather than a mandate for personal-information tracking: first, that consumers have heterogeneous preferences regarding privacy and the benefits tracking can provide, and second, that the law should respect those heterogeneous preferences by allowing consumers to decide for themselves whether to be tracked.

Altering rules that substantially inhibit opting out are inconsistent with respect for heterogeneous individual tracking preferences. For example, reverting defaults that require consumers to opt out every time they open up their browsers or fire up their mobile devices[247] might effectively keep

exempt from the interchange fee restrictions placed on traditional debit cards and its issuer, American Express, makes money from high interchange fees rather than from traditional overdraft and monthly account fees. *See* Andrew Kahr, *Amex's Bluebird Is a No-Fee Checking Account, Not a 'Prepaid Card,'* AMERICAN BANKER (Oct. 16, 2012), http://www.americanbanker.com/bankthink/american-express-bluebird-is-a-no-fee-checking-account-not-a-prepaid-card-1053516-1.html.

246. So too there might be more competition over overdraft fees absent the legal default rule, because some consumers may perceive that the default gives them sufficient control over overdraft fees and they do not need to shop for a bank that does not charge these fees.

247. *Cf.* Paul Schwartz, *Property, Privacy and Personal Data*, 117 HARV. L. REV. 2055, 2098–107 (2004) (proposing a default rule that a consumer's personal data cannot be transferred to third parties, complemented by altering rules requiring that the consumer consent to each

consumers in the default position (provided that firms did not manage to make the opt-out process so easy and routine that consumers would opt out mindlessly[248]). But this would be difficult to justify normatively. Although prohibiting differential treatment of consumers depending on whether they have agreed to be tracked[249] could prevent the most blatant ways that firms might maneuver consumers into a Track-Me position, it might also be normatively problematic. Where tracking currently funds the website, app, or device at issue, such a rule could have substantive effects on the availability of these, particularly for consumers who have limited financial means. Such substantive effects may be an appropriate trade-off for increased information privacy, but such effects go beyond providing consumers with choices about tracking.

Framing rules that would be dramatic enough to be effective are also normatively problematic. In a world cluttered with information and decisions, commanding attention requires something more drastic than a neutral description of the default and opt-out positions, such as those currently required for the financial-information defaults. But imagine a mandated disclosure likening tracking to stalking or spying, conveyed through pictures similar to the graphic cigarette warnings recently proposed by the Food and Drug Administration ("FDA").[250] While potentially effective in convincing consumers to stick with a Don't-Track-Me default, this disclosure would be incompatible with the policy goal of allowing people to sort themselves into their desired positions freely, without a strong policymaker push in any particular direction.

Framing rules that require more complex disclosures or other forms of consumer education might be another strategy, one that could in theory reduce consumer preference uncertainty that fuels susceptibility to firm framing manipulations. But in a quickly changing, complex decision

subsequent transfer of her data at the time the data is transferred, with the aim that the default would be sticky).

248. *Cf.* Wang et al., *supra* note 22 ("Some of our participants reported that they began to ignore our nudges after several days. Future work might investigate addressing this habituation effect.").

249. *See* Paul M. Schwartz & Daniel Solove, *Notice & Choice: Implications for Digital Marketing to Youth*, 2ND NPLAN/BMSG MEETING ON DIGITAL MEDIA AND MARKETING TO CHILDREN 4 (2009), *available at* http://digitalads.org/documents/Schwartz_Solove_Notice_Choice_NPLAN_BMSG_memo.pdf (proposing altering rules prohibiting "mak[ing] the provision of access, information, services, or transactions contingent upon a person's" consent to tracking).

250. *See, e.g.*, R.J. Reynolds Tobacco Co. v. Food & Drug Admin., 696 F.3d 1205, 1221–22 (D.C. Cir. 2012) (describing the Food and Drug Administration's proposed cigarette warnings).

environment, the utility of such education is likely to be nil.[251] Only simple, easily actionable messages tend to be effective in public education campaigns.[252] But a policymaker who believes that consumers ought to take a variety of positions with respect to tracking cannot send a simple "just say no" or "just say yes" message.

b) Mis-sorting

A further obstacle to developing more effective altering and framing rules is a practical one—most such rules are unlikely to sort consumers well. Consider an altering rule designed to counter firm manipulation by making it more difficult for a consumer to opt out of a Don't-Track-Me position, such as requiring a consumer to send a signed letter through the U.S. Postal Service.[253] Consumers who are in the habit of using traditional mail would find this altering rule no more than a speed bump, whereas consumers who conduct their lives online are more likely to be impeded by such a rule. Disparate opt-out rates of these two groups then would not reflect underlying preferences about tracking.

Alternately, consider a requirement that consumers pass an online test proving they have an understanding of the default and opt-out positions as a condition of opting out. Although this process might be informative to those who completed it, many consumers dislike being tested and would be deterred from attempting to opt out on this basis. Others have weak reading or comprehension abilities that would impair test performance. Yet consumers who dislike or perform poorly on written tests are not less likely

251. *Cf.* Lauren E. Willis, *Against Financial Literacy Education*, 94 IOWA L. REV. 197 (2008).

252. *See* JESSICA ASCHEMANN-WITZEL ET AL., *Lessons for Public Health Campaigns from Analyzing Commercial Food Marketing Success Factors: A Case Study*, *in* BIOMED CENTRAL PUBLIC HEALTH 9 (2012).

253. *Cf. Long-Term Care HIPAA Notice*, METLIFE, https://www.metlife.com/about/privacy-policy/HIPAA-privacy-policy/long-term-care-notice.html (last visited Nov. 29, 2013):

> You have the right to request a restriction or limitation on Personal Health Information we use or disclose about you for treatment, payment or health care operations, or that we disclose to someone who may be involved in your care or payment for your care, like a family member or friend. . . . To request a restriction, you must make your request in writing to Privacy Coordinator, MetLife, P.O. Box 937, Westport, CT 06881-0937. In your request, you must tell us (1) what information you want to limit; (2) whether you want to limit our use, disclosure or both; and (3) to whom you want the limits to apply (for example, disclosures to your spouse or parent).

to benefit from opting out than consumers who enjoy and perform well on written tests.

More broadly, any steps that consumers must take to opt out of a default will have differential effects depending on, for example, consumer habits, confidence, technological savvy, and susceptibility to biases. Some consumers might find an opt-out process confusing, while others have no trouble completing it. Similarly, consumers approach choices with different discount functions, and not all consumers are affected by omission bias.[254] Likewise, while most of us are procrastinators, consumers do not all procrastinate on the same things or to the same extent.[255] Heterogeneity in consumer responses to transaction barriers or in consumer biases means altering rules will affect consumers differently, yet this heterogeneity is unlikely to correspond well with consumers' well-informed privacy preferences. Today, for example, tech-savvy consumers are more likely than the average consumer to employ various techniques that prevent tracking, although tech-savvy consumers are no more concerned about privacy generally.[256]

c) Commercial Speech Doctrine Limits

Given the likelihood that firms will engage in marketing that drives judgment and decision biases to favor the Track-Me position, framing rules prohibiting this type of marketing might seem an appropriate response. Policymakers could even require firms to frame the default and opt-out positions in particular ways.

Under current First Amendment doctrine, commercial speech has a substantial degree of protection.[257] Framing rules generally cannot prohibit

254. *See, e.g.*, Christopher F. Chabris et al., *Individual Laboratory-Measured Discount Rates Predict Field Behavior*, 37 J. RISK & UNCERTAINTY 237, 263 (2008) (finding discount functions vary between subjects and correlate with behavior); Jonathan Baron & Ilana Ritov, *Omission Bias, Individual Differences, and Normality*, 94 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 74, 74 (2004) (finding that not everyone is affected by the omission bias, and some even display an action bias).

255. *See, e.g.*, Ernesto Reuben et al., Procrastination and Impatience 16–18 (Jan. 28, 2008) (unpublished manuscript), *available at* http://ssrn.com/abstract=1077708 (finding propensity to procrastinate varies, as does the degree to which people anticipate their own procrastination).

256. *See, e.g.*, RAINIE ET AL., *supra* note 4, at 24 (finding lower-income Internet users report many more problems with the misuse of their personal information, including online stalking and harassment, than higher income consumers); Yong Jin Park, *Digital Literacy and Privacy Online*, 40 COMMC'N RES. 215 (2011) (finding consumers in all socioeconomic groups nearly equally concerned about online privacy).

257. Any governmental restriction on non-misleading commercial speech must (1) directly advance a substantial governmental interest and (2) be narrowly tailored to serve that

firms from using particular speech, graphics, or vignettes to convey commercial messages if those messages are not misleading.[258] Although telling consumers they will not be tracked when they are being tracked is misleading,[259] courts are unlikely to find any of the marketing ploys described above misleading. The government is also limited in the disclosures that it can require firms to make. For example, while a disclosure likening tracking to stalking or spying might be effective in convincing consumers to stick with a Don't-Track-Me default, forcing firms to give such a disclosure might violate current constitutional limits on the regulation of commercial speech.[260]

Without the ability to prevent firm speech that frames the default and opt-out positions or to require disclosures that can reach consumers through dramatic messages, firms have the upper hand; they can better reach the consumer at the point of decision and can frame the decision using anything short of demonstrably misleading speech.

d) The First and Last Mover

Finally, practical limits on regulation will inhibit policymakers' efforts to manipulate consumer judgment and decision biases or to prevent firms from doing so. It is true that firm ploys to keep consumers in Track-Me defaults, move them out of Don't-Track-Me defaults, or move them back into Track-Me defaults will sometimes fall flat or even backfire. Consumers are a diverse and fickle lot; what one consumer finds acceptable another finds out-of-bounds, and a single consumer might find a path-breaking firm's actions disquieting at first but unremarkable if the rest of the market moves in the same direction.[261] However, firms can send a diverse set of marketing messages (informed by behavioral tracking data) and only need one of these to work with any particular consumer. Firms can also experiment with risky

interest. *See* Cent. Hudson Gas & Elec. Corp v. Pub. Serv. Comm'n of New York, 447 U.S. 557, 564 (1980).

258. *See* Willis, *supra* note 34, at 1161 (explaining how current commercial speech doctrine restricts rules prohibiting firm framing manipulations).

259. *See* ScanScout Consent Order, *supra* note 125.

260. *See, e.g.*, R.J. Reynolds Tobacco Co. v. Food & Drug Admin., 696 F.3d 1205, 1221–22 (D.C. Cir. 2012).

261. For example, Google's February 2012 privacy policy allowing third-party data sharing was met with controversy, but few noticed when Microsoft followed suit in October after running the "Scroogled" campaign criticizing Google. *Compare* Jeff Blagdon, *Google's New Controversial Privacy Policy Now in Effect*, VERGE (Mar. 1, 2012), http://www.theverge.com/2012/3/1/2835250/google-unified-privacy-policy-change-take-effect, *with* Edward Wyatt & Nick Wingfield, *As Microsoft Shifts Its Privacy Rules, an Uproar Is Absent*, N.Y. TIMES, Oct. 19, 2012, http://www.nytimes.com/2012/10/20/technology/microsoft-expands-gathering-and-use-of-data-from-web-products.html.

approaches on a small scale or change course quickly if one approach backfires.[262]

Policymakers are not nearly so agile. Policymakers can set framing rules requiring that firms provide consumers with particular disclosures, but firms can run circles around disclosures. Witness the failure of the behavioral-advertising privacy icon today. This simple, universal, and widely-used icon leads to an explanation of how to opt out of receiving behavioral advertising, but while most consumers have received advertising with this icon attached, very few consumers know what the icon means, and even fewer have clicked on it.[263] Policymakers are unlikely to develop any framing rules that present the opportunity to opt out of a default in clearer terms than an icon like this.[264]

Firms are ultimately both the first movers and the last movers and can use these positions to whipsaw default schemes. Firms use marketing to shape consumer perceptions before consumers encounter the opt-out decision and firms then shape the frame at the moment those firms are providing consumers with the legally required opt-out choice. A framing rule can control one aspect of a frame, but firms can place that aspect within a larger frame that determines how effective the legal framing will be. Policymakers can set altering rules, but firms implementing those rules have the final say on how the entire opt-out process is designed. An altering rule does not completely set how the default will be altered; it merely sets one aspect of that process, and firms control the rest. Yet, it is the entire process that affects the stickiness or slipperiness of the default. If the law could

262. *Cf. Facebook Backs Down, Reverses on User Information Policy*, CNN (Feb. 18, 2009), http://www.cnn.com/2009/TECH/02/18/facebook.reversal; *see also* Calo, *supra* note 244.

263. *See* Ur et al., *supra* note 49 (about ten percent of study participants had some idea that the AdChoices icon was related to advertising; none actually knew what it meant or how it functioned); Pedro Giovanni Leo et al., *What Do Online Behavioral Advertising Disclosures Communicate to Users?* (Carnegie Mellon Univ., Working Paper CMU-CyLab-12-008 2012), *available at* http://www.cylab.cmu.edu/research/techreports/2012/tr_cylab12008.html (finding that over half of surveyed consumers believed that clicking on the AdChoices icon would cause more advertising to pop-up and more respondents believed clicking the icon would let them purchase advertisements than understood that the icons had something to do with behavioral advertising); Matthew Creamer, *Despite Digital Privacy Uproar, Consumers Are Not Opting Out*, ADAGE (May 30, 2011), http://adage.com/article/digital/digital-privacy-uproar-consumers-opting/227828 (reporting click-through rate for the AdChoices icon is only .002%–.005%).

264. Mozilla has suggested a system by which websites or apps would present four of a set of eight possible icons that would encapsulate the key features of the privacy practices of the site or app. *See Privacy Icons*, MOZILLAWIKI, https://wiki.mozilla.org/Privacy_Icons (last visited Nov. 29, 2013). This system seems hopelessly complex, given the failure of even a single icon to be well-understood.

respond to each firm maneuver—and then hold everything constant—altering and framing rules might be able to fine-tune the power of a default. But the law cannot require stasis.

2. Will Competition Change the Calculus?

Is it possible that competition could do what government regulation cannot, creating a robust market for privacy, complete with consumers who make well-informed decisions about tracking that reflect their preferences? Some firms already engage in privacy pitches, and one can imagine others: "With our cellphone, no one can track you."; "Google tracks you. We Don't.";[265] "We're a social network, not an advertising network." Firms seeking to compete based on promises not to track consumers would not be limited by the same normative, legal, and practical constraints facing lawmakers. They could unabashedly push the merits of privacy and the demerits of tracking, and they could respond quickly to their competitors' marketing strategies with their own attempts to bend consumer judgment and decision biases in their favor. But will they? And if so, does the default matter?

Although some privacy-based competition has appeared in the marketplace, thus far it appears to be of limited effectiveness. A small cadre of privacy sophisticates may choose websites, apps, and devices based on reliable promises not to track consumers, but most consumers have difficulty distinguishing firms and products on privacy grounds.[266] Competition thus centers on privacy image rather than privacy reality.[267] For example, while the TRUSTe privacy certification seal increases consumer trust in websites,[268]

265. Kunur Patel, *How Do You Brand Consumer Privacy?*, ADAGE (Feb. 13, 2012), http://adage.com/article/digital/brand-consumer-privacy/232694 (quoting advertising copy used by Internet search engine DuckDuckGo).

266. *Cf.* Joseph Bonneau & Sören Preibusch, *The Privacy Jungle: On the Market for Data Protection in Social Networks*, *in* THE 8TH WORKSHOP ON THE ECONOMICS OF INFORMATION SECURITY (2009), http://preibusch.de/publications/Bonneau_Preibusch__Privacy_Jungle__2009-05-26.pdf (finding that the "economically rational choice for a [social networking] site operator is to make privacy control available to evade criticism from privacy fundamentalists, while hiding the privacy control interface and privacy policy to maximise sign-up numbers and encourage data sharing from the pragmatic majority of users"); Joseph Turow, *Behavior Aside, Consumers Do Want Control of Their Privacy*, ADAGE (Jan. 29, 2013), http://adage.com/article/guest-columnists/behavior-consumers-control-privacy/239376 (demonstrating how firm marketing may convince consumers that they are protecting consumer privacy even when a closer read of privacy policies reveals this protection to be minimal).

267. *See* Patel, *supra* note 265 (describing widespread advertising by web firms, including Google, touting their privacy-protectiveness).

268. Jensen et al., *supra* note 30.

one study found that websites using this seal engage in more privacy-invasive practices than firms without the seal.[269] Microsoft's recent "Scroogled" marketing campaign, which paints Google as a privacy-invader,[270] may turn out to be mere optics. Microsoft's campaign criticizes Google for tracking consumers, but Microsoft also reserves the right to track consumers in the fine print of its contracts.[271] Consumers who realize that privacy marketing is an unreliable indicator of privacy practices may simply assume all firms track them.[272]

A legal default might change this dynamic somewhat, particularly a Don't-Track-Me default that required firms to obtain express consumer consent before tracking. A request for consent would be a conspicuous admission that the firm tracks consumers, which would help consumers distinguish firms based on the firms' tracking practices, a predicate for genuine privacy-based competition.

But because the returns to firms from tracking are only likely to grow,[273] it is likely to be more profitable for most firms to join the firms that engage in tracking rather than compete against them based on privacy. For example, while Apple's Safari web browser blocks third-party cookies, Apple's iPhone tracks users for advertising and other purposes.[274] To the extent that

269. Benjamin Edelman, *Adverse Selection in Online 'Trust' Certifications*, *in* PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ELECTRONIC COMMERCE (2009), *available at* http://www.benedelman.org/publications/advsel-trust.pdf.

270. Michael Learmonth, *Microsoft Debuts New Commercials on Privacy, with Google in the Crosshairs; Bringing up Privacy to Draw a Distinction with Rival*, ADAGE (Apr. 22, 2013), http://adage.com/article/digital/microsoft-launches-privacy-tv-campaign-google-crosshairs/241001 (describing Microsoft's use of privacy in its marketing to distinguish itself from Google, and noting that "ad-supported web services are a money-losing side show for Microsoft but a profitable core business for Google"); Geoff Duncan, *Why Do Not Track May Not Protect Anybody's Privacy*, DIGITAL TRENDS (June 9, 2012), http://www.digitaltrends.com/mobile/why-do-not-track-may-not-protect-anybodys-privacy (suggesting that Microsoft set "Do-Not-Track" as the default for IE 10 both as a marketing strategy and to undercut Google's advertising revenue).

271. Strange, *supra* note 39.

272. Tony Vila et al., *Why We Can't Be Bothered to Read Privacy Policies*, *in* ECONOMICS OF INFORMATION SECURITY 143 (L. Jean Camp & Stephen Lewis eds., 2004) (explaining how privacy market functions like a lemons market).

273. *See* James Temple, *Rules Against Online Tracking Don't Go Far Enough*, SAN FRANCISCO CHRON., Mar. 7, 2012, http://www.sfgate.com/business/article/Rules-against-online-tracking-don-t-go-far-enough-3387373.php ("Targeting ads based on search queries, sites visited, stories read and social connections forms the core of the multimillion-dollar business models of many online companies, including Google, Yahoo and Facebook.").

274. *See, e.g.*, Rosa Golijan, *Apple's Tracking Your iPhone Again, but Says You Can 'Limit' It*, NBC.COM (Oct. 12, 2012), http://www.nbcnews.com/technology/apples-tracking-your-iphone-again-says-you-can-limit-it-1C6427715.

consumers understand that the iPhone tracks them, marketing for Safari that stresses the importance of privacy could hurt iPhone sales. Similarly, online retailers might want consumers to stay in a Don't-Track-Me position with respect to geolocational cellphone tracking to prevent their brick-and-mortar competitors from using this tracking to attract market share. However, online retailers are likely to want to use Internet-use tracking for their own purposes. Marketing that encourages consumers to adopt a Don't-Track-Me position for geolocational tracking could imperil marketing efforts to keep consumers in a Track-Me default on the Internet.

Further, the web browser and plugin makers that currently seek to satisfy consumer preferences not to be tracked might fare worse in a world with a legally enforceable Don't-Track-Me option. With the power to opt out of a Track-Me default or stay in a Don't-Track-Me default, consumers might view these intermediaries as no longer necessary to effectuate their privacy preferences.

While it is impossible to know whether a robust privacy market will develop, the "cookie wall" experience in the Netherlands is instructive. Dutch law made plain which firms were tracking consumers online because only those firms had to ask consumers to opt out of the Don't-Track-Me default.[275] But virtually all firms (and even nonprofits[276]) responded to the law by requiring consumers to opt out as a condition of using the firm's website, not by competing on promises not to track.[277]

VI. CONSEQUENCES OF PRIVACY POLICY BY DEFAULT

None of this tells us whether or when Track-Me or Don't-Track-Me is the best position for consumers. Tracking's benefits to society certainly exceed the costs in some situations. But a default scheme surrounded by firm manipulation of consumer biases or confusion to keep consumers in or lead consumers to the Track-Me position creates only the façade of choice. That façade cannot justify the conclusion that Track-Me is the right position. To assess whether and when firms ought to track consumers would require a far more difficult inquiry than checking whether a consumer has clicked a button opting out.[278]

275. Helberger, *supra* note 225.

276. *See, e.g.*, *Cookies*, TUDELFT, http://cookie.tudelft.nl/ (last visited Aug. 1, 2013).

277. *See* Helberger, *supra* note 225.

278. *Cf.* Ayres, *supra* note 16, at 2096 (explaining that where a default and accompanying altering rules do not produce "the appropriate separating equilibrium"—meaning that the right consumers stick with the default and the right consumers opt out—"lawmakers will

Privacy by default seems like an elegant, low-cost way to resolve concerns about personal information tracking without imposing a position on consumers. Leaving tracking decisions to individual consumers also sidesteps difficult tradeoffs between incommensurable values and politically perilous substantive judgment calls that policymakers would rather not make. If all policymakers are aiming for with a notice-and-choice regime of information privacy defaults is to avoid political heat, they may succeed. But if they seek to use tracking defaults as a way to set norms, guide consumers to individually or socially desired positions, or inform consumers through the opt-out decision process, they are likely to fail.

More generally, nudges may not be an effective way to help people make better choices about information privacy. Nudges can be powerful when no one is pushing back. But a push can easily overwhelm a nudge. Existing research supporting such nudges is performed in artificial conditions where firms do not have an opportunity to intervene.[279] But firms can use the same mechanisms and conditions that make nudges work to make nudges fail. Nudges' success in the lab shows that people's privacy decisions are heavily influenced by framing—which, ironically, is also evidence that nudges may not work in practice, given that firms can reframe nudges. For these experiments to inform public policy, researchers must anticipate and account for firms' dynamic responses to the proposed nudges.[280]

Yet, the broader effects of tracking defaults on the politics of information privacy are uncertain. On the one hand, tracking defaults could be a useful, collectively educational, political way-station on the road to better information-privacy regulation. A default makes tracking more visible than it would be were Track-Me the only option. While tracking defaults are unlikely to directly lead individuals to make well-informed decisions, defaults could foment social discussion and debates that would inform the populace.[281] In turn, this could create political pressure for better regulation.

On the other hand, Don't-Track-Me defaults might delay or derail better information-privacy regulation for two reasons. First, by creating the façade

need to face the more traditional decision of whether to suspend freedom of contracting altogether and make the socially-preferred rule mandatory").

279. *See supra* note 22.

280. As described, one set of lab experiments aimed at external validity by briefly redirecting consumers' attention away from a disclosure intended to nudge the consumer to engage in privacy-protective decisions. It found that even a fifteen-second delay was enough to negate the effects of the disclosure. Adjerid et al., *supra* note 234.

281. *Cf.* David Adam Friedman, *Micropaternalism*, 88 TUL. L. REV. 75 *passim* (2013) (making similar argument about supersize soda bans and their impact on discussion and debate about food and health, even if the bans fail in the courts).

of robust choice, courts, commentators, and consumers themselves are more likely to blame consumers for any adverse consequences that might flow from staying in the default position or from opting out. Experience with opting out of the civil justice system is instructive here. Some sellers permit consumers to "opt out" of arbitration and class-action waiver clauses in the fine print of their standard-form consumer contracts (and thus opt back into the default civil justice system) while keeping the good or service.[282] These firms have apparently calculated that the costs resulting from the few consumers who will exercise this choice are well worth the benefits the firms receive. These benefits could include deterring self-blaming consumers from challenging the fine-print clauses, convincing courts that the contracts cannot be unconscionable if consumers can opt out of these clauses,[283] or arguing in the political process that substantive regulation of these clauses is unnecessary because consumers can opt out. So too in the privacy realm, Track-Me defaults with which consumers stick en masse and Don't-Track-Me defaults from which consumers opt out en masse might defuse pressure, whether directly from consumers or through the courts or the political process, for more meaningful reform.

Second, a notice-and-choice regime of defaults not only reflects the current understanding of privacy as an individual choice, but re-inscribes it. The model conveys the message that the problem is one of accommodating heterogeneous consumer privacy preferences or heterogeneous consumer calculi about the right tradeoff to make between their privacy preferences

282. *See, e.g.*, *Discover Gift Cardholder Agreement 2010*, DISCOVER CARD, http://www.discovercard.com/shopcenter/giftcard-terms.shtml (last visited Aug. 1, 2013) ("You may reject this Arbitration of Disputes section but only if we receive a written notice of rejection from you within 30 days of your receipt of the card."); *Comcast Agreement for Residential Services*, COMCAST, http://www.comcast.com/Corporate/Customers/Policies/SubscriberAgreement.html (last visited Aug. 1, 2013) ("IF YOU DO NOT WISH TO BE BOUND BY THIS ARBITRATION PROVISION, YOU MUST NOTIFY COMCAST IN WRITING WITHIN 30 DAYS OF THE DATE THAT YOU FIRST RECEIVE THIS AGREEMENT. . . ."); *T-Mobile Terms & Conditions*, T-MOBILE, http://www.t-mobile.com/Templates/Popup.aspx?PAsset=Ftr_Ftr_TermsAndConditions (last visited Aug. 1, 2013) ("YOU MAY CHOOSE TO PURSUE YOUR CLAIM IN COURT AND NOT BY ARBITRATION IF YOU OPT OUT OF THESE ARBITRATION PROCEDURES WITHIN 30 DAYS. . . .").

283. This move has been successful in many courts. *See, e.g.*, Clerk v. ACE Cash Express, Inc., No. 09–05117, 2010 WL 364450 (E.D. Pa. Jan. 29, 2010) (holding class action waiver not unconscionable in part because plaintiff failed to exercise opt-out right); Guadagno v. E*Trade Bank, No. CV 08–03628 SJO (JCX), 2008 WL 5479062 (C.D. Calif. Dec. 29, 2008) (same); Crandall v. AT&T Mobility, LLC, No. 07-750-GPM, 2008 WL 2796752 (S.D. Ill. July 18, 2008) (same); Honig v. Comcast of Georgia, LLC, 537 F. Supp. 2d 1277 (N.D. Ga. Jan. 31, 2007) (holding arbitration clause not unconscionable because consumer failed to exercise opt-out right).

and the value they place on the benefits tracking can provide. Other conceptions of privacy—for example, as a social judgment about when the anonymity required for individual experimentation, reflection, and flourishing necessary for innovation and a liberal democratic society[284] trumps the utility of the free flow of personal information to society[285]—might lead to different policy responses. But it may be that so long as the public understands privacy as a right of individuals "to determine for themselves when, how, and to what extent information about them is communicated to others,"[286] better forms of regulation will remain unimagined.

284. *See, e.g.*, Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904 (2013) (arguing that privacy is necessary for both innovation and democracy); ALLEN, *supra* note 5; Schwartz, *supra* note 18, at 1647–66.

285. *See* Tene & Polonetsky, *supra* note 6.

286. WESTIN, *supra* note 8, at 7.

WHY THEY CAN WATCH YOU: ASSESSING THE CONSTITUTIONALITY OF WARRANTLESS UNMANNED AERIAL SURVEILLANCE BY LAW ENFORCEMENT

Brandon Nagy[†]

TABLE OF CONTENTS

I. **INTRODUCTION** 136

II. **USE OF UNMANNED AIRCRAFT SYSTEMS IN AERIAL SURVEILLANCE** 137

A. LAW ENFORCEMENT USE OF UAS SURVEILLANCE 139
B. NON-LAW ENFORCEMENT USE OF UASS 141

III. **DEVELOPING A FOURTH AMENDMENT PRIVACY FRAMEWORK FOR UAS SURVEILLANCE** 143

A. EARLY JURISPRUDENCE: THE PHYSICAL TRESPASS DOCTRINE—*OLMSTEAD V. UNITED STATES* 143
B. THE "REASONABLE EXPECTATION OF PRIVACY"—*KATZ V. UNITED STATES* 146
C. THE AERIAL SURVEILLANCE CASES 148
1. *Publicly Navigable Airspace as a Lawful Vantage Point*—California v. Ciraolo 149
2. *Constitutionally Protected Zones and Technology Beyond the Naked Eye*—Dow Chemical Co. v. United States 150
3. *Lawful Vantage Points and Objectively Reasonable Expectations of Privacy*—Florida v. Riley 152

© 2014 Brandon Nagy. This Note is the first-place winner of the *Berkeley Technology Law Journal*'s annual student writing competition.

† Law Clerk, the Honorable Senior Judge James A. Teilborg, Federal District of Arizona, Phoenix. J.D., Sandra Day O'Connor College of Law, Arizona State University (2013). The opinions expressed in this Note belong solely to the Author and do not represent the views of the Honorable Senior Judge Teilborg or the Federal District of Arizona. The Author thanks the *Berkeley Technology Law Journal* editorial board and staff for their editorial assistance and the opportunity to publish this Note. Additionally, the Author thanks Edith Cseke for her treasured support. The Author is available via email at brnagy@asu.edu.

4. *Constitutionality of Aerial Surveillance after* Riley*,* Dow*, and* Ciraolo 155
D. MODERN SURVEILLANCE CASES WITH UNMANNED AERIAL SURVEILLANCE IMPLICATIONS 155
1. *Peering into a Constitutionally Protected Zone: The Home*—Minnesota v. Carter 156
2. *Limiting Surveillance Technology*—Kyllo v. United States 158
3. *Visual Observation Without Trespass is Not a "Search"*—United States v. Jones 162
E. FAA REGULATIONS GOVERNING UAS 164
F. OTHER LEGISLATION GOVERNING UAS 166
IV. **LAW ENFORCEMENT UAS SURVEILLANCE WOULD LIKELY BE FOUND CONSTITUTIONAL IN A NUMBER OF SITUATIONS** 168
A. ESTABLISHING A CONSTITUTIONALLY PROTECTED ZONE 169
B. ATTACKING THE LAWFULNESS OF A UAS SURVEILLANCE VANTAGE POINT 170
C. ATTACKING THE SPECIFIC UAS TECHNOLOGY USED 172
V. **CONCLUSION** 174

I. INTRODUCTION

"In the far distance a helicopter skimmed down between the roofs, hovered for an instant like a bluebottle, and darted away again with a curving flight. It was the Police Patrol, snooping into people's windows."[1] Seventy years after penning these words, George Orwell's classic cautionary tale remains relevant. Today, the federal government and law enforcement agencies nationwide operate unmanned aircraft systems ("UASs") for, among other things, aerial surveillance similar to that which Orwell described. To paraphrase Justice Alito, dramatic technological changes have created a time in which popular expectations of privacy are in flux and may ultimately produce significant changes in popular attitudes.[2] These new UAS technologies promise to increase security and convenience at the cost of privacy, but many people find trading privacy for security worthwhile.[3] Other people, however, believe that warrantless law enforcement UAS surveillance violates their privacy and should be unconstitutional under the Fourth Amendment. The U.S. Supreme Court has not ruled on an aerial surveillance

1. George Orwell, *Nineteen Eighty-Four* 2 (Plume Printing 2003) (1949).
2. United States v. Jones, 132 S. Ct. 945, 962 (2012) (Alito, J., concurring).
3. *Id.*

case since 1986, let alone a UAS surveillance case. Instead, a patchwork of other Fourth Amendment precedent and Federal Aviation Administration ("FAA") regulations leave the public and law enforcement guessing about the extent of Fourth Amendment privacy protections from warrantless UAS surveillance. Despite the lack of clear constitutional guidance, both public and law enforcement UAS use continues to grow.

This Note begins by providing background information on the character and capabilities of UASs, as well as describing their current implementation throughout the United States. Next, it describes the current Fourth Amendment privacy framework and current FAA regulations as they affect UAS surveillance to construct an analytical framework. Last, this Note concludes that although UAS surveillance will likely be found constitutional, three potential arguments challenging the constitutionality of UAS surveillance may persuade the Supreme Court otherwise: (1) establishing the victim was in a constitutionally protected zone, (2) attacking the lawfulness of the vantage point from which the UAS surveilled, and (3) attacking the specific UAS technology used.

II. USE OF UNMANNED AIRCRAFT SYSTEMS IN AERIAL SURVEILLANCE

Commonly referred to as a "drone" or "unmanned aerial vehicle," the Federal Aviation Administration ("FAA") defines an "unmanned aircraft" simply as "a device used or intended to be used for flight in the air that has no onboard pilot."[4] This basic definition includes all classes of aircraft controllable in three axes, such as airplanes, helicopters, airships, and translational lift aircraft.[5] This simple description belies the immense diversity of shape and function of unmanned aircraft: "us[ing] aerodynamic forces to provide vehicle lift, [an unmanned aircraft] can fly autonomously or be piloted remotely, can be expendable or recoverable, and can carry a lethal or nonlethal payload."[6] A more expansive moniker, unmanned aircraft system ("UAS"), more accurately reflects the necessity of various accoutrements to operate an unmanned aircraft: "[t]he term [UAS] means an unmanned

4. FEDERAL AVIATION ADMINISTRATION (FAA), INTERIM OPERATIONAL APPROVAL GUIDANCE 08-01, UNMANNED AIRCRAFT SYSTEMS OPERATIONS IN U.S. NATIONAL AIRSPACE SYSTEM 4 (Mar. 13, 2008).

5. *Id.*

6. U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-06-610T, UNMANNED AIRCRAFT SYSTEMS: IMPROVED PLANNING AND ACQUISITION STRATEGIES CAN HELP ADDRESS OPERATIONAL CHALLENGES 5 (2006).

aircraft and associated elements (including communication links and the components that control the unmanned aircraft) that are required for the pilot in command to operate safely and efficiently in the national airspace system."[7] The configurability of UASs allows for a stunning diversity of potential public and private uses, including law enforcement, surveillance, monitoring forest fires, weather research, scientific data collection, farm fence inspection, construction-site photography, and fishery protection.[8]

Like conventional aircraft, UASs can operate at all levels of airspace, generally based on their size. A UAS can broadly be categorized as "large" or "small." "Small" UASs "typically weigh less than 55 pounds, fly below 400 feet above ground level, can stay airborne for several hours, and can be used for reconnaissance, inspection, and surveillance."[9] "Large" UASs, such as the now infamous Predator and Reaper systems, fly higher, for a longer duration, and cost significantly more than small UAS.[10] Consequently, small UASs are "expected to comprise the majority of the UAS[s] that will operate in the national airspace" and "offer a simple and cost effective solution" for law enforcement purposes.[11] Currently, law enforcement agencies and other interested purchasers can "choose from about 146 different types of small UAS[s] being manufactured by about 69 different companies in the U.S."[12] Additionally, some of the smallest camera-equipped UASs are marketed as toys.[13] Given the extremely low cost when compared to other small UASs, it is conceivable that law enforcement agencies could repurpose these "toys" as aerial surveillance tools alongside other, more capable UASs. With such myriad UASs featuring drastically different prices and capabilities, it is perhaps not surprising that both law enforcement and private parties expect to use UAS surveillance in a variety of operations.

---

7. FAA Modernization and Reform Act (FAAMA) of 2012, Pub. L. No. 112-95 §331(9), 126 Stat. 72 [hereinafter FAAMA].

8. U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-12-981, UNMANNED AIRCRAFT SYSTEMS, MEASURING PROGRESS AND ADDRESSING POTENTIAL PRIVACY CONCERNS WOULD FACILITATE INTEGRATION INTO THE NATIONAL AIRSPACE SYSTEM 10 (2012) [hereinafter GAO 2012 UAS REPORT].

9. *Id.* at 5.

10. *Id.*

11. *Id.* at 5 n.8.

12. *Id.* at 11.

13. For example, the Parrot AR.Drone2.0 weighs less than two pounds, comes equipped with a high-definition video camera, and can fly for twelve minutes up to 165 feet from the controller. *AR.Drone2.0*, PARROT.COM, http://ardrone2.parrot.com/ardrone-2/altitude (last visited Dec. 15, 2012) (offering the Parrot AR.Drone2.0 for purchase on Dec. 15, 2012 for $269.99 on Amazon.com).

A. LAW ENFORCEMENT USE OF UAS SURVEILLANCE

Privacy activists are not idly speculating when they worry about law enforcement use of warrantless UAS surveillance—it is already occurring. Currently, several law enforcement agencies operate UASs within the United States, primarily focusing on developing and testing surveillance technologies.[14] Customs and Border Protection ("CBP"), for example, owns nine UASs that it operates "to provide reconnaissance, surveillance, targeting, and acquisition capabilities" for its border security missions.[15] Similarly, without providing operational details, the FBI has admitted that, since late 2006, it has occasionally used UASs for domestic surveillance missions without a warrant.[16]

Non-federal law enforcement agencies across the country also operate a variety of UASs with various cameras for law enforcement surveillance purposes: the Seattle police department's Draganflyer X6 carries still, video, thermal, and low-light cameras;[17] the North Little Rock police department's Rotomotion SR30 carries day zoom or infrared cameras;[18] the Miami-Dade police department's Honeywell RQ-16A Micro Air Vehicle carries electro-optical or infrared cameras;[19] and the Texas Department of Public Safety's Wasp Air Vehicle carries two electro-optical and a thermal nighttime

14. *See* Jennifer Lynch, *These Drones Are Made For Watchin'*, ELECTRONIC FRONTIER FOUNDATION (EFF) (Aug. 16, 2012), https://www.eff.org/deeplinks/2012/08/these-drones-are-made-watchin (discussing documents released to the EFF by the FAA pursuant to FOIA requests).

15. GAO 2012 UAS REPORT, *supra* note 8, at 8–9 n.10.

16. Ryan Gallagher, *FBI Fesses Up: We've Used Domestic Drones 10 Times*, Slate (July 26, 2013), http://www.slate.com/blogs/future_tense/2013/07/26/fbi_admits_in_letter_to_rand_paul_that_it_s_used_domestic_drones_10_times.html.

17. *Draganflyer X6 Description/Technical Specifications*, EFF, https://www.eff.org/document/seattle-pd-2010aircraft-system (last visited Dec. 15, 2012). On February 7, 2013, however, Seattle's mayor "ordered the police department to abandon its plan to use drones after residents and privacy advocates protested." *Seattle Mayor Ends Police Drone Efforts*, USA TODAY (Feb. 7, 2013), http://www.usatoday.com/story/news/nation/2013/02/07/seattle-police-drone-efforts/1900785. The Seattle Police Department has agreed to return the UAS to the manufacturer. *Id.*

18. The Rotomotion SR30 weighs twenty-five pounds and can fly for seventy-five minutes. *North-Little-Rock-PD-2011_Aircraft_System*, EFF, https://www.eff.org/document/north-little-rock-pd-2011aircraftsystem (last visited Dec. 15, 2012).

19. The Honeywell RQ-16A weighs 16.3 pounds and can fly for forty minutes up to 10,000 feet. *MiamiDade-PD-2011-ESA56 Aircraft System*, EFF, https://www.eff.org/document/miamidade-pd-2011-esa56-aircraft-system (last visited Dec. 15, 2012); *T-Hawk*, HONEYWELL (June 2010), http://aerospace.honeywell.com/~/media/UWSAero/common/documents/myaerospacecatalog-documents/Defense_Brochures-documents/T-Hawk_Unmaned_Micro_Air_Vehicle.pdf.

camera.[20] Seeing potential demand, UAS manufacturers continue to develop and market new models, such as AeroVironment's Qube, designed specifically for law enforcement surveillance.[21]

As public concern over the privacy implications of law enforcement UAS operations increases, other departments pursuing UASs, such as the Alameda County, California, Sheriff's Office, publicly downplay surveillance missions in favor of other UAS applications: "[the Sheriff's Office] object[s] to the term surveil. We have no intention of doing that."[22] Instead, "[UAS surveillance] will provide real-time situational analysis for first responders to include search and rescue missions, tactical operations, disaster response, recovery and damage assessment, explosive ordnance response, wildland and structure fire response and response to Hazmat incidents."[23] Despite rising public concern[24] and the FAA regulatory hurdles,[25] the relative simplicity of use and low cost of operation allows UASs to provide an attractive option for airborne law enforcement activities.[26] For the price of a patrol car (approximately $30,000 to $50,000, far less than the cost of a manned aircraft), an agency can augment its airborne capabilities, especially surveillance.[27] It is no wonder, then, that industry forecasts predict state and local law enforcement agency orders will represent the greatest proportion of near-term growth in the small UAS market.[28]

20. *Texas-Dept-Public-Safety-2008 Communications*, EFF, https://www.eff.org/document/texas-dept-public-safety-2008-communications (last visited Dec. 15, 2012).

21. Isolde Raftery, *Anticipating Domestic Boom, Colleges Rev up Drone Piloting Programs*, NBCNEWS.COM (Jan. 29, 2013), http://openchannel.nbcnews.com/_news/2013/01/29/16726198-anticipating-domestic-boom-colleges-rev-up-drone-piloting-programs. AeroVironment describes the Qube as "public safety small UAS" that has been designed to be "transported in the trunk of a police vehicle . . . or carried in a backpack" and to "provide immediate situational awareness to first responders, giving them a birds-eye view of the situation, day or night, to save lives and protect property." *Qube: Public Safety Small UAS*, AeroVironment, http://www.avinc.com/uas/small_uas/qube/ (last visited Feb. 28, 2013).

22. David Kravets, *California County Inching Toward Drone Deployment?*, WIRED MAG. (Feb. 14, 2013), http://www.wired.com/threatlevel/2013/02/alameda-drone-deployment.

23. David Kravets, *California Eyeing Drone Surveillance*, WIRED MAG. (Dec. 4, 2012), http://www.wired.com/2012/12/california-drones.

24. *See infra* Section III.F (discussing other UAS legislation).

25. *See infra* Section III.E (discussing FAA Regulations).

26. GAO 2012 UAS REPORT, *supra* note 8, at 11.

27. *Id.*

28. *See id.*

B. NON-LAW ENFORCEMENT USE OF UASS

It is difficult to measure the extent of current private use of UASs, in part because the FAA permit statistics have aggregated all non-military users.[29] In the first half of 2012, law enforcement, academic institutions, and other government-related entities received a total of 342 permits to operate UASs (generally allowing operations for twelve to twenty-four months).[30] During the same time period, private sector entities, specifically UAS manufacturers, received just eight "experimental" permits granting UAS operations for research and development purposes only.[31]

In addition, many private users of UASs are simply not required to apply for FAA permits, which prevents an exact measurement of private UAS use.[32] More precisely, the FAA "may not promulgate any rule or regulation regarding a model aircraft . . . [if] the aircraft is flown strictly for hobby or recreational use."[33] A model aircraft can weigh up to fifty-five pounds and has no specific limitations on its flight characteristics.[34] Therefore, a private actor, acting solely as a hobbyist, may operate a small UAS that is essentially the same aircraft that law enforcement agencies must operate with a permit.[35] Although it is unclear exactly what "hobby or recreational" use entails,[36] private commercial users likely must obtain the proper permits before operating a UAS.[37] In fact, the FAA has previously noted that the aerial photography industry and others conducting remote sensing activities had "mistakenly interpreted FAA advisory circular 91-57, *Model Aircraft Operating Standards*, for permission to operate small UAS for research or compensation or hire purposes."[38] Given the varied and growing range of potential

29. *Id.* at 6–7.
30. *Id.*
31. *Id.* at 7.
32. FAAMA, *supra* note 7, § 336.
33. *Id.*
34. *Id.*
35. GAO 2012 UAS REPORT, *supra* note 8, at 31.
36. There is confusion within the UAS industry and FAA about the applicability of the hobbyist guidelines. For example, "[a] farmer can be a modeler if they operate their aircraft as a hobby or for recreational purposes," but a UAS manufacturer's employee cannot "fly as a hobbyist [his] own [UAS] over property that [he] own[s]." Spencer Ackerman, *Domestic-Drone Industry Prepares for Big Battle with Regulators*, WIRED MAG. (Feb. 13, 2013), http://www.wired.com/2013/02/drone-regulation.
37. FAAMA, *supra* note 7, § 336.
38. FAA, ORDER 1110.150, SMALL UNMANNED AIRCRAFT SYSTEM AVIATION RULEMAKING COMMITTEE 1 (Apr. 10, 2008), *available at* http://www.faa.gov/documentLibrary/media/Order/1110.150.pdf.

commercial uses[39] and the wide availability of extremely low cost but capable "toy" UASs,[40] it is certainly possible that significant private commercial and other non-"hobby or recreational" UAS activity occurs without proper FAA permits.

In the future, the UAS industry expects the private UAS sector to boom.[41] Indeed, Jim Williams of the FAA has predicted that "the potential market for government and commercial drones could generate 'nearly $90 billion in economic activity' over the next decade."[42] Because the military is not expected to significantly expand its UAS fleet and sales to law enforcement are limited by the fact that there are only 18,000 agencies in the United States, much of this expansion in the UAS market is forecast to stem from the private farming sector.[43] Indeed, twenty-one public universities, each focused on agriculture, have already received FAA clearance to operate UASs.[44] Anticipating that "10,000 commercial drones will be operating" within the United States by 2020, three colleges have already begun offering degrees in piloting UASs.[45] Additionally, "many others—including community colleges—offer training for remote pilots."[46] These programs focus on teaching non-military applications suitable both for law enforcement, such as tracking fleeing criminal suspects, and for private parties, such as "[m]onitoring livestock and oil pipelines, spotting animal poachers, . . . and delivering packages for UPS and FedEx."[47] If these expectations come to even partial fruition, then the U.S. public can expect non-law enforcement UAS use to become a fact of life. Even now, with dozens of federal, state, and local law enforcement agencies using and pursuing UASs, a constitutional framework to understand the privacy implications of UAS surveillance is clearly necessary.

39. The GAO lists numerous examples of commercial use: pipeline, utility, and farm fence inspections; vehicular traffic monitoring; real-estate and construction-site photography; telecommunication signal relay; fishery protection and monitoring; and crop dusting. GAO 2012 UAS REPORT, *supra* note 8, at 10.

40. *See, e.g.*, PARROT.COM, *supra* note 13.

41. Ackerman, *supra* note 36.

42. *Id.*

43. Spencer Ackerman, *Drone Boosters Say Farmers, Not Cops, Are the Biggest U.S. Robot Market*, WIRED MAG. (Feb. 5, 2013), http://www.wired.com/2013/02/drone-farm.

44. *Id.*

45. Raftery, *supra* note 21.

46. *Id.*

47. *Id.*

III. DEVELOPING A FOURTH AMENDMENT PRIVACY FRAMEWORK FOR UAS SURVEILLANCE

The word "privacy" is not found within the U.S. Constitution, but for the past hundred years, the U.S. Supreme Court has been using the Fourth Amendment to protect a nebulous conception of privacy from technological advances that threaten to undermine it. The Fourth Amendment proclaims in part that "the right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated."[48] As technology has evolved, so too has the Supreme Court's jurisprudence on what law enforcement applications of these new technologies constitute an "unreasonable search and seizure." In this way, the Supreme Court has used the Fourth Amendment to determine "what limits there are upon . . . technology to shrink the realm of guaranteed privacy."[49]

Because the Fourth Amendment protects against "unreasonable" searches (in contrast to "reasonable" searches), jurisprudence focuses on two questions: first, whether a Fourth Amendment search occurred, and second, whether the search was unreasonable. If a court concludes that no search occurred, then the action cannot have violated the Fourth Amendment.[50] The two questions are important to distinguish because, "with few exceptions, the question of whether a warrantless search of a home is reasonable and hence constitutional must be answered no."[51] Consequently, whether or not a Fourth Amendment search occurred will often be the dispositive question in determining whether or not the Fourth Amendment protects a person's privacy from a technological advancement in law enforcement surveillance.

A. EARLY JURISPRUDENCE: THE PHYSICAL TRESPASS DOCTRINE—*OLMSTEAD V. UNITED STATES*

Early Supreme Court jurisprudence on visual surveillance linked common-law trespass to the question of whether or not a search occurred. In this way, the common law was clear: visual surveillance of a person or home was lawful because "the eye cannot by the laws of England be guilty of a

48. U.S. CONST. amend. IV.

49. Kyllo v. United States, 533 U.S. 27, 34 (2001).

50. *See id.* at 31 (noting that whether or not a "Fourth Amendment 'search' " has occurred is "the antecedent question").

51. *Id.* (citing Illinois v. Rodriguez, 497 U.S. 177, 181 (1990); Payton v. New York, 445 U.S. 573, 586 (1980)).

trespass."[52] In 1928, however, this static trespass doctrine was challenged by a technological advance in law enforcement surveillance—telephone wiretapping.

In *Olmstead v. United States*, the Supreme Court evaluated whether or not telephone wiretapping constituted an unreasonable search.[53] While investigating Olmstead's suspected Prohibition-era bootleg activities, federal prohibition officers tapped his telephone by inserting "small wires along the ordinary telephone wires from the residence [of the defendant] and those leading from the chief office."[54] The federal prohibition officers did not trespass upon any property of the defendant because the taps were inserted into telephone wires located in the basement of the large office building and along streets near the house.[55]

In order to adapt existing Fourth Amendment jurisprudence to the novel technology of telephone wiretapping, the *Olmstead* litigants offered competing analogies.[56] The Government argued that the Court should analogize wiretapping to "traditional visual surveillance from a lawful vantage point."[57] Olmstead, however, argued that the government's actions were more like opening postal mail, which violates Fourth Amendment protections.[58] By a 5–4 majority, the Court endorsed the government's analogy to visual surveillance and concluded that Olmstead had not been searched.[59]

Writing for the one-justice majority, Chief Justice Taft explained two rationales. First, unlike "material things—the person, the house, his papers, or his effects," Olmstead's conversations were immaterial.[60] Therefore, not even a liberal interpretation of the Fourth Amendment in the "interest of liberty" could "justify enlargement of the language employed . . . to apply the

52. Boyd v. United States, 116 U.S. 616, 628 (1886) (quoting Entick v. Carrington, 95 Eng. Rep. 807 (Gr. Brit. 1765)).

53. Olmstead v. United States, 277 U.S. 438 (1928).

54. *Id.* at 456–57.

55. *Id.*

56. Luke M. Milligan, *The Real Rules of "Search" Interpretations*, 21 WM. & MARY BILL RTS. J. 1, 9 (2012).

57. *Id.* at 9 (citing *Olmstead*, 277 U.S. at 447–52 (citing Hester v. United States, 265 U.S. 57 (1924)) (holding that visual surveillance from open fields does not violate the Fourth Amendment)).

58. *Id.* (citing *Olmstead*, 277 U.S. at 464 ("It is urged that the language of Mr. Justice Field in *Ex Parte Jackson* . . . offers an analogy."); *Ex parte* Jackson, 96 U.S. 727 (1878) (holding that the Fourth Amendment is applicable to sealed letters)).

59. *Olmstead*, 277 U.S. at 466.

60. *Olmstead*, 277 U.S. at 464.

words search and seizure as to forbid hearing or sight."[61] Chief Justice Taft's second rationale, however, relied on prior trespass case law: "[no precedential cases] hold the Fourth Amendment to have been violated as against a defendant, unless there has been . . . an actual physical invasion of his house 'or curtilage' for the purpose of making a seizure."[62] Thus, by concluding that "[t]he intervening [telephone] wires [were] not part of [the defendant's] house or office, any more than are the highways along which they stretched,"[63] the federal prohibition agent's wiretapping activities could not be a Fourth Amendment search. Like visual surveillance from a lawful vantage point, the wiretaps were not a physical intrusion into a constitutionally protected area (such as a house, office, or curtilage).

In a separate dissent that has been characterized as "one of the most famous dissents in Supreme Court history,"[64] Justice Brandeis characterized wiretapping as a "[s]ubtler and more far-reaching means of invading privacy" and prophetically warned that "[t]he progress of science in furnishing the government with means of espionage is not likely to stop with wire tapping."[65] Arguing that the "makers of our Constitution" conferred upon the people, as against the government, "the right to be let alone," Brandeis concluded that every "unjustifiable intrusion by the government upon the privacy of the individual, whatever the means employed, must be deemed a violation of the Fourth Amendment."[66]

Like Justice Brandeis in his dissent, the public did not react favorably to *Olmstead*'s holding that law enforcement could wiretap telephone conversations without a warrant. Consequently, in 1934, Congress made wiretapping a federal crime (thereby rendering warrantless wiretapping illegal and the fruits of it inadmissible) by enacting § 605 of the Federal Communications Act.[67] Despite this, *Olmstead*'s analogy to a physical trespass continued to control, largely unchanged, for nearly forty years. For example, in *Lee v. United States*, the Court analogized an audio technological advance (a small radio transmitter worn by a conversant to transmit the conversation to remote law enforcement agents) to tools for improved visual surveillance (presumably from a lawful vantage point):

---

61. *Id.* at 465.
62. *Id.* at 466.
63. *Id.* at 465.
64. DANIEL J. SOLOVE & PAUL M. SCHWARTZ, INFORMATION PRIVACY LAW 260 (4th ed. 2011).
65. *Olmstead*, 277 U.S. at 473–74 (Brandeis, J., dissenting).
66. *Id.* at 478.
67. SOLOVE & SCHWARTZ, *supra* note 64, at 260–61.

> [Defendant] was talking confidentially and indiscreetly with one he trusted, and he was overheard. This was due to aid from a transmitter and receiver, to be sure, but with the same effect on his privacy as if [the law enforcement agent] had been eavesdropping outside an open window. The use of bifocals, field glasses or the telescope to magnify the object of a witness' vision is not a forbidden search or seizure . . . .[68]

In 1961, the *Olmstead* physical trespass analogy was tweaked by the holding in *Silverman v. United States* that an actual trespass (as defined by local law) was not necessary as long as there was "an actual intrusion into a constitutionally protected area."[69] In *Silverman*, government agents eavesdropped on private conversations within the defendant's house by attaching a "spike mike" listening device to a heating duct running into the defendant's house. This action differed from the wiretapping in *Olmstead* because, in effect, the agent's actions turned the entire heating duct of the house into a microphone. Calling this action an "usurp[tion of] part of the [defendant's] house or office," the Court concluded that the eavesdropping constituted an "unauthorized physical encroachment" (in essence, a constructive trespass).[70] While adjusting the *Olmstead* rule, the *Silverman* Court included strong language reminding government agents of the importance of the Fourth Amendment: "[a]t the very core [of the Fourth Amendment] stands the right of a man to retreat into his own home and there be free from unreasonable governmental intrusion."[71] Although *Silverman* refrained from using "privacy" to expand the Fourth Amendment's protections beyond concepts of trespass, its sentiment nonetheless hinted at the future direction of Fourth Amendment privacy jurisprudence.

B. THE "REASONABLE EXPECTATION OF PRIVACY"—*KATZ V. UNITED STATES*

In 1967, the Supreme Court drastically changed gears by rejecting *Olmstead*'s physical trespass analogy and accepting the idea of defining a Fourth Amendment search based on the victim's intentions or expectations.[72] In *Katz v. United States*, the Court evaluated whether or not placing an "electronic listening and recording device" on the outside of a public

68. Lee v. United States, 343 U.S. 747, 753–54 (1952).

69. Silverman v. United States, 365 U.S. 505, 512 (1961).

70. *Id.* at 510–12.

71. *Id.* at 511 (citing Boyd v. United States, 116 U.S. 616, 626 (1886); Entick v. Carrington, 95 Eng. Rep. 807 (Gr. Brit. 1765)).

72. *See* Katz v. United States, 389 U.S. 347, 353 (1967).

telephone booth for the purposes of overhearing a conversation within the closed public telephone booth constituted an unreasonable search.[73] Framing the issue in terms of a physical trespass on a constitutionally protected area, the lower courts upheld the government's actions based on *Olmstead* and its progeny.[74] Consequently, the Court granted certiorari "in order to consider the constitutional questions" of whether a public telephone booth is a "constitutionally protected area" and whether a "physical penetration of a constitutionally protected area" is necessary to be a search and seizure violative of the Fourth Amendment.[75] The Court, however, rejected "this formulation of the issues" and broke with *Olmstead* by stating that the "correct solution to Fourth Amendment problems is not necessarily promoted by the incantation of the phrase 'constitutionally protected area.' "[76] To be explicitly clear in its rejection of *Olmstead*, the Court found that the resolution of Fourth Amendment issues "cannot turn upon the presence or absence of a physical intrusion into any given enclosure" and that the " 'trespass' doctrine . . . can no longer be regarded as controlling."[77]

Justice Stewart, writing the majority opinion, offered a new kind of Fourth Amendment search analysis in place of *Olmstead*:

> [T]he Fourth Amendment protects people, not places. What a person knowingly exposes to the public, even in his own home or office, is not a subject of Fourth Amendment protection. But what he seeks to preserve as private, even in an area accessible to the public, may be constitutionally protected.[78]

Under this formulation of the Fourth Amendment, the Court must look to the subjective intent of the defendant inside the phone booth, rather than to the physical characteristics of the phone booth. Here, Katz manifested intent to preserve his privacy by entering the telephone booth, closing the door, and paying the fee. Thus, Katz was "surely entitled to assume that the words he utter[ed] into the mouthpiece [would] not be broadcast to the world. To read the Constitution more narrowly is to ignore the vital role that the public telephone has come to play in private communication."[79]

---

73. *Id.* at 348.

74. *Id.* at 349–50; *see* Goldman v. United States, 316 U.S. 129, 134–35 (1942) (finding the warrantless use of a "detectaphone" placed on the outside of a wall to overhear conversations originating within the adjacent, enclosed office constitutional).

75. *Katz*, 389 U.S. at 350.

76. *Id.*

77. *Id.* at 353.

78. *Id.* at 351 (citations omitted).

79. *Id.* at 352.

By allowing the evolving role and use of technology in society—here the telephone—to impact its decision, the *Katz* majority attempted "to incorporate an objective and evolving privacy standard into [Fourth Amendment search] decisionmaking."[80] The majority opinion, however, lacked clear standards to guide future courts.[81] Justice Harlan, in his concurrence, supplemented the majority's "people-not-places" discussion with a two-pronged "reasonable expectation of privacy" test: first, whether the person "exhibited an actual (subjective) expectation of privacy," and second, whether that "expectation [is] one that society is prepared to recognize as 'reasonable.' "[82] Thus, to be a Fourth Amendment search (and therefore require a warrant), the government action must impinge on both a person's subjective expectation of privacy and society's objectively reasonable expectation of privacy. Justice Harlan's two-part test would come to be known as the *Katz* test[83] and remains the rule today.[84] Furthermore, supplementing his own two-prong test, Justice Harlan expressed that: "a man's home is, for most purposes, a place where he expects privacy, but objects, activities, or statements that he exposes to the 'plain view' of outsiders are not 'protected' because no intention to keep them to himself has been exhibited."[85] Although left undefined in his concurrence, Justice Harlan's "plain view" analysis would become critical to understanding a person's "reasonable expectation of privacy" in the aerial surveillance cases that would later arise.[86]

C. THE AERIAL SURVEILLANCE CASES

During the 1980s, the Supreme Court granted certiorari for three cases—*Ciraolo*,[87] *Dow*,[88] and *Riley*[89]—presenting questions about the constitutionality of warrantless government aerial surveillance for domestic enforcement

80. Milligan, *supra* note 56, at 15.

81. *Id.* at 17.

82. *Katz*, 389 U.S. at 361 (Harlan, J., concurring).

83. *See, e.g.*, Minnesota v. Carter, 525 U.S. 83, 97 (1998) (Scalia, J., concurring) (commenting on "the Katz test (which has come to mean the test enunciated by Justice Harlan's separate concurrence in Katz)").

84. *But see, e.g.*, United States v. Jones, 132 S. Ct. 945, 952 (2012) ("the Katz reasonable-expectation-of-privacy test has been *added to*, not *substituted for*, the common-law trespassory test.").

85. *Katz*, 389 U.S. at 361 (Harlan, J., concurring).

86. Travis Dunlap, *We've Got Our Eyes on You: When Surveillance by Unmanned Aircraft Systems Constitutes a Fourth Amendment Search*, 51 S. TEX. L. REV. 173, 185–86 (2009).

87. California v. Ciraolo, 476 U.S. 207, 209 (1986).

88. Dow Chem. Co. v. United States, 476 U.S. 227, 229 (1986).

89. Florida v. Riley, 488 U.S. 445, 449–51 (1989).

purposes. In each case, the Court applied the two-pronged *Katz* test and determined that no "subjective expectation of privacy that society was willing to accept as reasonable" existed.[90] Consequently, the Court held that no Fourth Amendment search had taken place and the warrantless aerial surveillance at issue was constitutional.

1. Publicly Navigable Airspace as a Lawful Vantage Point—California v. Ciraolo

In the 1986 case of *California v. Ciraolo*, the Supreme Court evaluated whether aerial surveillance conducted with the naked eye was a Fourth Amendment search requiring a warrant.[91] Acting on an anonymous tip that Ciraolo was growing marijuana in his backyard, California police officers began an investigation.[92] The officers' attempts to peer into Ciraolo's backyard (part of the curtilage of his home) from ground level were stymied by two fences (a six-foot outer fence and a ten-foot inner fence) that completely enclosed Ciraolo's yard.[93] Undeterred, the officers "secured a private airplane and flew over [Ciraolo's] house at an altitude of 1,000 feet, within navigable airspace."[94] From that height, the officers observed marijuana plants growing within Ciraolo's curtilage and "photographed the area with a standard 35mm camera."[95]

Applying *Katz*'s two-prong "reasonable expectation of privacy" test, the Court determined that Ciraolo satisfied the first prong when he manifested a subjective expectation of privacy by erecting the two fences.[96] The 5–4 majority, instead, focused on the second prong: "whether the government's intrusion infringes upon the personal and societal values protected by the Fourth Amendment."[97] As part of this analysis, Chief Justice Burger, writing the majority opinion, initially noted that the fact that an "area is within the curtilage does not itself bar all police observation" and that the Fourth Amendment does not "require law enforcement officers to shield their eyes when passing a home on public thoroughfares."[98] Further, Burger explained that an individual's ineffective measures to "restrict some views of his

90. Paul McBride, *Beyond Orwell: The Application of Unmanned Aircraft Systems in Domestic Surveillance Operations*, 74 J. AIR L. & COM. 627, 642 (2009).

91. *Ciraolo*, 476 U.S. at 209.

92. *Id.*

93. *Id.*

94. *Id.*

95. *Id.*

96. *Id.* at 211–12 (quoting Rawlings v. Kentucky, 448 U.S. 98, 105 (1980)).

97. *Id.* at 212 (quoting Oliver v. United States, 466 U.S. 170, 181–83 (1984)).

98. *Id.* at 213.

activities" does not protect against an officer's lawful "observations from a public vantage point . . . [that] render[] the activities clearly visible."[99] Burger then examined the lawfulness of the vantage point, noting that the police made the observation within "publicly navigable airspace," in a "physically nonintrusive manner," and that the marijuana plants were "readily discernible to the naked eye."[100] Based on these factors, Burger determined that no Fourth Amendment search occurred.[101] Burger's third caveat—the naked eye—would at first appear to be a bright-line for aerial surveillance decisions; however, its companion case, *Dow*, immediately added shades of grey.

2. *Constitutionally Protected Zones and Technology Beyond the Naked Eye*—Dow Chemical Co. v. United States

In the 1986 case of *Dow Chemical Co. v. United States*, the Supreme Court evaluated whether aerial surveillance of an industrial manufacturing complex conducted with an aerial mapping camera was a Fourth Amendment search requiring a warrant.[102] The Environmental Protection Agency ("EPA") sought an inspection of a 2000-acre Dow Chemical Company manufacturing and industrial facility, but was denied access.[103] The facility's perimeter fencing and elaborate security effectively blocked the facility from ground-level view.[104] Although numerous buildings in the facility were covered, some manufacturing equipment and various piping conduits were visible from an aerial vantage point. Consequently, the EPA hired a commercial aerial photographer, using a "standard floor-mounted, precision aerial mapping camera," to photograph the facility at altitudes of 1200, 3000, and 12,000 feet (always within publicly navigable airspace).[105]

On the same day as the *Ciraolo* decision, Chief Justice Burger also delivered the *Dow* decision, which tackled two issues not found in *Ciraolo*: first, whether the common-law "curtilage doctrine" applied to an industrial manufacturing facility, and second, whether the enhanced photography (capable of discerning piping as little as one half of an inch wide) constituted a Fourth Amendment search.[106] On the first issue, the Court noted that the

99. *Id.*

100. *Id.* at 213–14 (defining publicly navigable airspace, the court observed that "[a]ny member of the public flying in this airspace who glanced down could have seen everything that these officers observed").

101. *Id.*

102. Dow Chem. Co. v. United States, 476 U.S. 227 (1986).

103. *Id.* at 229.

104. *Id.*

105. *Id.*

106. *Id.* at 235.

"curtilage area immediately surrounding a private house has long been given protection as a place where the occupants have a reasonable and legitimate expectation of privacy that society is prepared to accept."[107] The Court further defined curtilage as "the area which extends the intimate activity associated with the sanctity of a man's home and the privacies of life."[108] After contrasting curtilage with the lesser protections afforded an "open field,"[109] the Court determined that Dow's "industrial curtilage" was more analogous to an "open field" because "[t]he intimate activities associated with family privacy and the home and its curtilage simply do not reach the outdoor areas or spaces between structures and buildings of a manufacturing plant."[110] As an "open field," the Court held that Dow did not have an objectively reasonable expectation of privacy (the second part of the *Katz* test) because the facility was "open to the view and observation of persons in aircraft lawfully in the public airspace immediately above or sufficiently near the area for the reach of cameras."[111]

Having already found Dow did not have an objective expectation of privacy, the Court only briefly considered the second issue of the enhanced photography. Despite the district court's finding that the aerial photographer used a $22,000 camera described as "the finest precision aerial camera available" mounted in a specialty airplane,[112] the Supreme Court dismissed the camera as a "conventional, albeit, precise commercial camera commonly used in mapmaking"[113] that could be "readily duplicate[d] by "any person with an airplane and an aerial camera."[114] Consequently, although the photographs "undoubtedly" gave the EPA "more detailed information than naked-eye views," the enhanced photographs did not constitute a Fourth Amendment search.[115] The Court concluded its discussion by discriminating between acceptable enhancements, like the technology used by the EPA, and sophisticated surveillance "not generally available to the public, such as satellite technology," the use of which may be prohibited by the

107. *Id.*

108. *Id.* at 236 (quoting Oliver v. United States, 466 U.S. 170, 180 (1984)) (internal quotation marks omitted).

109. *Id.* at 236 (citing *Oliver*, 466 U.S. at 178).

110. *Id.* at 236.

111. *Id.* at 239.

112. Dow Chem. Co. v. United States, 536 F. Supp. 1355, 1357 n.2 (E.D. Mich. 1982), *rev'd*, 749 F.2d 307 (6th Cir. 1984), *aff'd*, 476 U.S. 227 (1986).

113. *Id.* at 238.

114. *Id.* at 231.

115. *Id.* at 238.

Constitution.[116] This closing thought—the potential relevance of a surveillance technology's availability to the public—would prove to be dispositive fifteen years later in *Kyllo v. United States*.[117]

3. Lawful Vantage Points and Objectively Reasonable Expectations of Privacy—Florida v. Riley

In 1989, only three years after *Ciraolo* and *Dow*, the Supreme Court again evaluated the constitutionality of aerial surveillance and issued its most recent opinion on the subject.[118] In *Florida v. Riley*, the Supreme Court evaluated "[w]hether surveillance of the interior of a partially covered greenhouse in a residential backyard from the vantage point of a helicopter located 400 feet above the greenhouse constitutes a 'search' for which a warrant is required under the Fourth Amendment."[119] Similar to the facts in *Ciraolo*, police began their investigation after receiving an anonymous tip that Riley was growing marijuana on his property.[120] Riley, like the defendant in *Ciraolo*, had concealed his activities from ground level observation.[121] Unlike the defendant in *Ciraolo*, however, Riley also attempted to prevent aerial observation of his activities by growing the marijuana within a greenhouse "covered by corrugated roofing panels, some translucent and some opaque."[122] The investigating officer, unable to see inside the greenhouse from ground level, circled twice over the greenhouse in a helicopter at an altitude of 400 feet.[123] The officer, with his naked eye, observed that two panels (approximately ten percent) of the greenhouse roof were missing and within that gap he could see marijuana plants growing.[124]

Writing for a plurality of the Court, Justice White found that *Ciraolo*'s fixed-wing aircraft analysis should be applied to the helicopter aerial surveillance of Riley's property. Noting that Riley met the first *Katz* prong (subjective expectation of privacy) because the greenhouse was within his curtilage and covered, the Court examined whether Riley's expectation of

116. *Id.* Because there is no further discussion of the type of curtilage being surveilled, "at least to the degree here" appears to relate only to sight-enhancing equipment and not to a distinction between industrial curtilage and Ciraolo's residential curtilage. *See id.*

117. Kyllo v. United States, 533 U.S. 27 (2001).

118. McBride, *supra* note 90, at 645.

119. 488 U.S. 445, 447–48 (1989).

120. *Id.* at 448.

121. *Id.*

122. *Id.*

123. *Id.*

124. *Id.*

privacy was objectively reasonable.[125] The Court determined that Riley had no reasonable expectation of privacy from helicopter surveillance conducted within the navigable airspace of a fixed-wing plane because the two missing roof panels would have enabled the public to observe the greenhouse from the same height.[126] Because the helicopter surveilled Riley from 400 feet, an altitude below the 500-foot threshold for fixed-wing craft, the Court next examined the lawfulness of flying a helicopter at 400 feet.[127] Finding that operation of a helicopter at 400 feet is lawful "if the operation is conducted without hazard to persons or property on the surface,"[128] the Court examined the record and found no "interfere[nce] with respondent's normal use of the greenhouse" because the helicopter surveillance created no "undue noise, . . . wind, dust, or threat of injury."[129] Consequently, this helicopter surveillance did not violate the law, breached no objectively reasonable expectation of privacy, and was not a Fourth Amendment search.[130]

Of particular note is that within its "lawfulness" analysis (here, key to the second prong of the *Katz* test), the plurality's application of property concepts related to trespass, nuisance, and interference hearkens back to *Olmstead*[131] and its rejected physical-trespass doctrine. This was not the Court's first post-*Katz* "reference to concepts of real or personal property law" in order to determine if a reasonable expectation of privacy exists.[132] Still, it is notable because in a Fifth Amendment takings case forty-three years prior, the Court characterized the navigable airspace (defined as "airspace above the minimum safe altitudes of flight prescribed by the [FAA]") above private property as "a public highway."[133] In doing so, the Court abolished the common-law "ancient doctrine that . . . ownership of the land extend[s] to the periphery of the universe" and declared that recognizing "private claims [of] the airspace would . . . clog these highways, seriously interfere with their control and development in the public interest, and

125. *Id.* at 450.
126. *Id.* at 450–51.
127. *Id.*
128. *Id.* at 451 n.3 (citing the controlling FAA regulations found at 14 CFR § 91.79 (1988)).
129. *Id.* at 452.
130. *Id.*
131. Olmstead v. United States, 277 U.S. 438, 466 (1928).
132. Rakas v. Illinois, 439 U.S. 128, 143–44 & n.12 (1978) (holding that, to be objectively reasonable, an expectation of privacy must have "a source outside of the Fourth Amendment, either by reference to concepts of real or personal property law or to understandings that are recognized and permitted by society.").
133. U.S. v. Causby, 328 U.S. 256, 260–61 (1946) (internal quotations and citations omitted).

transfer into private ownership that to which only the public has a just claim."[134] Indeed, noticing the apparent inconsistency in *Riley*'s plurality decision, Justice Brennan—joined in dissent by Justices Marshall and Stevens—explicitly questioned the plurality's reliance on such concepts: "Riley might have a cause of action in inverse condemnation, but that is not what the Fourth Amendment is all about," and "[i]f indeed the purpose of the restraints imposed by the Fourth Amendment is to 'safeguard the privacy and security of individuals,' then it is puzzling why it should be the helicopter's noise, wind, and dust that provides the measure of whether this constitutional safeguard has been infringed."[135]

The Court, however, declined to establish the "lawfulness" of a flight (a determination apparently based, in part, on analysis of an aerial surveillance victim's property rights) as a bright-line rule: "[t]his is not to say that an inspection of the curtilage of a house from an aircraft will always pass muster under the Fourth Amendment simply because the plane is within the navigable airspace specified by law."[136] Instead, the Court noted that Riley had adduced no empirical evidence "to suggest that helicopters flying at 400 feet are sufficiently rare in this country" to support his claim of an objectively reasonable expectation of privacy.[137] This additional rationale may have been added to persuade Justice O'Connor to concur because her concurrence heavily criticizes the plurality's reliance on lawfulness based on compliance with FAA regulations.[138] Justice O'Connor reasoned that relying on FAA regulations promulgated to "promote air safety" would abdicate Fourth Amendment determinations to the executive branch and that "the fact that a helicopter could conceivably observe the curtilage at virtually any altitude or angle, without violating FAA regulations, does not in itself mean that an individual has no reasonable expectation of privacy from such observation."[139] Justice O'Connor then opined that the relevant *Katz* test was "whether the helicopter was in the public airways at an altitude at which members of the public travel with sufficient regularity that Riley's expectation of privacy from aerial observation was not one that society is prepared to recognize as reasonable."[140] Emphasizing that "[i]f the public

134. *Id.*

135. *Riley*, 488 U.S. at 462 (Brennan, J., dissenting) (quoting Justice White's majority opinion in Camara v. Municipal Court, 387 U.S. 523, 528 (1967)).

136. *Id.* at 451 (majority opinion).

137. *Id.*

138. *Id.* at 452 (O'Connor, J., concurring).

139. *Id.* at 452–54.

140. *Id.* at 454 (internal quotations and citation omitted).

rarely, if ever, travels overhead at such altitudes, the observation cannot be said to be from a vantage point generally used by the public," but agreeing with the plurality that Riley had not introduced evidence demonstrating such flights were rare, O'Connor concurred that Riley lacked an objective expectation of privacy.[141]

4. *Constitutionality of Aerial Surveillance after* Riley*,* Dow*, and* Ciraolo

In sum, post-*Riley*, *Ciraolo*, and *Dow*, the constitutionality of warrantless aerial surveillance of a home's curtilage appears to be decided on a case-by-case assessment of the two-pronged *Katz* test. The reasonable expectation of privacy prong, however, appears to depend primarily on two factors: (1) the lawfulness of the flight, and (2) whether the aerial surveillance victim can adduce sufficient evidence that such flights were sufficiently rare. *Riley* and *Ciraolo* demonstrate that the type of craft may not be dispositive unless the craft actually interferes with the victim's use of her property. *Dow* does not rely on human observation, e.g. the surveillance can be via camera. *Riley* clarifies that this first factor, an extension of earlier lawful vantage point analyses, will be heavily influenced by FAA guidelines (and perhaps other applicable regulatory guidelines), and, potentially, property concepts insofar as the aerial surveillance inhibits the victim's use of the surveilled property. Also of note is that the regularity of overhead flights, a factor made relevant by *Riley*, appears to be an empirical burden placed on the victim of surveillance, not on the law enforcement agency (potentially creating a presumption that flights in navigable airspace are sufficiently regular). Thus, the Supreme Court's guidance, as derived from *Riley*, *Dow*, and *Ciraolo*, appears to be that aerial surveillance, whether by naked eye or assisted by some technology, conducted from lawful airspace that is traversed by aircraft with sufficient regularity, and that does not interfere with the victim's use of her property, is not a Fourth Amendment search and does not require a warrant.[142]

D. MODERN SURVEILLANCE CASES WITH UNMANNED AERIAL SURVEILLANCE IMPLICATIONS

In the twenty-three years since *Riley*, the Court has not commented specifically on the constitutionality of warrantless aerial surveillance despite the many questions left unanswered by *Riley*, *Dow*, and *Ciraolo*. For example,

141. *Id.* at 455.

142. *See id.* (majority opinion); California v. Ciraolo, 476 U.S. 207 (1986); Dow Chem. Co. v. United States, 476 U.S. 227 (1986).

by speaking only to aerial surveillance of curtilage, *Riley* did not opine on the constitutionality of aerial surveillance of the inside of a home. Additionally, while the aerial surveillance cases appear to allow the use of cameras, they do not explain the extent to which law enforcement agencies may use technology, such as infrared, electro-optical, or night-vision cameras, to enhance their observational powers without a warrant. Clues to the answers to these questions, however, can be found in three recent, notable cases: *Minnesota v. Carter*,[143] *Kyllo v. United States*,[144] and *United States v. Jones*.[145]

1. Peering into a Constitutionally Protected Zone: The Home—Minnesota v. Carter

As troubling as aerial surveillance of a home's curtilage can be, how much more unsettling would it be if law enforcement used an aerial vantage point to peer inside of the home? The Supreme Court has not directly answered that question, but the Court's 1998 decision in *Minnesota v. Carter* suggests that at least some warrantless surveillance of the inside of a home from an aerial vantage point may be constitutional.[146] In *Carter*, the Court evaluated whether temporary invitees into a residence had a reasonable expectation of privacy and whether a police officer's observation of the interior of a residence through a gap in a "drawn window blind" constituted a Fourth Amendment search.[147] Acting on a tip, a police officer peered into the ground-level window of a garden (basement) apartment from the vantage point of a "grassy area just outside the apartment's window" that was "frequently" used by the public for walking and playing.[148] "[T]hrough a gap in the closed blind," the officer observed Carter bagging a white, powdery substance.[149] Carter exited the apartment and was apprehended.[150] Police then returned to the apartment, executed a search warrant, and discovered that Carter had never been inside that apartment prior to that day and had been invited "for the sole purpose of packaging the cocaine" (an endeavor lasting approximately two and one-half hours).[151]

143. Minnesota v. Carter, 525 U.S. 83 (1998).
144. Kyllo v. United States, 533 U.S. 27 (2001).
145. United States v. Jones, 132 S.Ct. 945 (2012).
146. *See Carter*, 525 U.S. at 91.
147. *Id.* at 85, 91.
148. *Id.* at 85; *id.* at 103 (Breyer, J., concurring).
149. *Id.* at 85, 86 (majority opinion).
150. *See id.* at 85.
151. *Id.* at 86.

Delivering the opinion of the Court, Chief Justice Rehnquist rejected the Minnesota Supreme Court's use of "standing" doctrine to determine Carter's "legitimate expectation of privacy" and reiterated that the *Katz* test governs this area.[152] Further explaining that "in order to claim the protection of the Fourth Amendment," the Court's previous holding in *Rakas v. Illinois* required Carter to "demonstrate that he personally ha[d] an expectation of privacy in the place searched, and that his expectation [was] reasonable."[153] The Court next observed that although the "Fourth Amendment protects people, not places," the extent of Fourth Amendment protection "depends . . . upon whether the person who claims the protection of the Amendment has a legitimate expectation of privacy in the invaded place."[154] Deciding that the text of the Fourth Amendment only protects "people in 'their' houses," where "they" includes only residents and overnight guests, the Court held that Carter could not claim a Fourth Amendment violation because he was "merely present with the consent of the householder" to conduct "a business transaction."[155] Because Carter was found to be merely a non-overnight invitee (and the lessee of the apartment was not a party to the case), the Court did not decide whether the officer's observations of the interior of the apartment "constituted a 'search' " and would have violated the reasonable expectation of privacy of a resident or overnight guest.[156] The Court's interpretation of the Fourth Amendment as applicable only to "people in 'their' houses,"[157] however, strongly suggests that the Court would have found that the lessee had a reasonable expectation of privacy under the *Katz* test and therefore a Fourth Amendment search had occurred. Consequently, the Court would have had to determine whether or not the officer's warrantless search (observations through a gap in the closed blinds) was reasonable and hence constitutional.

Although the majority did not evaluate that question, Justice Breyer, in his individual concurrence, characterized the officer's observations as being "made 'from a public area outside the curtilage of the residence' " and concluded that although a Fourth Amendment search occurred, it was not unreasonable.[158] Analogizing to the aerial observations in *Riley* and *Ciraolo*,

152. *Id.* at 87–88 (citing Katz v. United States, 389 U.S. 347, 351 (1967)).
153. *Id.* at 88 (citing Rakas v. Illinois, 439 U.S. 128, 143–44 & n.12 (1978)).
154. *Id.* (quoting *Katz*, 389 U.S. at 351; *Rakas*, 439 U.S. at 143).
155. *Id.* at 89–91.
156. *Id.* at 86, 91.
157. *See id.* at 89.
158. *Id.* at 103–04 (Breyer, J., concurring). Note, however, that three years later, in *Kyllo v. United States*, Justice Breyer appears to have changed his mind about when a Fourth

Justice Breyer found that despite "[t]he precautions that the apartment's dwellers took to maintain their privacy" (e.g. closing the blinds of the ground-level apartment), "an ordinary passerby standing in" the same place as the officer (a "place used by the public") could have "see[n] through the window into the kitchen."[159] Perhaps signaling his opinion on a future surveillance case, Justice Breyer concluded his concurrence by opining that "there is a benefit to an officer's decision to confirm an informant's tip by observing the allegedly illegal activity from a public vantage point," such as "sav[ing] an innocent apartment dweller from a physically intrusive, though warrant-based, search if the constitutionally permissible observation revealed no illegal activity."[160]

In sum, the Court's holding and analysis in *Carter* suggests that law enforcement may warrantlessly use aerial surveillance to peer inside of a home if the fruits of the surveillance are used against non-residents and non-overnight guests. While the use of aerial surveillance of the inside of a home against a resident remains unsettled, Justice Breyer, at least, appears to prefer that law enforcement use aerial surveillance to confirm illegality over subjecting an innocent resident to a warrant-based, but intrusive, physical search of the house.[161] Even if a future majority of the Court were to disagree with Justice Breyer's preference, his concurrence reinforces the analogy and reasoning of *Riley* and *Ciraolo*, thereby providing a jurisprudential roadmap suggesting little difference between a lawful vantage point on the ground and one in the air, or between observations of a house's interior and its curtilage. Consequently, *Carter* offers the possibility that the constitutionality of aerial surveillance of a resident within her house could depend on the frequency and lawfulness of public access to the vantage point.

2. Limiting Surveillance Technology—Kyllo v. United States

As troubling as aerial surveillance with ordinary cameras can be, how much more unsettling would it be if law enforcement used specialized sensors to record details with telescopic magnification, film in darkness, or "see" though walls? The Supreme Court has not directly addressed the use of such sensors in the aerial surveillance context, but the Court's 2001 decision in *Kyllo v. United States* suggests there are limits on the power to use technology to warrantlessly enhance a law enforcement agency's

Amendment search occurs. 533 U.S. 27, 32 (2001) (joining the Kyllo majority opinion in characterizing visual observation without trespass as "no 'search' at all").

159. *Id.* at 104.

160. *Id.* at 105–06.

161. *Id.* at 103–06.

observational powers.[162] In *Kyllo*, the Court evaluated whether warrantless law enforcement use of a thermal-imaging device to detect relative amounts of heat within a private home constituted a Fourth Amendment search.[163] Suspecting that Kyllo was using high-intensity heat lamps to grow marijuana indoors, an agent of the U.S. Department of the Interior used a thermal imager[164] to scan the private home.[165] The scan lasted only a few minutes and was performed at 3:20 a.m. from a parked car on a public road in front of and behind Kyllo's house.[166] The scan revealed that "the roof over the garage and a side wall . . . were relatively hot compared to the rest of the home and substantially warmer than neighboring homes."[167] Using this and other information, the agent successfully sought a search warrant of Kyllo's home, found an indoor marijuana growing operation, and arrested Kyllo.[168]

Kyllo unsuccessfully moved to suppress the evidence, entered a conditional guilty plea, and appealed.[169] The Ninth Circuit remanded for an evidentiary hearing to determine "the intrusiveness of [the] thermal imaging."[170] The district court found that the thermal imager was a "non-intrusive device" because it only showed "a crude visual image of the heat being radiated from the outside of the house" and did not "show any people or activity within the walls of the structure, . . . penetrate walls or windows to reveal conversations or human activities," or reveal "any intimate details of the home."[171] A divided Ninth Circuit panel affirmed the denial of a motion to suppress, holding that Kyllo had not shown a "subjective expectation of privacy because he had made no attempt to conceal the heat escaping from his home" and that Kyllo had no objectively reasonable expectation of privacy because the imager only exposed "amorphous hot spots" and not "any intimate details of Kyllo's life."[172]

---

162. Kyllo v. U.S., 533 U.S. 27, 27–28 (2001).

163. *Id.* at 29.

164. "Thermal imagers detect infrared radiation, which virtually all objects emit but which is not visible to the naked eye. The imager converts radiation into images based on relative warmth—black is cool, white is hot, shades of gray connote relative differences; in that respect, it operates somewhat like a video camera showing heat images." *Id.* at 29–30.

165. *Id.* at 29.

166. *Id.* at 29–30.

167. *Id.*

168. *Id.*

169. *Id.* at 30.

170. *Id.*

171. *Id.* (internal quotations and citation omitted).

172. *Id.* at 30–31 (internal quotations and citations omitted).

Writing for a 5–4 majority, Justice Scalia began the Court's opinion by restating the core of the Fourth Amendment as "the right of a man to retreat into his own home and there be free from unreasonable government intrusion"[173] and that warrantless searches are presumptively unconstitutional.[174] The Court then admitted, however, that "the antecedent question whether or not a Fourth Amendment 'search' has occurred is not so simple under our precedent."[175] Attempting to explain the governing precedent, the Court (joined by Justice Breyer) specifically cited to Justice Breyer's concurrence in *Carter* as an example of an incorrect interpretation of precedent ("[o]ne might think that the new validating rationale would be that examining the portion of a house that is in plain public view, while it is a 'search' despite the absence of trespass, is not an 'unreasonable' one under the Fourth Amendment") and clarified that such a rationale is mistaken because "in fact we have held that visual observation is no 'search' at all."[176] Next, citing *Ciraolo* and *Riley*, among others, as examples of non-searches, the Court contrasted the instant case as "more than naked-eye" surveillance of a home.[177] The Court then differentiated the enhanced aerial photography in *Dow* from the instant case by noting that in *Dow*, it was "important that [it was] *not* an area immediately adjacent to a private home, where privacy expectations are most heightened."[178]

Recognizing that absent *Dow*, the Court had not previously evaluated "how much technological enhancement of ordinary perception . . . is too much," the Court acknowledged technological advances had affected the "degree of privacy" protected by the Fourth Amendment: "[f]or example, as [*Riley* and *Ciraolo*] make clear, the technology enabling human flight has exposed to public view (and hence, we have said, to official observation) uncovered portions of the house and its curtilage that once were private."[179] Before confronting "what limits there are upon th[e] power of technology to shrink the realm of guaranteed privacy," the Court criticized the *Katz* reasonable expectation test as "circular, . . . subjective[,] unpredictable," and difficult to apply to certain areas, such as automobiles, the curtilage, and

173. *Id.* at 31 (citing Silverman v. United States, 365 U.S. 505, 511 (1961)).

174. *Id.* (citing Illinois v. Rodriguez, 497 U.S. 177, 181 (1990); Payton v. New York, 445 U.S. 573, 586 (1980)); *see* Dow Chem. Co. v. United States, 476 U.S. 227, 234–35, 239 (1986).

175. *Kyllo*, 533 U.S. at 31.

176. *Id.* at 32.

177. *Id.* at 33.

178. *Id.* at 33 (quoting *Dow*, 476 U.S. at 237 n.4 (emphasis in original)).

179. *Id.* at 33–34 (internal citations omitted).

uncovered portions of a home.[180] In contrast to the difficult *Katz* analysis used for those areas, the Court explained that, at a minimum, a physical intrusion into the interior of the home would violate a reasonable expectation of privacy in the home.[181] Consequently, the Court offered a "ready criterion" to identify a search of the interior of a home: "obtaining by sense-enhancing technology any information regarding the interior of the home that could not otherwise have been obtained without physical 'intrusion into a constitutionally protected area' constitutes a search—at least where . . . the technology in question is not in general public use."[182]

During its analysis of potential rules, the Court specifically rejected the Government's and four-Justice dissent's proposed rule distinguishing "through-the-wall" from "off-the-wall" surveillance as being a "mechanical interpretation of the Fourth Amendment" rejected in *Katz*.[183] Claiming that such an approach would conclude, inter alia, that "a satellite capable of scanning from many miles away would pick up only visible light emanating from a house," the Court reasoned that such a rule "would leave the homeowner at the mercy of advancing technology."[184] Additionally, the Court rejected the Government's argument that the thermal imaging search was constitutional because it did not "detect private activities occurring in private areas."[185] Finding this "intimate details" argument based on *Dow* unpersuasive, the Court claimed "[t]he Fourth Amendment's protection of the home has never been tied to measurement of the quality or quantity of information obtained."[186] Instead, the Court believed that adoption of its new, firm, bright-line rule best assured the "preservation of that degree of privacy against government that existed when the Fourth Amendment was adopted."[187] Utilizing its new rule, the Court held that the thermal imager was not in "general public use," that it revealed information that would have "been unknowable without physical intrusion," and therefore that the warrantless surveillance was a presumptively unreasonable Fourth Amendment search.[188]

180. *Id.* at 37.
181. *Id.*
182. *Id.* at 34 (quoting Silverman v. United States, 365 U.S. 505, 512 (1961)).
183. *Id.* at 35–37; *id.* at 41–44 (Stevens, J., dissenting).
184. *Id.* at 35 (majority opinion).
185. *Id.* at 37.
186. *Id.*
187. *Id.* at 34.
188. *Id.* at 40.

In sum, the Court's holding and analysis suggests several limits to law enforcement use of advanced sensors to enhance warrantless aerial surveillance of the home. Most specifically, infrared cameras and thermal imaging may not be directed at a home, whether from the air or the ground, at least until such devices are in general public use. More generally, the use of any sense enhancing technology, such as radar or ultrasound technologies, which reveals details of the home otherwise unknowable without physical intrusion,[189] is presumptively unconstitutional. However, the use of visible light enhancing sensors against a home (e.g. high-magnification cameras and night-vision optics) is less clear. The Court's emphasis on physical intrusion, combined with dicta claiming visual observation (absent trespass) is no search at all, suggests that aerial surveillance, even using visual enhancing technologies, would be constitutional as long as conducted from a lawful (non-trespassory) vantage point. The Court's analysis, however, characterizes *Ciraolo* and *Riley* as "naked-eye"[190] surveillance of the home—a characterization suggesting that enhanced visual surveillance may be distinguishable from precedent. Additionally, the Court disparages a satellite's visible light scanning technology—essentially an extremely high magnification camera—as being similarly offensive as a thermal imager.[191] While aerial and satellite surveillance have many differences, certainly a similarity could be found in aerial use of high-magnification cameras. In any event, *Kyllo*'s most important implication for aerial surveillance is its "not in general public use" caveat. This implicit recognition that society's reasonable expectations of privacy change as technology spreads indicates that increasing public use of UAS will increase the potential constitutionality of warrantless law enforcement use.

3. Visual Observation Without Trespass is Not a "Search"—United States v. Jones

Aerial surveillance need not be directed solely at stationary targets, such as homes and their curtilage. A UAS's camera can be a powerful tool for tracking moving targets, but the Supreme Court has not directly addressed whether it is constitutional. The Court's 2012 decision in *United States v. Jones*, however, suggests that law enforcement agencies are free to use UASs to track suspects.[192] In *Jones*, the Supreme Court evaluated whether police

189. *Id.* at 36 (discussing several goals of police scientific research).
190. *Id.* at 33.
191. *Id.* at 35.
192. *See* United States v. Jones, 132 S. Ct. 945 (2012).

attachment of a Global-Positioning System ("GPS") device to a vehicle, and subsequent monitoring of the vehicle's movements on public streets, constituted a Fourth Amendment search.[193] Suspecting Jones of trafficking in narcotics, police, without a valid warrant,[194] attached a GPS device to the undercarriage of a vehicle used by Jones.[195] Police monitored Jones movements via the GPS device for the next twenty-eight days, ultimately indicting Jones.[196] During trial, the district court partially suppressed the GPS data (suppressing only the data from times when the vehicle was in Jones' garage), but the D.C. Circuit reversed the conviction after holding that warrantless use of the GPS device violated the Fourth Amendment.[197]

Delivering the opinion of the Court (a five-justice majority, with the remaining four justices concurring for different reasons), Justice Scalia continued the Court's resurgent use of property concepts to analyze Fourth Amendment searches.[198] Finding that the text of the Fourth Amendment "reflects its close connection to property," the Court traced property's role in search jurisprudence from the eighteenth century through today, but noted that *Katz* and its progeny "ha[d] deviated from that exclusively property-based approach."[199] Nonplussed by *Katz*, the Court declared that "Fourth Amendment rights do not rise or fall with the *Katz* formulation" and "the *Katz* reasonable-expectation-of-privacy test has been *added to*, not *substituted for*, the common-law trespassory test."[200] Thus, the Court concluded that, as a physical intrusion on private property, the police installation of the GPS device here "would have been considered a 'search' within the meaning of the Fourth Amendment when it was adopted" and was therefore unconstitutional absent a warrant.[201]

Evidently believing its use of *Katz* as a supplement to a property-based analysis would "assur[e] preservation of that degree of privacy against government that existed when the Fourth Amendment was adopted,"[202] the majority criticized the concurrence's exclusive reliance on *Katz* analysis,

193. *Id.* at 948–49.

194. Police obtained a warrant authorizing installation of the GPS device in D.C. and within ten days of the warrant's issuance. Installation occurred, however, in Maryland on the eleventh day. *Id.*

195. *Id.*

196. *Id.*

197. *Id.* at 948–49.

198. *See id.* at 949–53.

199. *Id.* at 949–50.

200. *Id.* at 950, 952.

201. *Id.* at 949.

202. *Id.* at 950 (quoting Kyllo v. United States, 533 U.S. 27, 31 (2001)).

despite the concurrence reaching the same result.[203] Specifically, the *Jones* majority explained that while its property-based rule governed trespass situations, situations "without trespass would *remain* subject to *Katz* analysis."[204] This clarification of the extent of the Court's retrenchment to a property-based analysis specifically affects visual, and therefore aerial, surveillance. In the context of non-trespassory situations, the Court reiterated that, to date, the Court had not "deviated from the understanding that mere visual observation does not constitute a search."[205] The Court continued to explain that because "[a] person traveling in an automobile on public thoroughfares has no reasonable expectation of privacy in his movements from one place to another,"[206] "traditional surveillance" of that person, which could include "aerial assistance," is a constitutional visual observation, not a Fourth Amendment search.[207]

In sum, the Court's holding and analysis in *Jones* provides that law enforcement agencies can use aerial surveillance to track a suspect's public movements. Perhaps more importantly, however, the dicta reaffirms that warrantless visual surveillance (including aerial surveillance), absent trespass, is not a Fourth Amendment search. Although the *Jones* court does not speculate on what form an aerial surveillance trespass may take, aerial surveillance conducted from a lawful vantage point, if consistent with *Riley* and *Ciraolo* (e.g. sufficiently routine and within FAA guidelines), almost certainly remains constitutional.

E. FAA REGULATIONS GOVERNING UAS

Any future evaluation of the constitutionality of UAS surveillance, just like any other visual surveillance, will almost surely require determining the lawfulness of the vantage point. Given *Riley* and *Ciraolo*'s references to FAA guidelines in making this determination,[208] current and future FAA UAS guidelines are highly likely to play a vital, and perhaps dispositive, role in any future ruling. The FAA requires special permits for UAS operations in the national airspace system[209] because UASs are not compliant with

203. *Id.* at 953–54; *id.* at 957–64 (Alito, J., concurring).

204. *Id.* at 953 (majority opinion) (emphasis in original).

205. *Id.* (citing *Kyllo*, 533 U.S. at 31–32).

206. *Id.* (citing United States v. Knotts, 460 U.S. 276, 281 (1983)).

207. *Id.* at 953–54.

208. *See* Florida v. Riley, 488 U.S. 445, 449–51 (1989); California v. Ciraolo, 476 U.S. 207, 209 (1986).

209. *See* FAA, INTERIM OPERATIONAL APPROVAL GUIDANCE 08-01, UNMANNED AIRCRAFT SYSTEMS OPERATIONS IN THE U.S. NATIONAL AIRSPACE SYSTEM 2 (Mar. 13, 2008).

fundamental safety requirements, such as an onboard pilot to see and avoid other aircraft.[210] These permit applications make no mention of privacy concerns, instead requesting only that the applicant "demonstrate that injury to persons or property along the flight path is extremely improbable."[211]

Despite dozens of law enforcement agencies successfully filing for UAS permits,[212] the pace of UAS integration into the national airspace has apparently been too slow for Congress, which passed the FAA Modernization Act ("FAAMA") of 2012.[213] Signed by President Obama in February 2012, the FAAMA requires, among other things, the FAA to increase the pace of integration of drone flights in U.S. airspace.[214] In response to law enforcement agencies "mistakenly" operating their small UASs under FAA advisory circular 91-57, *Model Aircraft Operating Standards*,[215] the order also created a "Small [UAS] Aviation Rulemaking Committee" and tasked it with creating and promulgating relevant regulations.[216] The FAA has not yet completed this task, but the FAAMA requires the FAA "to publish a final rule governing small UAS[s] in the *Federal Register* by August 2014."[217]

Of note is that the FAAMA focuses on the "safe and routine operation of civil [UASs] in the national airspace system."[218] The FAAMA's sole mention of "privacy" occurs in the context of air passenger screening,[219] not civil UAS operations. Indeed, the Government Accountability Office ("GAO"), Congress' research arm, in its report on the FAAMA, discusses potential privacy implications and claims that "FAA officials and others have suggested that regulating privacy issues in connection with equipment carried on UAS, such as surveillance sensors that do not affect safety, is outside FAA's mission, which is primarily focused on aviation safety."[220] Consequently, the FAAMA and any resulting future FAA regulations for law enforcement UAS operations are unlikely to address privacy or Fourth Amendment implications. Thus, Justice O'Connor's criticism of the Court's

210. *See, e.g.*, Right-of-Way Rules: Except Water Operations, 14 CFR 91.113(b) (2013) (containing "see-and-avoid" provision).
211. *Id.* at 8.
212. *See* Lynch, *supra* note 14.
213. FAAMA, *supra* note 7.
214. *See id.* § 332 (codified as 49 U.S.C. § 40101 et seq.).
215. FAA, *supra* note 38.
216. *Id.*
217. GAO 2012 UAS REPORT, *supra* note 8, at 28.
218. FAAMA, *supra* note 7, § 332(B).
219. *Id.* § 826.
220. GAO 2012 UAS REPORT, *supra* note 8, at 36.

use of FAA regulations in *Riley*[221] remains relevant to evaluations of UAS operations twenty-three years later.

F. OTHER LEGISLATION GOVERNING UAS

In addition to FAA regulations, law enforcement UAS operations may be affected by other federal and state legislation. Unlike FAA regulations, which are primarily focused on safety regulations, legislatures can directly protect citizens' privacy by proscribing certain law enforcement uses of UAS. At the moment, such legislation appears to currently enjoy rather broad bipartisan support at both the state[222] and federal levels.[223] Indeed, the 2012 Republican Party platform included a plank supporting a restriction on domestic law enforcement UAS surveillance.[224]

State legislatures have not been idle. On February 6, 2013, by overwhelming bipartisan votes, both houses of the Virginia legislature approved a two-year moratorium on law enforcement UAS use within the state.[225] Since that time, forty-two states have proposed legislation limiting law enforcement use of UAS for surveillance and eight of those states have enacted such legislation.[226] Florida's enacted bill, for example, is entitled the "Freedom from Unwarranted Surveillance Act" and, as of July 1, 2013, protects citizens from "privacy-invasive technology" by requiring a judicial

221. Florida v. Riley, 488 U.S. 445, 452 (O'Connor, J., concurring).

222. *See* Catherine Crump & Jay Stanley, *Why Americans Are Saying No to Domestic Drones*, SLATE (Feb. 11, 2013), http://www.slate.com/articles/technology/future_tense/2013/02/domestic_surveillance_drone_bans_are_sweeping_the_nation.html (describing how "states and cities are increasingly passing legislation to stop the use of surveillance drones").

223. *See* Sandra Fulton, *Even Amidst a Host of Congressional Priorities, Drones Makes the Cut*, AMERICAN CIVIL LIBERTIES UNION (Mar. 20, 2013), https://www.aclu.org/blog/technology-and-liberty-criminal-law-reform/even-amidst-host-congressional-priorities-drones (describing a "rare show of bipartisanship" where members of Congress from both sides of the aisle have raised privacy concerns over drone use).

224. GOP, 2012 Republic Platform, *available at* http://www.gop.com/2012-republican-platform_We/#Item11 (last visited Feb. 28, 2013) ("The Fourth Amendment: Liberty and Privacy: Affirming 'the right of the people to be secure in their houses, papers, and effects, against unreasonable searches and seizures,' we support pending legislation to prevent unwarranted or unreasonable governmental intrusion through the use of aerial surveillance or flyovers on U.S. soil, with the exception of patrolling our national borders. All security measures and police actions should be viewed through the lens of the Fourth Amendment; for if we trade liberty for security, we shall have neither.").

225. ACLU, *Virginia House of Delegates and Senate Approve Two Year Moratorium on Drones*, ACLU.ORG (Feb. 6, 2013), http://www.aclu.org/criminal-law-reform/virginia-house-delegates-and-senate-approve-two-year-moratorium-drones.

226. ACLU, *Status of Domestic Drone Legislation in the States*, ACLU.ORG, https://www.aclu.org/blog/technology-and-liberty/status-domestic-drone-legislation-states (last updated Aug. 28, 2013).

warrant supported by probable cause before law enforcement may use UAS for surveillance.[227] Not every state's legislature, however, has proposed or enacted protections as broad as those of Florida: Idaho's enacted bill, for example, includes a warrant exception for "controlled substance investigations."[228] Although a patchwork of state laws could protect many Americans from warrantless UAS surveillance, only federal legislation can provide a consistent floor of privacy protection to all Americans.

Towards that goal, in 2012, both Democrat and Republican congresspeople proposed various bills restricting law enforcement use of UAS for surveillance.[229] More recently, on February 13, 2013, Representative Ted Poe introduced a House bill, the "Preserving American Privacy Act of 2013," which purports to generally require a judicial warrant for law enforcement UAS surveillance.[230]

Despite the proclaimed privacy-protection intent of such legislation, the exact effects of proposed legislation remain unclear. For example, Representative Poe's bill would allow law enforcement to use UASs to gather "covered information" for use as evidence only if pursuant to either a judicial warrant, presumably supported by probable cause, or a judicial "order," based merely on "specific and articulable facts showing a reasonable suspicion of criminal activity and a reasonable probability that the operation of a [UAS] will provide evidence of such criminal activity."[231] Law enforcement would also have the ability to seek a renewable judicial order authorizing UAS surveillance over a "stipulated public area" (such as a public park) for up to forty-eight hours, but the bill does not provide any standard by which the judicial officer is to base this order upon.[232] Perhaps most notably, the bill defines "covered information" as "(A) information that is reasonably likely to enable identification of an individual; or (B) information

227. 2013 Fla. Laws 2013-0092 c1, *available at* http://flsenate.gov/Session/Bill/2013/0092/BillText/er/HTML.

228. 2013 Idaho Sess. Laws SB 1134 Engrossment 2, § 21-213, 2(a), 27–28, *available at* http://www.legislature.idaho.gov/legislation/2013/S1134E2.pdf.

229. Andrea Stone, *Drone Privacy Bill Would Put In Safeguards On Surveillance*, HUFFINGTON POST (Aug. 1, 2012), http://www.huffingtonpost.com/2012/08/01/drone-privacy-bill_n_1728109.html.

230. Preserving American Privacy Act of 2013, H.R. 637, 113th Cong. (2013). As of Sept. 11, 2013, H.R. 637 has spent nearly seven months languishing in the congressional committee to which it was assigned. GOVTRACK.US, http://www.govtrack.us/congress/bills/113/hr637.

231. H.R. 637 § 3119c.

232. *Id.*

about an individual's property that is not in *plain view*."[233] "Plain view," however, is already the line separating constitutional and unconstitutional warrantless surveillance.[234]

Regardless of their actual effects, the "other legislation" described above attempts to protect privacy by directly proscribing certain law enforcement activities. FAA regulations, by contrast, indirectly affect privacy through constitutional implications because they define the lawfulness of the vantage point.[235] This "other legislation," then, operates not as a guide to the potential constitutionality of warrantless UAS surveillance, but rather as an independent and alternative method of protecting privacy should constitutional arguments fail.

IV. LAW ENFORCEMENT UAS SURVEILLANCE WOULD LIKELY BE FOUND CONSTITUTIONAL IN A NUMBER OF SITUATIONS

Those concerned that increasing law enforcement UAS operations erode personal privacy at the expense of the Fourth Amendment do, indeed, have much to worry about. Current Supreme Court jurisprudence, although not directly addressing the Fourth Amendment implications of UAS surveillance, suggests that law enforcement UAS use, with certain limits, would be found constitutional if challenged under the Fourth Amendment. Because warrantless Fourth Amendment searches are presumptively invalid, the challenger will likely only need to convince the Court that a search occurred; that, however, is no easy feat. For a search to occur, *Katz* requires that the person had both a subjective and an objectively reasonable expectation of privacy. *Kyllo* and *Jones* both explain that a trespassory search or one that would have required intrusion but for the use of advanced technology necessarily satisfy the *Katz* test. *Jones* and *Kyllo*, however, have strongly reaffirmed the concept that visual surveillance from a lawful vantage point is not a search at all because the object of surveillance was in "plain view." Citing to *Ciraolo*, *Dow*, and *Riley*, both *Kyllo* and *Jones* made no distinction between traditional aerial surveillance and on-the-ground surveillance. The rule appears clear: as long as the vantage point is lawful, visual surveillance is not a search. Thus, those wishing for warrantless UAS surveillance to be declared an unconstitutional Fourth Amendment search must indirectly

233. *Id.* § 3119a (emphasis added).
234. *See supra* Sections III.B–D (discussing Fourth Amendment Privacy jurisprudence).
235. *See supra* Section III.E (discussing FAA Regulations).

attack the surveillance with one of three main litigation strategies: (1) establishing that the victim was in a constitutionally protected zone, (2) attacking the legality of the vantage point from which the UAS surveilled, or (3) attacking the specific UAS technology used.

A. ESTABLISHING A CONSTITUTIONALLY PROTECTED ZONE

While the Fourth Amendment protects people and not places,[236] the place where people are when visually surveilled unmistakably matters. *Jones* makes clear that in public, a person has no reasonable expectation of privacy from visual surveillance.[237] The Supreme Court has long held that "an expectation of privacy in commercial premises . . . is different from, and indeed less than, a similar expectation in an individual's home,"[238] so seeking refuge in a commercial premise affords little, if any, protection from visual surveillance. *Carter*, meanwhile, explains that a person has no reasonable expectation of privacy in *another person's* home unless invited as an overnight guest.[239] This leaves only a person's *own* home, but privacy there, too, is not absolute. *Riley* and *Ciraolo*, as reiterated in *Kyllo*, make clear that flight technology has "uncovered portions of the house and its curtilage that once were private."[240]

It is unlikely that the Court would find that there is a reasonable expectation of privacy for any activity in the home that can be observed from a public (e.g. non-trespassory and lawful) vantage point. Justice Brennan, in his dissent in *Riley*, even noted that there was nothing in the Court's opinion to suggest that law enforcement would conduct a Fourth Amendment search if they looked, from their helicopter, through an open window into a room viewable only from the air.[241] This view is supported by the Court's emphasis in *Kyllo* that trespass is the basis for evaluating surveillance of the home and reiteration that visual surveillance is not a search.[242] Moreover, *Riley*, *Ciraolo*, and Justice Breyer's concurrence in *Carter* each found anti-observation countermeasures (respectively: building two fences, partially covering the roof of a greenhouse, and incompletely closing window blinds) insufficient to render law enforcement observation a Fourth Amendment search because

236. Katz v. United States, 389 U.S. 347, 351 (1967).
237. *See supra* Section III.D.3 (discussing *United States v. Jones*).
238. New York v. Burger, 482 U.S. 691, 700 (1987).
239. *See supra* Section III.D.1 (discussing *Minnesota v. Carter*).
240. Kyllo v. United States, 533 U.S. 27, 33–34 (2001).
241. Florida v. Riley, 488 U.S. 445, 462 (1989) (Brennan, J., dissenting).
242. *Kyllo*, 533 U.S. at 32.

the *Katz* test evaluates a person's subjective expectation of privacy differently than society's objective reasonable expectation.

The precedent in this area may not control because the Court could find some conceptual difference between manned aerial visual observations and UAS visual observations, even if made with a standard camera. The Court in *Riley* cautioned that it was *not* saying "that an inspection of the curtilage of a house from an aircraft will always pass muster under the Fourth Amendment simply because the plane is within the navigable airspace specified by law."[243] It may be that the Court would determine that a search occurs if a UAS observes the interior of a home through a closed, but not completely opaque, skylight, even if it is surveilling from a lawful vantage point. UAS observation through an above-ground-floor window at an angle impossible from a manned aircraft may offend the Court's interpretation of the Fourth Amendment. As Justice Alito succinctly stated in his concurrence in *Jones*, "technology can change [privacy] expectations."[244] While technology has tended to erode reasonable expectations of privacy, and the existing case law suggests that visual observations from lawful vantage points are not searches, *Kyllo* made clear that some technology can sufficiently offend a reasonable expectation of privacy.

B. ATTACKING THE LAWFULNESS OF A UAS SURVEILLANCE VANTAGE POINT

Riley and *Ciraolo* establish the lawfulness of a vantage point as critical to whether or not a Fourth Amendment search occurred. *Kyllo* and *Jones*'s return of property concepts to the equation further confirmed that visual surveillance must not be trespassory. A UAS surveilling from the same vantage point as *Ciraolo*'s fixed-wing craft or *Riley*'s helicopter would likely be considered to be similarly lawful. However, it is easy to imagine (and perhaps expect) that law enforcement agencies will deploy UAS to surveil from vantage points impossible in conventional aircraft simply because UASs can provide the capability to do so.

In his dissent in *Riley*, Justice Brennan could have been describing current UAS technology and FAA regulations when he asked the reader to imagine "a helicopter capable of hovering just above an enclosed courtyard or patio without generating any noise, wind, or dust at all—and, for good measure, without posing any threat of injury."[245] Law enforcement use under then-

243. *Id.* at 451 (majority opinion).
244. United States v. Jones, 132 S. Ct. 945, 962 (2012) (Alito, J., concurring).
245. *Riley*, 488 U.S. at 462–63 (Brennan, J., dissenting).

controlling FAA regulations meant that the vantage point was lawful and therefore observations from that vantage point, even into the house, were not a Fourth Amendment search.[246] Current FAA regulations, which still emphasize safety over privacy, are not materially different from those in *Riley*. Even assuming the presence of the UAS did not interfere with the use of the home (an important factor in *Riley* and the FAA regulations), under *Riley*, it would appear that this type of law enforcement UAS surveillance remains constitutional.

In contrast, *Kyllo* and *Jones*'s emphasis on a property-based approach suggests that the Court might entertain the argument that the physical presence of the UAS over the property at sufficiently low heights renders it a trespass, and, therefore, a Fourth Amendment search. *United States v. Causby* may have abolished the common-law "ancient doctrine . . . that ownership of the land extend[s] to the periphery of the universe,"[247] but the Court could find that a resident retains some right of exclusion from the airspace above her property that is not usable by conventional craft. It is difficult to imagine that the airspace between a house's curtilage and the height of its roof is a "public highway" such that granting the property owner some right of exclusion "seriously interfere[s] with [its] control and development in the public interest."[248]

It is also possible that UAS surveillance from a non-trespassory, lawful vantage point could be found to be sufficiently intrusive. Imagine that law enforcement officers, like those in *Carter*, wish to peer inside of an apartment several stories above ground to confirm a tip that criminal activity was occurring. Unable to walk up to the window and peer in, the officers warrantlessly launch a small UAS from the street below. The UAS quickly ascends to the height of the apartment's window and relays video of the activities within to the officers on the ground. Assuming the UAS flight was within FAA guidelines and that no trespass or physical intrusion into the apartment occurred, the vantage point was clearly lawful and appears constitutional under existing precedent. Once FAA regulations open the skies to low-altitude UAS flight, such a holding effectively creates an Orwellian state: using UAS, law enforcement could surveil from the position of any possible viewer at any possible vantage point.

246. *Id.*

247. *See* United States v. Causby, 328 U.S. 256, 260 (1946) (internal quotations and citations omitted).

248. *Id.* at 261.

If the Court wishes to avoid such Orwellian effects, it may be persuaded by a *Kyllo*-based argument. More specifically, because the activity within the apartment could only have been seen via the angle of the UAS or through an actual physical law enforcement intrusion into the apartment, the Court could use *Kyllo*'s rule to interpret that a Fourth Amendment search occurred as long as the Court first finds that UASs are not within "general public use."[249] However, if such an angle of observation could have been achieved from an adjacent building, regardless of police law enforcement agent's actual ability to physically observe from that location, it is unlikely that *Kyllo* would be found applicable because the general public could have been afforded the same view.

Alternatively, the Court may choose to differentiate the hypothetical above from *Riley* and *Ciraolo* because a traditional aircraft would have been incapable of observing from the angle the UAS observed from. This potential differentiation underscores Justice O'Connor's criticism of judicial reliance on FAA regulations for a determination of the lawfulness of a vantage point. A future Court could find grounds to declare that the lawfulness of a vantage point is not determined merely by compliance with FAA regulations because the FAA is primarily concerned with safety and not privacy or Fourth Amendment protections, as highlighted by its recent mandate from Congress in the FAAMA.[250] In such a case, law enforcement attempting to introduce evidence so obtained may need to heed Justice O'Connor's suggestion and demonstrate that UAS flights at that altitude are routine enough to prevent an objectively reasonable expectation of privacy. Unfortunately this argument for the unconstitutionality of certain UAS surveillance may expire soon: the FAAMA's stated goal is to rapidly increase the routine private and governmental use of UAS in the United States.[251]

C. ATTACKING THE SPECIFIC UAS TECHNOLOGY USED

The third type of potential attack on the constitutionality of UAS visual surveillance is predicated on *Kyllo*'s "general public use" criteria for advanced surveillance technologies. A challenger of law enforcement UAS surveillance could attempt to assert that because the "general public" does not use UASs either specifically for surveillance or generally for any use, law enforcement's use of UASs should be held unconstitutional. Because the Court carefully crafted the *Kyllo* rule to establish a bright line at surveillance of the home, this

249. Kyllo v. United States, 533 U.S. 27, 34 (2001).
250. FAAMA, *supra* note 7, § 332(B).
251. FAAMA, *supra* note 7.

argument is unlikely to be persuasive against UAS surveillance outside the home. Even when applied to UAS surveillance of the home, however, the window of opportunity to challenge UAS surveillance is closing rapidly. Even if a challenger to law enforcement UAS surveillance could adduce sufficient evidence demonstrating the general public does not use UASs right now, private use of UASs is expected to grow.[252] Moreover, Congress, via the FAAMA, has declared its goal to increase private UAS operations until they are routine.[253]

If a general challenge to law enforcement UAS surveillance is destined to fail, perhaps opponents could successfully challenge law enforcement use of specific optics. *Kyllo* has already specifically held that infrared cameras (i.e., thermal imaging) are not sufficiently used by the public. This determination is of course subject to change as the technology proliferates to the public. For now, at least, law enforcement cannot warrantlessly use infrared cameras mounted on a UAS for surveillance purposes. Indeed, any future sensor capable of intruding into the home and not in general public use (perhaps radar and sonar based technologies), would be found unconstitutional under *Kyllo* regardless of whether it was mounted on a UAS or deployed on the ground.

The use of advanced electro-optical cameras allowing many thousands times magnification[254] may also be foreclosed by a lack of general public use. Indeed, *Kyllo*'s dicta mentions satellite scanning technologies based upon advanced electro-optical cameras, but is unclear on what their import is. Perhaps it is a hint that the Court would entertain an argument on the constitutionality of similar cameras mounted on UASs? If so, the Court could also use the naked eye of *Ciraolo* and *Riley* and the "common" aerial camera in *Dow* (as well as its specific fact of industrial, not residential, curtilage) as points of differentiation. In any case, if law enforcement agencies take care to only equip their UAS with cameras featuring abilities readily available to the public, then warrantless law enforcement UAS surveillance is very unlikely to run afoul of *Kyllo* and will remain constitutional.

252. GAO 2012 UAS REPORT, *supra* note 8, at 5.

253. FAAMA, *supra* note 7.

254. For example, DARPA has recently announced development of a 1.8 gigapixel camera "that is supposedly the highest-resolution surveillance system in the world." Nicole Lee, *DARPA's 1.8-gigapixel cam touts surveillance from 20,000 feet* (video), ENGADGET.COM (Jan. 28, 2013), http://www.engadget.com/2013/01/28/darpa-argus-is-surveillance. The electro-optical camera is described as being able to "spot a six-inch object within a ten square mile radius from 20,000 feet in the air." *Id.*

V. CONCLUSION

Unmanned aircraft systems, once exclusively the province of the military, are rapidly spreading to civilian use. Federal, state, and local law enforcement agencies have discovered that UASs are a relatively inexpensive and highly effective augmentation of their airborne enforcement arsenals. Consequently, many law enforcement agencies are taking advantage of UAS diversity and configurability to expand their aerial surveillance efforts. These efforts, however, raise many Fourth Amendment implications for the privacy of those surveilled. Many people would like the Supreme Court to find UAS surveillance unconstitutional, but current Fourth Amendment conventional aerial-surveillance jurisprudence appears likely to extend to UAS surveillance operations. If so, then the Fourth Amendment may offer little privacy protection because most surveillance applications do not actually qualify as Fourth Amendment searches in the first place. Regardless, dicta in current precedent raise the possibility that the Court will not turn a deaf ear towards all constitutional challenges of law enforcement UAS surveillance. Specifically, the Court may be persuaded by arguments based on the heightened expectations of privacy in constitutionally-protected zones, the vantage point from which a UAS surveilled, or the specific technology used. In case the Supreme Court is not so persuaded, those worried about losing privacy to warrantless law enforcement UAS surveillance should continue to seek legislation proscribing such conduct. Such legislation may not affect the constitutionality of warrantless law enforcement UAS surveillance, but it can certainly protect Americans' privacy. As the deployment and operation of UASs accelerates, so too should the debate over the accompanying privacy implications.

REGULATING DISRUPTIVE INNOVATION

Nathan Cortez[†]

ABSTRACT

Disruption theory tells us that certain innovations can undermine existing products, firms, or even entire industries. Classic examples include the Kodak camera, the Bell telephone, and the Ford Model T. Modern examples abound. The market entrant's innovation ultimately displaces industry incumbents. Regulators, too, are challenged by such disruptive innovations. The new product, technology, or business practice may fall within an agency's jurisdiction but not square well with the agency's existing regulatory framework. Call this "regulatory disruption."

Most scholars intuit that regulators should be cautious rather than firm in such situations. Tim Wu, in *Agency Threats*, argues that agencies confronting disruptive innovations should avoid traditional rulemaking and adjudication, and instead rely on "threats" packaged in guidance documents, warning letters, and the like. Threats, he argues, are less burdensome, more flexible, and avoid regulation that is miscalibrated or premature. However, this Article argues that a flexible initial posture based primarily on "threats" can calcify, creating weak defaults that lead to suboptimal regulation in the long term. Regulatory inertia can be hard to break without an external shock, usually a tragedy or some other massive failure that reignites interest in regulation. As a case study, this Article shows how the FDA's approach to a disruptive technology (computerized medical devices) twenty-five years ago fits the threat framework strikingly well, and how it failed. The FDA's threats became stale and counterproductive—during a profound computer revolution, no less. This Article counterposes the FDA's approach to software with the FCC's approach to the Internet, which initially relied on threats, but later codified them via binding regulations and enforcement shortly thereafter.

This Article argues that agencies need not be so tentative with innovations. If agencies are concerned about regulating prematurely or in error, then they can experiment with timing rules, alternative enforcement mechanisms, and other variations on traditional interventions. If agencies do choose to proceed by making threats, then they should use them as a short-term precursor to more decisive, legally binding action, as the FCC did, and avoid relying on them as a long-term crutch, as the FDA did.

© 2014 Nathan Cortez.

† Nathan Cortez, Associate Professor and Associate Dean for Research, Southern Methodist University ("SMU"), Dedman School of Law. J.D., Stanford University, B.A., University of Pennsylvania. Many thanks to Nicholas Bagley, Jeffrey Bellin, Glenn Cohen, Lars Noah, Jordan Paradise, Frank Pasquale, Nicolas Terry, and David Zaring for commenting on earlier drafts. I also received valuable comments and questions from participants at the American Society of Law, Medicine, and Ethics ("ASLME") annual meeting, the SMU Law School Faculty Forum, and at the Harvard Law School's Petrie-Flom Center conference on "The Food and Drug Administration in the 21st Century" (May 4, 2013). I thank Courtney Eudy, Banee Pachuca, and Robert Thetford for their excellent research. An earlier version of this Article was selected for the "New Voices in Administrative Law" program at the 2013 AALS Annual Meeting (Jan. 5, 2013). Finally, this research was supported by generous funding from the A.J. and Ann Van Wynen Thomas Memorial Endowed Research Award.

TABLE OF CONTENTS

I. **INTRODUCTION** 176

II. **DISRUPTIVE INNOVATION** 182

III. **EMPTY THREATS?** 187

A. AGENCY THREATS 187
B. THE FDA AND THE QUARTER-CENTURY-LONG THREAT 191
C. THE FCC AND THE SIX-YEAR-LONG THREAT 196

IV. **THE REGULATORY TOOLKIT FOR DISRUPTIVE INNOVATIONS** 199

A. TIMING 201
B. FORM 206
1. Rulemaking 207
2. Adjudication 208
3. Guidances 209
4. New Forms? 213
5. Does Form Matter? 214
C. DURABILITY 217
1. Sunsets 218
2. Deadlines 220
D. ENFORCEMENT 221

V. **CONCLUSION** 227

I. INTRODUCTION

A persistent challenge for regulators is confronting new technologies or business practices that do not square well with existing regulatory frameworks. These innovations can depart in important ways from older incumbents. For example, the innovation might present unanticipated benefits and risks. It might disturb carefully crafted equilibria between regulators, industry, and consumers. The innovation might puncture prevailing regulatory orthodoxies, forcing regulators to reorient their postures or even rethink their underlying statutory authority. The quintessential example is the Internet, which rumpled not just one, but several regulatory frameworks, including those of the Federal Communications Commission ("FCC"), the Federal Trade Commission ("FTC"), and the Food and Drug Administration ("FDA").[1]

1. For the FCC's experience adapting its longstanding regulatory framework to Internet providers, see *infra* Section III.C. For the FTC's experience adapting its longstanding frameworks for privacy, marketing fraud, and antitrust violations to online

Periodically, our economy generates these "disruptive innovations."[2] Joseph Bower and Clayton Christensen introduced the idea to describe new technologies that undermine and eventually displace established products, firms, or even entire industries. Iconic examples include automobiles, personal computing, and cellular phones. But the idea's explanatory power extends to many other products and industries. As a result, disruption theory has inspired prolific writing in the business academy and now creeps into other disciplines.[3]

Here, this Article uses the idea to refer to innovations that disrupt existing regulatory schemes, not necessarily industry incumbents (though they may do that too). Call this "regulatory disruption." Legal scholars have examined disruptive innovation in various disciplines, such as civil

activities, see, for example, Babette E.L. Boliek, *FCC Regulation Versus Antitrust: How Net Neutrality Is Defining the Boundaries*, 52 B.C. L. REV. 1627, 1652–82 (2011) (comparing the suitability of FCC regulation versus FTC antitrust jurisdiction in crafting net neutrality policies); Steven Hetcher, *The FTC as Internet Privacy Norm Entrepreneur*, 53 VAND. L. REV. 2041, 2053 (2000) (using public choice theory to explain the FTC's approach to online privacy); Jay P. Kesan & Andres A. Gallo, *Optimizing Regulation of Electronic Commerce*, 72 U. CIN. L. REV. 1497, 1605–22 (2004) (considering in detail the FTC's regulation of online consumer fraud from a microeconomic and game-theoretical perspective). For an amusing juxtaposition that highlights the friction between new technologies and older regulatory frameworks, see Rep. George W. Gekas & James W. Harper, *Annual Regulation of Business Focus: Regulation of Electronic Commerce*, 51 ADMIN. L. REV. 769, 772, 777 (1999) (juxtaposing the heading "The Internet: Everything's New!" with the heading "Government: Everything's Old!"). For the FDA's early experience adapting its longstanding framework for regulating pharmaceutical labeling and advertising to new online promotional practices, see Peter S. Reichertz, *Legal Issues Concerning the Promotion of Pharmaceutical Products on the Internet to Consumers*, 51 FOOD & DRUG L.J. 355, 357–61 (1996); Marc J. Scheineson, *Legal Overview of Likely FDA Regulation of Internet Promotion*, 51 FOOD & DRUG L.J. 697 (1996).

2. *See* Joseph L. Bower & Clayton M. Christensen, *Disruptive Technologies: Catching the Wave*, HARV. BUS. REV., Jan.–Feb. 1995, at 43, 45 (1995). Bower and Christensen did not discuss the term "disruptive innovation" in their 1995 article, but Christensen's follow-up book, CLAYTON M. CHRISTENSEN, THE INNOVATOR'S DILEMMA: WHEN NEW TECHNOLOGIES CAUSE GREAT FIRMS TO FAIL (1997), helped introduce and popularize the term.

3. For example, Christensen and colleagues have applied disruption theory to health care. *See generally* CLAYTON M. CHRISTENSEN ET AL., THE INNOVATOR'S PRESCRIPTION: A DISRUPTIVE SOLUTION FOR HEALTH CARE (2009). An early application of disruption theory to health care in the legal literature is Lesley H. Curtis & Kevin A. Schulman, *Overregulation of Health Care: Musings on Disruptive Innovation Theory*, 68 L. & CONTEMP. PROBS. 195 (2006). For a thoughtful analysis of why health information technologies have yet to disrupt the U.S. health care system, per Christensen's criteria, see Nicolas P. Terry, *Information Technology's Failure to Disrupt Healthcare*, 13 NEV. L.J. 722 (2013). For a skeptical review of *The Innovator's Prescription*, see J.D. Klienke, *Perfection in PowerPoint*, 28 HEALTH AFF. 1223 (2009) ("Although the book excavates several fascinating nuggets about wrenching changes in other industries, it attempts to force each into Christensen's franchise-in-progress, 'disruption theory.' ").

procedure,[4] environmental law,[5] and intellectual property.[6] Health law scholarship predominates the legal literature on disruptive innovation,[7] perhaps because our dysfunctional health care system begs for transformation.[8] And, to be sure, there is no shortage of legal scholarship examining technological innovation more generally.[9] But no one has yet applied disruption theory to the field in which it should be most useful—administrative law. Administrative law scholarship contains a rich, sprawling discourse on how agencies should regulate new markets.[10] These questions, incidentally, can be particularly vexing with categorically novel technologies and business practices. Contemporary debates in administrative law thus map well onto disruption theory, and for that reason it is worth merging the two here.

The one scholar that comes closest to merging disruption theory and administrative law theory is Tim Wu, in his recent essay *Agency Threats*.[11] Wu

4. *See, e.g.*, Cassandra Burke Robertson, *The Facebook Disruption: How Social Media May Transform Civil Litigation and Facilitate Access to Justice*, 65 ARK. L. REV. 75 (2012).

5. *See, e.g.*, Joel B. Eisen, *Can Urban Solar Become a "Disruptive" Technology?: The Case for Solar Utilities*, 24 NOTRE DAME J.L. ETHICS & PUB. POL'Y 53 (2010).

6. *See, e.g.*, Fred von Lohmann, *Fair Use as Innovation Policy*, 23 BERKELEY TECH. L.J. 829 (2008).

7. Indeed, of the forty-eight articles in the Westlaw database "Journals and Law Reviews" ("JLR") that cite one or more of Clayton Christensen's publications on "disruptive innovation," roughly half of them are in the health law field (search conducted on Oct. 8, 2013).

8. Terry, *supra* note 3.

9. The vast majority of this scholarship focuses on intellectual property, for obvious reasons. But some does not. *See, e.g.*, Ronald J. Gilson, *Locating Innovation: The Endogeneity of Technology, Organizational Structure, and Financial Contracting*, 110 COLUM. L. REV. 885 (2010).

10. The scholarship is too numerous to cite, but see, for example, Peter Huber, *The Old-New Division in Risk Regulation*, 69 VA. L. REV. 1025 (1983) (considering flaws in how Congress and federal regulators consider older, well-known risks versus newer, unkown ones); Gregory N. Mandel, *History Lessons for a General Theory of Law and Technology*, 8 MINN. J.L. SCI. & TECH. 551 (2007) (deriving various lessons for policymakers after considering past responses to new technologies); Thomas O. McGarity & Karl O. Bayer, *Federal Regulation of Emerging Genetic Technologies*, 36 VAND. L. REV. 461 (1983) (considering whether then-existing regulatory frameworks were adequate to regulate new genetic engineering technologies); Monroe E. Price, *The Newness of New Technology*, 22 CARDOZO L. REV. 1885 (2001) (examining how legal institutions address "the threat of the new," focusing mostly on information technologies).

11. Tim Wu, *Agency Threats*, 60 DUKE L.J. 1841 (2011). Daniel Gervais also tackles the question of regulating what he calls "inchoate technologies," distinguishing these from more "stable" technologies. Daniel Gervais, *The Regulation of Inchoate Technologies*, 47 HOUS. L. REV. 665, 671 (2010). Like Wu, his treatment is somewhat Internet-centric. Gervais argues for a general skepticism towards regulating inchoate technologies, and is less concerned with the timing and form of intervention. As such, Gervais seems primarily concerned with *whether* to regulate inchoate technologies, not how. *Id.* at 669. After these, the next closest to discussing

does not invoke disruption theory per se, but does address how regulators should confront novel technologies, business models, and practices.[12] In doing so, Wu engages a broader scholarly debate on the proper form of agency policymaking.[13] He argues that agencies should forgo traditional policymaking tools like rulemaking and adjudication in favor of informal "threats"—packaged in guidance documents, warning letters, and the like—when confronting industries that face conditions of rapid change and high uncertainty.[14] An informal, elastic "threat regime" is preferable, he argues, when agencies face dynamic rather than static industries characterized by disruptive innovation, unexpected market entries, new business models, and other exogenous shocks.[15] More traditional regulatory interventions like rulemaking and adjudication, the argument goes, risk being badly miscalibrated, or premature, or both.[16]

This Article tests the argument for agency threats and offers some important limitations. It argues that a regulatory threat works best as a temporary stopgap that presages more traditional regulatory intervention, not as a long-term strategy. Otherwise, the threat risks becoming stale or even counterproductive. Resource-strapped agencies that employ threats (most do) should guard carefully against relying on them as a long-term crutch. At some point, regulators must regulate.

In short, this Article demonstrates that agencies need not be so timid when confronting new technologies—even disruptive ones. If agencies are concerned about imposing regulation that is miscalibrated or premature, then they can reduce the cost of errors by using timing rules, alternative enforcement mechanisms, and other variations that might "soften" traditional regulation without undermining it long-term. In essence, regulators can experiment with binding approaches that can be more finely calibrated to the novel technology or business practice. The public interest demands that agencies maintain their fortitude in the face of regulatory disruption. And, somewhat counterintuitively, new technologies can benefit

disruptive technologies and regulatory responses is Donald Labriola, *Dissonant Paradigms and Unintended Consequences: Can (and Should) the Law Save Us from Technology?*, 16 RICH. J.L. & TECH. 1 (2009), although the article focuses on behavioral and social psychology rather than regulatory theory.

12. Wu uses the phrase "disruptive innovation" just once in *Agency Threats*, *supra* note 11, at 1848, and does not cite Christensen's work. However, he has cited it in previous writing. *E.g.*, TIM WU, THE MASTER SWITCH: THE RISE AND FALL OF INFORMATION EMPIRES 20 (2010).

13. *See also infra* Section IV.B.

14. Wu, *supra* note 11, at 1842.

15. *Id.* at 1848–49.

16. *Id.* at 1849.

from decisive, well-timed regulation. Some early regulatory interventions might even become, in Daniel Carpenter's words, "market-constituting," by enabling a robust market that otherwise might not exist, especially for credence goods that are difficult for consumers to evaluate.[17]

To make these points, this Article evaluates how a traditional regulatory agency, the FDA, has struggled to confront a technology that constantly evolves and reinvents itself—software. The FDA, responsible for regulating most new medical products, has spent the last century navigating waves of innovation in the food, drug, and medical device industries.[18] Indeed, one of the agency's major charges is to act as a gatekeeper for new medical products.[19]

But computerized medical devices have confounded the agency over the last twenty-five years, pushing the FDA far beyond its regulatory comfort zone. The FDA first confronted computerized medical devices in the mid-1980s, particularly after a series of patient deaths traced to the first radiation machines controlled by software.[20] Following these deaths, the FDA published a 1987 draft guidance that explained what types of software the agency would and would not regulate, including the requirements that might apply.[21] But the FDA relied on the draft guidance for the next eighteen years, withdrawing the policy unceremoniously in 2005,[22] leaving nothing in its

17. Daniel Carpenter, *Confidence Games: How Does Regulation Constitute Markets?*, *in* GOVERNMENT AND MARKETS: TOWARD A NEW THEORY OF REGULATION 164, 170 (Edward J. Balleisen & David A. Moss eds., 2010). As Carpenter explains:

> [D]rugs are types of credence goods, whose quality consumers can assess neither through inspection (as for "inspection goods" like a tomato) nor experience (as for experience goods like a job). Such goods, social scientists have demonstrated both theoretically and empirically, create "lemons problems." Because of informational shortcomings, consumers will continually purchase or consume inferior products when superior alternatives are available . . . .

Id. at 174.

18. *See* FOOD & DRUG LAW INST., FDA: A CENTURY OF CONSUMER PROTECTION (Wayne L. Pines ed., 2006) (discussing various historical accounts of the FDA's response to developments in the food, drug, device, and biologics industries).

19. *See generally* DANIEL CARPENTER, REPUTATION AND POWER: ORGANIZATIONAL IMAGE AND PHARMACEUTICAL REGULATION AT THE FDA (2010) (detailing an account of the FDA's gatekeeping role in pharmaceutical regulation).

20. For a fuller account of this incident, see *infra* Section III.B.

21. Draft Policy Guidance for Regulation of Computer Products, 52 Fed. Reg. 36,104 (Sept. 25, 1987) [hereinafter 1987 Notice of Draft Policy]; FDA Draft Policy for the Regulation of Computer Products (proposed Nov. 13, 1989) [hereinafter 1989 Draft Software Policy] (on file with author).

22. Annual Comprehensive List of Guidance Documents at the Food and Drug Administration, 70 Fed. Reg. 824,890 (Jan. 5, 2005).

place. The FDA never promulgated comprehensive software regulations and never even finalized the draft policy. All of this happened during a profound computer revolution no less, when software became increasingly ubiquitous and critical to patient safety. Today, the FDA repeats this pattern, releasing guidance after guidance to address software technologies in medicine.[23]

The FDA's approach to software over the last quarter century fits the agency threat framework strikingly well. The software industry has always been dynamic, not static. Many observers believe that recent software innovations will disrupt prevailing health industry norms and practices.[24] The FDA faces conditions of high uncertainty, to risk understatement. And the conventional wisdom has been that decisive software regulation by the agency would be premature.[25] So the FDA has relied on guidance documents to explain its expectations. It has not made rules. It has made threats.

But looking back, the threats embodied in the FDA's 1987 draft guidance were largely empty. The FDA's tentative, short-term approach to an emerging technology calcified into a long-term default, ultimately leading to suboptimal regulation for an entire industry. What started as a modest, flexible response by the FDA, consistent with the agency threat framework, became stale, ultimately undermining the agency's long-term authority over device software. This Article thus uses the FDA's approach to software, spanning a quarter century, to argue that agency threats are only useful as a short-term stopgap that presages more decisive intervention.

This Article then counterposes this example to the FCC's approach to the Internet. The FCC used threats to enforce principles of Internet neutrality against Internet service providers as an initial, short-term precursor to more decisive rules and enforcement. In contrast, the FDA's initial threats became a long-term crutch with no hard law backstop. These two approaches might be superimposed onto other examples, from the Federal Aviation Administration ("FAA") screening novel aircraft designs or flight software, to the Environmental Protection Agency ("EPA") considering new sources of pollution or anti-emissions technologies, to the Securities and

23. For example, in September 2013, the FDA published another guidance explaining its approach to the latest generation of device software, embodied in mobile medical applications for smartphones and tablets (called "medical apps" or more broadly, "mobile health" applications). FDA, Mobile Medical Applications (Sept. 25, 2013) [hereinafter Mobile Medical Applications], *available at* http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm263366.pdf.

24. *See, e.g.*, CHRISTENSEN ET AL., *supra* note 3; WILLIAM HANSON, SMART MEDICINE: HOW THE CHANGING ROLE OF DOCTORS WILL REVOLUTIONIZE HEALTH CARE 3–16 (2011) (describing uses of software technology in modern medicine).

25. *See infra* Sections IV.A, IV.B.

Exchange Commission ("SEC") engaging novel securities instruments or high-speed trading technologies. The FDA's own experience in relying primarily on guidance to confront novel things like genetically modified foods, nanotechnology, and xenotransplantation, demonstrates both the patterns and the stakes here.[26] Somewhat counterintuitively, both the public and the innovators can benefit from earlier, more decisive regulatory interventions, per the FCC's approach to net neutrality.

The Article begins in Part II by addressing "regulatory disruption," the idea that novel technologies or business practices can disturb existing regulatory frameworks. Part III then considers agency threats as a response, drawing on the FDA's long struggle to regulate computerized devices as a model to avoid, and the FCC's recent battle for net neutrality as the model to emulate. Part IV broadens this argument and assesses normatively how agencies should address regulatory disruption. Agencies facing disruption must make four related decisions, each of which is implicated when deciding whether to use "threats": (i) when to intervene, (ii) the form that intervention should take, (iii) how durable or transitory that intervention should be, and (iv) how rigorously to monitor and sanction noncompliance. "Threats" are not just a decision about form, but also represent decisions about timing, duration, and enforcement.

This Article argues that if agencies are concerned about regulating new technologies or business practices in error, or prematurely, then they can experiment with alternative timing and enforcement methods to reduce these risks while maintaining regulatory fortitude. Agencies can experiment with binding regulation. For example, agencies might use regulatory sunsets to better calibrate how long the intervention endures. Agencies could also specify rulemaking deadlines to make good on initial threats. And citizen suits or private rights of action could help correct systematic underenforcement by resource-strapped regulators.

II. DISRUPTIVE INNOVATION

In 1995, Joseph Bower and Clayton Christensen introduced the idea of "disruptive technologies" that depart fundamentally from existing ones, usually by being less complicated, more accessible, and less expensive.[27] Disruptive "technologies"—later broadened in the literature to disruptive

26. *See, e.g.*, Lars Noah, *Managing Biotechnology's [R]evolution: Has Guarded Enthusiasm Become Benign Neglect?*, 11 VA. J.L. & TECH., no 4, Spring 2006, at 1, 22–35, 45, 60–63.

27. Bower & Christensen, *supra* note 2.

"innovations"—undermine and then displace incumbents.[28] Classic examples include the Kodak camera, Bell telephones, Sony transistor radios, the Ford Model T, and Xerox photocopiers.[29] Contemporary examples include Cisco routers, Fidelity mutual funds, Google advertising, and the Southwest Airlines business model.[30] In each case, the market entrant introduced an innovation that eventually, though often not immediately, changed the entire market. The innovation "disrupts" the market in unforeseen ways.

This Article merges disruption theory with theories of regulation, using the idea of disruptive innovation to consider how regulators should respond to novel technologies or business practices that do not fit comfortably within their regulatory frameworks. The innovation typically falls within a specific agency's jurisdiction, but does not fit well within the agency's regulatory schemes that contemplate more established technologies or business practices. The innovation "disrupts" the regulatory framework, not necessarily industry incumbents, as envisioned by Bower and Christensen. "Regulatory disruption" occurs, then, when the "disruptee" is the regulatory framework itself.[31]

Certainly, there have been iconic innovations that did not fall within an existing agency's jurisdiction, eventually prompting Congress to create entirely new agencies. For example, the birth of commercial aviation during the barnstorming 1920s coincided with the birth of the Bureau of Air Commerce, the Civil Aeronautics Authority, and other precursors to the modern Federal Aviation Administration ("FAA").[32] The parallel evolutions of the automobile industry and the U.S. interstate highway system led Congress to create the National Highway Safety Agency, the precursor to the

28. CHRISTENSEN, *supra* note 2, at XV.

29. CHRISTENSEN ET AL., *supra* note 3, at 7.

30. *Id.*

31. Daniel Gervais addresses his analysis to "inchoate" technologies, which are fledgling technologies that have yet to mature. He does not focus his analysis explicitly on "disruptive" innovations, although he also seems to contemplate technologies that do not fit well into existing regulatory frameworks. *See* Gervais, *supra* note 11, at 671–72.

32. *See generally* Miranda Anger, *International Aviation Safety: An Examination of the U.S., EU, and the Developing World*, 72 J. AIR L. & COM. 141, 142–49 (2007); Frederick A. Ballard, *Federal Regulation of Aviation*, 60 HARV. L. REV. 1235 (1947). For a fascinating history of the government's unique role in innovation in the commercial airline industry, see David C. Mowery & Nathan Rosenberg, *The Commercial Aircraft Industry*, *in* GOVERNMENT AND TECHNICAL PROGRESS: A CROSS-INDUSTRY ANALYSIS 101–61 (Richard R. Nelson ed., 1982).

National Highway Traffic Safety Administration ("NHTSA"), and the Department of Transportation.[33]

But this Article is not concerned with innovations that have no regulatory "home," as these generally require congressional intervention to assign jurisdiction. Agencies cannot address them sua sponte.[34] The U.S. regulatory state has aged out of its formative years, when Congress created most modern regulatory agencies.[35] Today, there are hundreds of federal agencies, departments, and commissions,[36] many of which enjoy sweeping jurisdiction. For that reason, few innovations today will emerge without a regulatory "home" (or even multiple "homes") completely beyond every agency's jurisdiction. The contemporary challenge, then, is how existing agencies can confront these innovations given their broad but sometimes inert statutory frameworks.

33. Ralph Nader's famous book, UNSAFE AT ANY SPEED: THE DESIGNED-IN DANGERS OF THE AMERICAN AUTOMOBILE (1965), amplified growing public demand for auto safety regulation, leading Congress to pass the Highway Safety Act of 1966, Pub. L. No. 89–564, 80 Stat. 731 (codified as amended at 23 U.S.C. §§ 401–404 (2012)), the National Traffic and Motor Vehicle Safety Act of 1966, Pub. L. No. 89–563, 80 Stat. 718 (codified as amended at 49 U.S.C. §§ 30,101–30,126, 30,141–30,147, 30,161–30,169 (2012)), and the Department of Transportation Act, Pub. L. No. 89-670, 80 Stat. 931 (1966). In parallel, Congress confronted pollution by motor vehicles with the Motor Vehicle Air Pollution Control Act, Pub. L. No. 87-272, 79 Stat. 992 (1965), which vested regulatory authority with the U.S. Department of Health, Education, and Welfare ("HEW"). In 1969, Congress created the Council on Environmental Quality ("CEQ") with the National Environmental Policy Act of 1969, Pub. L. No. 91-190, 83 Stat. 852 (codified as amended at 42 U.S.C. §§ 4321–4347), which was later reorganized by executive order into the Environmental Protection Agency (EPA), Reorganization Plan No. 3 of 1970, 3 C.F.R. 1072 (1966–1970), *reprinted in* 84 Stat. 2086 (1970). For a history of the government's role in facilitating innovation in the automobile industry, see Lawrence J. White, *The Motor Vehicle Industry*, *in* GOVERNMENT AND TECHNICAL PROGRESS: A CROSS-INDUSTRY ANALYSIS, *supra* note 32, at 411–50.

34. *See* FDA v. Brown & Williamson Tobacco Corp., 529 U.S. 120 (2000) (striking down the FDA's attempt to regulate tobacco products without explicit congressional authority to do so).

35. *See generally* Jerry L. Mashaw, *Federal Administration and Administrative Law in the Gilded Age*, 119 YALE L.J. 1362 (2010) (describing an early history of administrative agencies); Jerry L. Mashaw, *Administration and "the Democracy": Administrative Law from Jackson to Lincoln, 1829–1861*, 117 YALE L.J. 1568 (2008); Jerry L. Mashaw, *Reluctant Nationalists: Federal Administration and Administrative Law in the Republican Era, 1801–1829*, 116 YALE L.J. 1636 (2007); Jerry L. Mashaw, *Recovering American Administrative Law: Federalist Foundations, 1787–1801*, 115 YALE L.J. 1256 (2006); Robert L. Rabin, *Federal Regulation in Historical Perspective*, 38 STAN. L. REV. 1189 (1986) (describing a broader period-by-period history of federal regulation).

36. *See Federal Agencies and Commissions*, THE WHITE HOUSE, http://www.whitehouse.gov/our-government/federal-agencies-and-commissions (last visited Sept. 19, 2013).

Perhaps the best modern example is the Internet. Since 1934, the FCC has had jurisdiction over communications networks.[37] But the Communications Act addressed a world of circuit-switching on telephone lines operated by monopoly "common carriers," not the future world of decentralized, digitized, packet-switching run by a constellation of smaller Internet service providers.[38] When this world began to emerge in the 1980s and 1990s, the FCC "self-consciously adopted a policy of non-regulation."[39] But the FCC's posture of separating "telecommunications services" from "information services" became less and less tenable as the Internet began to mature in the 1990s.[40] By then, it was clear that Internet service fell within the FCC's jurisdiction. But it did not square well with the FCC's longstanding framework.[41]

Another immediate example of disruptive innovation is the avalanche of novel securities instruments that contributed to the 2008 financial crisis. For years, financial firms had been experimenting with new and sometimes exotic securities products, particularly over-the-counter derivatives like mortgage-backed securities, credit default swaps, and collateralized debt obligations.[42] In the 1990s, the Commodities Futures Trading Commission ("CFTC") considered using its existing jurisdiction to regulate the derivatives market, until Congress intervened to bar it, at the behest of key regulators like Federal Reserve chairman Alan Greenspan and Treasury Secretary Robert

37. Communications Act of 1934, Pub. L. No. 73-416, 48 Stat. 1064, 1064 (codified as amended in scattered sections of 47 U.S.C. (2012)) ("An Act to provide for the regulation of interstate and foreign communication by wire or radio, and for other purposes.").

38. *See* Jonathan Weinberg, *The Internet and "Telecommunications Services," Universal Service Mechanisms, Access Charges, and Other Flotsam of the Regulatory System*, 16 YALE J. ON REG. 211, 211–12, 215–18 (1999) (juxtaposing Internet technologies with the regulatory framework created by the Communications Act of 1934, which created the FCC).

39. Philip J. Weiser, *The Future of Internet Regulation*, 43 UC DAVIS L. REV. 529, 531 (2009) (internal quotations omitted).

40. *See* Kevin Werbach, *The Network Utility*, 60 DUKE L.J. 1761, 1763–64 (2011); *see also* Weinberg, *supra* note 38, at 212, 232–34; Weiser, *The Future of Internet Regulation*, *supra* note 39, at 531.

41. *See infra* Section III.C (discussing the FCC's response in further detail).

42. *See* FINANCIAL CRISIS INQUIRY COMMISSION, THE FINANCIAL CRISIS INQUIRY REPORT, at xxiv (2011), *available at* http://fcic-static.law.stanford.edu/cdn_media/fcic-reports/fcic_final_report_full.pdf [hereinafter FCIC FINAL REPORT] (reviewing financial products and their contribution to the financial collapse of 2008); *see also* Brooksley Born, *Foreword: Deregulation: A Major Cause of the Financial Crisis*, 5 HARV. L. & POL'Y REV. 231 (2011). For an entertaining popular account of these novel securities and those who created them (and the few that actually understood them), see MICHAEL LEWIS, THE BIG SHORT: INSIDE THE DOOMSDAY MACHINE (2010).

Rubin.[43] Greenspan, a well-known laissez faire economist, pushed deregulation and self-regulation in the derivatives market, strongly resisting the CFTC's efforts to act on concerns that the market posed broad, systemic risks.[44] In the years leading up to the housing and stock market crash in 2008, the derivatives market would balloon to $673 *trillion* in notional amount.[45]

In the wake of the economic crisis, Congress created the Financial Crisis Inquiry Commission ("FCIC") to investigate the root causes of the financial collapse.[46] The FCIC attributed the collapse in part to the weak response by regulators to these novel financial products and practices. For example, the Federal Reserve Board had refused to use its authority to regulate aggressive new mortgage lending practices.[47] The Securities and Exchange Commission ("SEC") had failed to enforce disclosure requirements for mortgage-backed securities.[48] Congress had pushed to deregulate novel over-the-counter derivatives.[49] And regulators like the SEC, the Office of the Comptroller of the Currency ("OCC"), and the Office of Thrift Supervision ("OTS") exerted only superficial oversight of the banks engaged in these activities.[50] The FCIC's final report found that "[t]echnology has transformed the efficiency, speed, and complexity of financial instruments and transactions," but that the United States "had a 21st-century financial system with 19th-century safeguards."[51]

---

43. *See* Commodity Futures Modernization Act of 2000, Pub. L. No. 106-554, 114 Stat. 2763 (codified as amended in scattered sections of 7, 11, 12, 15 U.S.C.). Before that, Congress imposed a moratorium on the CFTC action. Omnibus Consolidated and Emergency Supplemental Appropriations Act, Pub. L. No. 105-277 § 760, 112 Stat. 2681-1, 2681-35 (1998). For a description of the role Rubin and Greenspan, among others, played in this matter, see FCIC FINAL REPORT, *supra* note 42, at 47. The over-the-counter derivatives market is one innovation that Wu concedes might have benefited from more traditional early regulation. Wu, *supra* note 11, at 1850.

44. Peter S. Goodman, *Taking a Hard New Look at a Greenspan Legacy*, N.Y. TIMES, Oct. 9, 2008, http://www.nytimes.com/2008/10/09/business/economy/09greenspan.html (quoting congressional testimony of Brooksley Born, then chair of the CFTC, that the opaque derivatives market could "threaten our regulated markets or, indeed, our economy without any federal agency knowing about it."); Born, *supra* note 42, at 236–38 (quoting Greenspan's testimony to Congress urging that regulation was unnecessary for over-the-counter derivatives).

45. FCIC FINAL REPORT, *supra* note 42, at xxiv.

46. Fraud Enforcement and Recovery Act of 2009, Pub. L. No. 111-21 § 5(a), 123 Stat. 1617, 1625.

47. FCIC FINAL REPORT, *supra* note 42, at xvii; Born, *supra* note 42, at 234–35.

48. FCIC FINAL REPORT, *supra* note 42, at 169–70; Born, *supra* note 42, at 235–36.

49. FCIC FINAL REPORT, *supra* note 42, at xxiv–xxv; Born, *supra* note 42, at 236–38.

50. FCIC FINAL REPORT, *supra* note 42, at xviii, 150–54, 173–174, 302–04; Born, *supra* note 42, at 238–42.

51. FCIC FINAL REPORT, *supra* note 42, at xv, xvii, xx.

These stories tell the modern predicament: how do regulators with broad statutory authority manage innovations as they mature, and as their problems become apparent?

III. EMPTY THREATS?

If certain innovations do not square well with existing regulatory frameworks, how should agencies respond, sans congressional intervention? This Part first considers the argument that agencies should rely on informal threats rather than more traditional modes of regulation. It then demonstrates how the FDA took precisely this approach with computerized devices twenty-five years ago, which led to suboptimal regulation of increasingly ubiquitous and dangerous products. This story cautions against over-relying on regulatory threats. Part III concludes by counterposing the FDA's approach to software with the FCC's use of threats to address Internet networks, which was disciplined by early enforcement and rulemaking to codify its threats. Unlike the FDA, the FCC did not over-rely on threats.

A. AGENCY THREATS

The agency threat framework argues that federal agencies should use informal "threats" against regulated industries under conditions of "high uncertainty."[52] For example, an agency might announce via guidance document, warning letter, or press release that it will take action against companies that employ novel technologies or business practices in a certain way.[53] But this type of regulatory elasticity, as demonstrated below, can harden into a long-term default position, leading to suboptimal regulation if not disciplined in some way.

An agency makes a "threat" when it gives "at least some warning of agency action related to either ongoing or planned behavior."[54] Agencies can package threats in many forms—guidance documents, interpretive rules, warning letters, press releases, official speeches, and even private meetings with regulated firms.[55] Threats can be very public, as a press release, or very

52. Wu, *supra* note 11, at 1848 (defining "high uncertainty" as scenarios "in which alternative future states of the world do not occur with quantifiable probability").

53. *Id.* at 1841, 1844.

54. *Id.* at 1844.

55. *Id.* Some of these vehicles themselves come in many forms. For example, guidance documents can include enforcement guidelines, policy statements, interpretive rules and memoranda, agency staff manuals and circulars, private ruling letters, informal advice, etc. *See* Final Bulletin, Agency Good Guidance Practices, 72 Fed. Reg. 3432, 3434 (Jan. 25, 2007); M. Elizabeth Magill, *Agency Choice of Policymaking Form*, 71 U. CHI. L. REV. 1383, 1386, 1391

private, as a face-to-face meeting.[56] Either way, the threat specifies behavior that the agency either desires or forbids.[57] Invoking a notorious example, Wu gives a nod to Don Corleone's use of threats in *The Godfather*.[58]

Regardless of their wrapping, threats are informal rather than formal, soft law rather than binding hard law, and are issued without much procedure (though they may be accompanied by quite a bit of ceremony).[59] This, indeed, is their appeal. Threats can issue without notice and comment or other procedural hurdles required by the Administrative Procedure Act ("APA") or other statutes.[60] As such, threats serve as a third alternative to rulemaking and adjudication, the two classic modes of regulation.[61] According to Wu, Peter Strauss, and other advocates of informal agency action, threats occupy "an important element in the hierarchy of agency law."[62] On this I agree.

Wu also contemplates threats as being unenforceable by agencies,[63] though ambiguity on this point seems to *be* the point. Otherwise, what is the agency threatening? The agency threat specifies either desired or forbidden behavior, and at least hints at potential repercussions. Threats, in short, are assertions that the agency will do something at some point given certain triggering activities. A threat that is unenforceable on its face would not seem to appeal to many agencies.

(2004). A 2007 Executive Order by the Bush administration defined "guidance document" as "an agency statement of general applicability and future effect, other than a regulatory action, that sets forth a policy on a statutory, regulatory, or technical issue or an interpretation of a statutory or regulatory issue." *See* Exec. Order No. 13,422, 3 C.F.R. 191–92 (2007) (rescinded by Exec. Order No. 13,497, 3 C.F.R. 218 (2010)).

56. Wu, *supra* note 11, at 1844.

57. *Id.*

58. Wu stops short of suggesting that agencies rely on some of the more aggressive tactics used by the mafia. *See id.* at 1857 (referencing THE GODFATHER (Paramount Pictures 1972)).

59. *Id.* at 1844. "Soft law" is described in Jacob E. Gersen & Eric A. Posner, *Soft Law: Lessons from Congressional Practice*, 61 STAN. L. REV. 573 (2008).

60. *See, e.g.*, Robert A. Anthony, *Interpretive Rules, Policy Statements, Guidances, Manuals, and the Like—Should Federal Agencies Use Them to Bind the Public?*, 41 DUKE L.J. 1311, 1328, 1334–35 (1992); Nathan Cortez, *Adverse Publicity by Administrative Agencies in the Internet Era*, 2011 BYU L. REV. 1371, 1380 (2011); Lars Noah, *Administrative Arm-Twisting in the Shadow of Congressional Delegations of Authority*, 1997 WIS. L. REV. 873, 874–76 (1997); Lars Noah, *The Little Agency That Could (Act with Indifference to Constitutional and Statutory Strictures)*, 93 CORNELL L. REV. 901, 904–05 (2008).

61. Wu, *supra* note 11, at 1841.

62. Gersen & Posner, *supra* note 59, at 626; Peter L. Strauss, *Publication Rules in the Rulemaking Spectrum: Assuring Proper Respect for an Essential Element*, 53 ADMIN. L. REV. 803, 804 (2001); Wu, *supra* note 11, at 1847; David Zaring, *Best Practices*, 81 N.Y.U. L. REV. 294, 294 (2006).

63. *See* Wu, *supra* note 11, at 1846.

Threats are useful, according to their proponents, under conditions of high uncertainty, and for dynamic rather than static states of industry.[64] Thus, for relatively stable industries in which business models, practices, and technologies are relatively well settled, agencies have enough information to regulate with certainty.[65] Legally binding rules make sense. Agencies can activate the lumbering machinery of rulemaking or adjudication without worrying that the effort will be for naught, or that "law" they produce will soon be obsolete. But in dynamic industries—characterized by disruptive innovation, unexpected market entries, new business models, and other exogenous shocks—agencies may lack sufficient information to regulate with certainty.[66] Doing so risks creating rules that are miscalibrated, or premature, or both.[67]

According to the threat framework, agencies face three choices when confronting dynamic states of industry: make law, make threats, or do nothing.[68] Again, making law via traditional rulemaking or adjudication can be premature or simply may generate flawed rules. And somewhat ironically, making law may create more uncertainty by triggering judicial challenges that can take years to resolve.[69] Doing nothing, or waiting, can also be undesirable because the industry may develop outside the public eye, and without considering the public interest.[70] Moreover, as the industry matures, the more settled its norms and business practices will become, which can be hard to reverse later.[71]

Threats, the argument goes, allow agencies to oversee the formative years of an industry without suffocating it.[72] Agency threats themselves can

64. *Id.* at 1848.

65. *Id.*

66. *Id.* at 1849.

67. *Id.* Although Wu does not discuss Clayton Christensen's "disruptive innovation" framework in *Agency Threats*, *supra* note 11, he does address it in a separate work. *See* WU, THE MASTER SWITCH, *supra* note 12, at 20.

68. Wu, *supra* note 11, at 1848–49. This list generally parallels the four choices David Super identifies when agencies must make decisions that appear to be unusually costly: (i) devote the resources necessary to make the decision, (ii) make a low-quality decision based on the limited decisional inputs it can muster, (iii) postpone a decision to when it has better decisional inputs, or (iv) do nothing. David A. Super, *Against Flexibility*, 96 CORNELL L. REV. 1375, 1407 (2011).

69. Wu, *supra* note 11, at 1849.

70. *Id.* at 1850. Gervais, conversely, argues that social norms and sometimes the technology itself will fill the regulatory void. Gervais, *supra* note 11, at 668.

71. Wu, *supra* note 11, at 1842, 1854.

72. *Id.* at 1851.

generate a public debate, which often contemplates the public interest.[73] And the lack of formality makes the threat quicker to issue and more flexible than rulemaking or adjudication.[74] If the threat turns out to be too harsh, or not harsh enough, the agency can easily retreat or recalibrate.

Proponents argue that the criticisms of agency threats are overblown. Rather than being a procedural end-run around the APA, an amplification of statutory authority, a way to skirt judicial review, or even an affront to the rule of law itself,[75] agency threats are precisely what agencies should use in conditions of high uncertainty. According to Wu, we should stop worrying and learn to love agency threats.[76]

This Article tests this argument by looking at two instances of agencies relying on threats in responding to disruptive technologies: the FDA's response to software and the FCC's response to the Internet. It is striking how closely the FDA's approach to software fits the threat framework. The medical device software industry has always been relatively dynamic.[77] The pace of innovation seems to be forever quickening, not abating, as evidenced by the latest generation of medical software for mobile devices.[78] To say that the FDA is facing conditions of high uncertainty is to risk understatement.[79] Moreover, frequently we hear the chorus that FDA rulemaking for software would be premature, due to the agency's lack of expertise, its lack of resources, and the state of the industry.[80]

---

73. *Id.* Gervais worries that quickly-developing social norms will evolve to circumvent any new regulatory framework. Gervais, *supra* note 11, at 673.

74. Wu, *supra* note 11, at 1851.

75. *Id.* at 1846–47. Wu focuses particular attention on Lars Noah's criticisms, *supra* note 60, which survey the range of problems and recent examples of abuse.

76. Wu, *supra* note 11, at 1842; *see also* Nicolas P. Terry, *Prescriptions Sans Frontières (or How I Stopped Worrying About Viagra on the Web but Grew Concerned About the Future of Healthcare Delivery)*, 4 YALE J. HEALTH POL'Y L. & ETHICS 183 (2004); DR. STRANGELOVE OR: HOW I LEARNED TO STOP WORRYING AND LOVE THE BOMB (Columbia Pictures 1964).

77. Nathan Cortez, Analog Agency in a Digital World (unpublished manuscript) (on file with author) (tracing the development of early medical device software to the present day, and the FDA's regulatory response).

78. *See, e.g.*, Nathan Cortez, *The Mobile Health Revolution*, 47 UC DAVIS L. REV. 1173 (2014) (surveying mobile medical applications and other mobile health technologies and examining their significance).

79. The FDA itself emphasizes the high uncertainty in its guidance for mobile medical applications, noting the "extensive variety of actual and potential functions of mobile apps, the rapid pace of innovation in mobile apps, and the potential benefits and risks to public health represented by these apps." Mobile Medical Applications, *supra* note 23, at 4.

80. *See, e.g.*, *FDA Public Workshop-Mobile Medical Applications Draft Guidance*, FDA, (Sept. 12–13, 2011), http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm267821.htm (last visited Oct. 29, 2013) [hereinafter *FDA Public Workshop*] (transcripts available at http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/uc

Consistent with the agency threat framework, the FDA did not make law. Nor did it do nothing. It published several guidance documents. It made threats. Moreover, these threats have at times generated public debate, sometimes even including lengthy public comment periods.[81] Finally, the latest wave of medical software developers seems to prefer some guidance to none,[82] although unsurprisingly it wants more clarity. In short, the FDA's approach to software over the last twenty-five years demonstrates how an agency relied on threats in precisely the circumstances described by Wu.

B. THE FDA AND THE QUARTER-CENTURY-LONG THREAT

Overburdened agencies like the FDA have long embraced informal threats. But there can be too much of a good thing. For twenty-five years, the FDA has relied on threats to address software in medical devices. This history cautions that threats can become stale and even counterproductive.

As with many regulatory interventions, the FDA's foray into device software was prompted by tragedy.[83] The story begins in 1985, when a Canadian company began selling the first therapeutic radiation machines controlled primarily by software.[84] Between 1985 and 1987, the machines massively over-radiated several patients in the United States and Canada, killing at least six.[85] The injuries were attributed to various software design flaws and user errors.[86] When the FDA investigated these incidents, it required the manufacturer to make certain corrections, which were not fully

m275908.htm and http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm275909.htm).

81. *Id.*

82. The mHealth Regulatory Coalition objected to the Bennett-Hatch amendment to the FDA user fee bill that would have prohibited FDA from finalizing its Draft Guidance for eighteen months. Letter from mHealth Regulatory Coalition to Senators Enzi and Harkin (May 17, 2012), http://mhealthregulatorycoalition.org/wp-content/uploads/2010/06/MRC-Letter-to-Senate-HELP-Committee-on-Proposed-Moratorium-FINAL.pdf.

83. The tragedy over thalidomide helped prompt the 1962 Drug Amendments. CARPENTER, *supra* note 19, at 256–60; J. Richard Crout et al., *FDA's Role in the Pathway to Safe and Effective Drugs*, *in* FDA: A CENTURY OF CONSUMER PROTECTION, *supra* note 18, at 159, 168–69.

84. *See generally* Nancy G. Leveson, *Medical Devices: The Therac-25*, *in* SAFEWARE: SYSTEM SAFETY AND COMPUTERS 515 app. A (1995) (providing a comprehensive story of the Therac-25 saga). The FDA had begun considering software before the Therac-25, as shown by its letters in response to congressional committees considering computers in medicine. *See Information Technologies in the Health Care System, Hearing before the Subcomm. on Investigations and Oversight of the H. Comm. on Sci. and Tech.*, 99th Cong. 196–204 (1986). But the Therac-25 saga seemed to push the agency towards announcing a policy towards software.

85. Leveson, *supra* note 84, at 515.

86. *Id.*

implemented until two years later, after a recall.[87] Yet, at the time, the FDA's response was considered "impressive," given that it had no policy for medical software.[88]

In 1986, FDA Commissioner Frank Young said in a speech that the agency would approach device software with the "least regulation consistent with the requirements of public health and safety."[89] This speech represented the first threat, though the threat was a soft, imprecise one. In 1987, the FDA published its first draft policy on software, explaining which types of software the agency would and would not regulate, and which regulatory requirements might apply.[90]

In 1989, the FDA updated the document, which became known as the "Draft Software Policy."[91] The policy confirmed that the FDA's "basic philosophy for computer products" was "to apply the least degree of regulatory control necessary to provide reasonable assurance of safety and effectiveness."[92] Like the 1987 version, the 1989 policy delineated which computer products the agency would and would not regulate.[93] It clarified how these new technologies might fit into the statutory definition of "device," which Congress wrote in 1976 when no one could have imagined today's versions.[94] Together, the 1987 and 1989 draft software policies constituted "threats," consistent with the framework.

Since that time, the FDA's approach to computerized medical devices has been the archetype of regulatory minimalism. The agency never finalized the 1989 Draft Software Policy; in fact, the FDA withdrew it without explanation in 2005.[95] In the absence of legally binding rules, lawyers have

87. *Id.* at 525, 544–48.

88. *Id.* at 553.

89. FDA & National Library of Medicine, Software Policy Workshop (Sept. 3–4, 1996), *reprinted in* FOOD AND DRUG LAW INSTITUTE, BIOMEDICAL SOFTWARE REGULATION 403 (Thomas E. Colonna & Jonathan S. Helfgott eds., 2009) [hereinafter FDA & NLM, 1996 Software Policy Workshop].

90. 1987 Notice of Draft Policy, *supra* note 21; *see also* Medical Devices; Medical Software Devices; Notice of Public Workshop, 61 Fed. Reg. 36,886 (July 15, 1996). The FDA had begun crafting the policy in 1985. *See* FDA & NLM, 1996 Software Policy Workshop, *supra* note 89, at 403.

91. 1989 Draft Software Policy, *supra* note 21; *see* Mobile Medical Applications, *supra* note 23, at 6.

92. E. Stewart Crumpler & Harvey Rudolph, *FDA Software Policy and Regulation of Medical Device Software*, 52 FOOD & DRUG L.J. 511, 513 (1997).

93. 1989 FDA Draft Software Policy, *supra* note 21.

94. Medical Device Amendments of 1976, Pub. L. No. 94-295 (1976) (codified as amended at 21 U.S.C. § 321 (2012)).

95. Annual Comprehensive List of Guidance Documents at the Food and Drug Administration, 70 Fed. Reg. 824,890 (Jan. 5, 2005).

had to advise clients based largely on the 1989 draft policy, even after the FDA withdrew the policy. The agency did pre-clear discrete software products on a case-by-case basis, which could serve as precedent for similar products.[96] But this piecemeal approach fell far short of establishing a definitive, cohesive philosophy towards software that companies could use to predict their regulatory obligations.

Ironically, after the FDA withdrew the 1989 guidance, it explained that "it would be impractical to prepare an overarching software policy to address all of the issues related to the regulation of all medical devices containing software" because "the use of computer and software products as medical devices grew exponentially and the types of products diversified and became more complex."[97] Thus, rather than providing more oversight as medical technology matured—during a profound computer revolution—the FDA provided less.

By and large, the FDA has avoided proceeding by rule here. It has promulgated very few prospective regulations governing software, and what little it has done addresses relatively low-risk devices.[98] For example, in 2011, the FDA finalized a rule governing medical device data systems. But these represent only a narrow slice of low-risk products that merely transfer, store, display, or convert medical device data, without doing much else.[99] Periodically, in the preambles to final rules, the FDA will acknowledge computer products. For example, when finalizing its Quality Systems Regulation ("QSR") for devices in 1996, the FDA observed that software design flaws and the failure to validate software after maintenance were the most common source of errors.[100] The QSR, which establishes good manufacturing practices for devices,[101] has been perhaps the one area in

96. For example, the FDA has created dozens of regulatory categories for devices that incorporate software, including medical calculators, cameras, lights, magnifiers, microscopes, monitors, recorders, reminders, scales, surgical tools, and a host of data systems that store, display, and manipulate data. Mobile Medical Applications, *supra* note 23, at 29–31 (listing distinct device categories codified by FDA).

97. FDA, Draft Guidance for Industry and Food and Drug Administration Staff: Mobile Medical Applications, 5 (July 11, 2011) [hereinafter 2011 Mobile Medical Applications Draft] (on file with author) (superseded by Mobile Medical Applications, *supra* note 23) (referring to the 2005 withdrawal of the 1989 Draft Software Policy).

98. For example, in 2011, the agency promulgated a rule classifying devices that electronically display, store, transfer, or convert medical device data, known as Medical Device Data Systems ("MDDS"). 76 Fed. Reg. 8637 (Feb. 15, 2011) (codified at 21 C.F.R. § 880.6310 (2013)).

99. *See* 21 C.F.R. § 880.6310(a); 76 Fed. Reg. 8643–44 (Feb. 15, 2011).

100. Medical Devices; Current Good Manufacturing Practice ("CGMP"), 61 Fed. Reg. 52,602, 52,617, 52,620 (Oct. 7, 1996).

101. 21 C.F.R. § 820.1 (2013).

which the FDA has provided firm standards for software. But the QSR is notable for giving manufacturers significant flexibility to design and manufacture devices according to customized specifications.[102] This is both a strength and a weakness—a strength in recognizing the diversity of medical devices and the implausibility of generating one-size-fits-all standards, but a weakness in decentralizing standards and delegating significant discretion to regulated firms. Of course, the FDA relies on guidance to explain how the QSR applies to software.[103]

Indeed, the FDA heavily relies on guidance to oversee software. Agency documents that summarize the FDA's approach generally cite to the same cluster of five guidances.[104] Together, these documents form a cascade of quasi-regulation, recommendations, and "current thinking,"[105] but offer few firm rules. Software does not stand on terra firma with the FDA. Looking back, the 1987 document was like the gateway drug that led to guidance after guidance for the next twenty-five years.

Unfortunately, the FDA's response to software has not been commensurate with how ubiquitous and critical to patient safety software has become. Widespread problems with device software persist. For example, between 2009 and 2011, the *New York Times* documented several hundred catastrophic injuries caused by software and user errors related to the newest generation of radiation machines.[106] The incidents bear striking similarity to the problems with radiation software that originally prompted the FDA's

102. *See Medical Devices Quality System (QS) Regulation/Medical Device Good Manufacturing Practices*, FDA, http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/QualitySystemsRegulations/default.htm (last updated Apr. 28, 2011).

103. *See, e.g.*, FDA, General Principles of Software Validation (Jan. 11, 2002), *available at* http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm085371.pdf.

104. *See* FDA, Cybersecurity for Networked Medical Devices Containing Off-the-Shelf (OTS) Software (Jan. 14, 2005), *available at* http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm077823.pdf; FDA, Guidance for the Content of Premarket Submissions for Software Contained in Medical Devices (May 11, 2005), *available at* http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm089593.pdf; FDA, General Principles of Software Validation, *supra* note 103; Guidance for Industry, FDA Reviewers and Compliance on Off-The-Shelf Software Use in Medical Devices (Sept. 9, 1999), *available at* http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm073779.pdf; FDA, Design Control Guidance for Medical Device Manufacturers (Mar. 11, 1997), *available at* http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm070642.pdf.

105. *See, e.g.*, FDA, *supra* note 103, at 1.

106. *See, e.g.*, *Radiation Boom*, N.Y. TIMES, http://topics.nytimes.com/top/news/us/series/radiation_boom/index.html (last visited Oct. 10, 2013) (linking to twenty-six articles in the series on medical radiation by Walt Bogdanich).

intervention two decades earlier. In one case, a hospital in Manhattan exposed a cancer patient to fatal doses of radiation for three consecutive days after the software repeatedly crashed.[107] The patient died two years later.[108] In another case, a Brooklyn hospital administered three and a half times the prescribed dose of radiation to a breast cancer patient for twenty-seven days due to user errors and misprogrammed software.[109] The *New York Times* documented hundreds of similar mistakes,[110] with one expert estimating that one in twenty radiation patients nationwide will suffer injury.[111]

Unfortunately, the FDA's posture towards software is reactive rather than proactive. To wit, the FDA subsequently investigated these radiation software failures reported by the *New York Times*, but only after they generated national attention.[112]

Yet, the FDA seems to be generally aware of the dangers posed by software products. In February 2010, the FDA revealed during a public meeting that it had received voluntary reports of 260 malfunctions, forty-four injuries, and six deaths related to health information technologies.[113] But because these were voluntary reports, and because even the FDA's mandatory reporting requirements suffer from dramatic under-reporting, the real numbers are no doubt much higher.[114]

These injuries are the byproducts of medical innovation. As alluring as medical innovation is, it is not an unmitigated good. The role of regulators like the FDA is to facilitate the benefits of new technologies while managing

107. Walt Bogdanich, *Radiation Offers New Cures, and Ways to Do Harm*, N.Y. TIMES, Jan. 23, 2010, http://www.nytimes.com/2010/01/24/health/24radiation.html.

108. *Id.*

109. *Id.*

110. *Id.* ("Over the previous eight years, despite hundreds of mistakes, the state issued just three fines against radiotherapy centers, the largest of which was $8,000."); Walt Bogdanich, *As Technology Surges, Radiation Safeguards Lag*, N.Y. TIMES, Jan. 27, 2010, http://www.nytimes.com/2010/01/27/us/27radiation.html.

111. Bogdanich, *supra* note 107 (quoting Dr. John J. Feldmeier of the University of Toledo).

112. *Id.*

113. Testimony of Jeffrey Shuren, Director, FDA's Center for Devices and Radiological Health, Health Information Technology ("HIT") Policy Committee, Adoption/Certification Workgroup (Feb. 25, 2010) (on file with author).

114. *See* Edward M. Basile & Beverly H. Lorell, *The Food and Drug Administration's Regulation of Risk Disclosure for Implantable Cardioverter Defibrillators: Has Technology Outpaced the Agency's Regulatory Framework?*, 61 FOOD & DRUG L.J. 251, 257–58 (2006) ("The [FDA] estimates that as few as one in every 100 medical device adverse events actually is reported to [the] FDA, although there is no hard data to support this estimate.").

their risks.[115] Doing so should support rather than undermine long-term markets for the technologies, preserve consumer trust, and level the playing field among competitors. Facilitating new technologies need not be in tension with managing their risks. The challenge for regulators is pursuing both during early periods of high uncertainty.

C. THE FCC AND THE SIX-YEAR-LONG THREAT

In contrast to the FDA's quarter-century-long threat, the FCC took a more disciplined approach when confronted with its own disruptive innovation. Unlike the FDA's initial "threat" in 1987, which lingered for decades, the FCC's initial "threat" in 2004—warning Internet service providers to maintain net neutrality[116]—was quickly reinforced by enforcement and, ultimately, by rulemaking just six years later. The differences here are instructive. Like the FDA, the FCC was confronted with a new technology that did not square well with its existing regulatory framework. And like the FDA's early posture towards software, the FCC's early posture towards the Internet was one of benign neglect.[117] The lack of high-profile problems seemed to justify this approach,[118] perhaps as it also justified FDA's early stance towards software.

The Internet developed gradually as a government and academic network between the late 1960s and the late 1980s.[119] Since 1934, jurisdiction over communications networks has resided with the FCC.[120] As early as 1966, the FCC foresaw the potential convergence between computers and communications, initiating what it called the *Computer I* inquiry to study how

115. For example, the Federal Food, Drug, and Cosmetic Act authorizes the FDA to pre-approve only new drugs that are safe and effective, which requires balancing their risks and benefits. 21 U.S.C. § 355 (2012).

116. Wu, *supra* note 11, at 1844–45, 1852 n.41. Wu originally coined the phrase "net neutrality." *See* Tim Wu, *Network Neutrality, Broadband Discrimination*, 2 J. TELECOM. & HIGH TECH. L. 141 (2003).

117. Weiser, *The Future of Internet Regulation*, *supra* note 39, at 531, 533 ("To date, the Internet has developed outside of the FCC's traditional regulatory model, enjoying freedom from regulatory oversight.").

118. *Id.* at 537.

119. *See generally* JANET ABBATE, INVENTING THE INTERNET (1999) (providing a detailed historical account of the Internet); *A Brief History of the Internet & Related Networks*, INTERNET SOCIETY, http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet-related-networks (last visited Oct. 10, 2013) (providing a briefer history). For a compelling account of the government support required to develop the Internet and personal computing, see Steve Bickerstaff, *Shackles on the Giant: How the Federal Government Created Microsoft, Personal Computers, and the Internet*, 78 TEX. L. REV. 1 (1999).

120. Communications Act of 1934, Pub. L. No. 73-416, 48 Stat. 1064 (codified as amended in scattered sections of 47 U.S.C.) ("An Act to provide for the regulation of interstate and foreign communication by wire or radio, and for other purposes.").

computers might both operate on and replicate traditional telephone networks.[121] As a result of the inquiry, the FCC "chose to quarantine data processing from regulated telecommunications, rather than tackle the public policy considerations of the nascent computer utility directly."[122]

During the twentieth century, federal communications law developed discrete regulatory regimes for discrete services, such as telephony, broadcast television, cable television, and satellite.[123] Internet Protocol transmissions did not fit neatly, if at all, into these regulatory categories.[124] The FCC's regulatory framework, based on the hoary Communications Act of 1934, contemplated a world of circuit-switching on legacy telephone lines run by monopoly "common carriers"; it did not anticipate (how could it?) the decentralized, digitized, packet-switching Internet that later emerged.[125]

During the 1980s and 1990s, the FCC "self-consciously adopted a policy of non-regulation toward the Internet during its emergence as an important commercial network."[126] But as the Internet began to grow exponentially in the mid 1990s, the FCC's posture of separating "telecommunications services" from "information services" became less and less tenable.[127] Indeed, dividing the communications world by service category has become obsolete with computer-to-computer communications that can now replicate other categories of services.[128] Thus, Internet service clearly fell within the FCC's jurisdiction. But it did not square well with the FCC's longstanding framework, based as it was on service categories. The categorical lines evaporated with Internet technologies.

The FCC understandably struggled with whether to apply traditional regulatory frameworks to a technology as dynamic as the Internet.[129] Even 1996 legislation addressing Internet networks became outdated by the late

121. Regulatory and Policy Problems Presented by the Interdependence of Computer and Communications Services and Facilities, 7 F.C.C.2d 11 (1966); *see also* Bickerstaff, *supra* note 119, at 7–13 (recounting the FCC's launch of the *Computer I* inquiry).

122. Werbach, *supra* note 40, at 1763.

123. Weinberg, *supra* note 38, at 213.

124. *See id.*

125. Weinberg, *supra* note 38, at 215–18 (juxtaposing Internet technologies with the regulatory framework created by the Communications Act of 1934, which created the FCC).

126. Weiser, *supra* note 39, at 531 (internal quotation marks omitted).

127. Weinberg, *supra* note 38, at 212, 232–34; Weiser, *supra* note 39, at 531; Werbach, *supra* note 40, at 1764.

128. Weinberg, *supra* note 38, at 232–34.

129. Weiser, *supra* note 39, at 549–50.

1990s.[130] Internet scholars rehashed the familiar debate about whether rulemaking or adjudication would be more effective.[131]

Eventually, Internet technologies and business practices matured enough to warrant binding regulation.[132] But the FCC proceeded initially by making a threat. In 2004, FCC Chairman Michael Powell gave a speech at a popular telecommunications conference, warning the broadband industry that it should preserve net neutrality in various ways, such as allowing broadband users to attach different devices to their networks and use Internet applications of their choice.[133]

This threat turned out to be far from empty. Just one year after Powell's speech, the FCC initiated an investigation against a telephone company that violated one of these principles.[134] Of course, the FCC did not allege that the company violated Powell's speech. Rather, it cited violations of the Communications Act of 1934.[135] The FCC backed up its threat.

Moreover, the FCC elaborated on Internet freedoms in a policy statement shortly after the settlement[136] and promulgated a legally binding rule just six years later, codifying the thrust of Powell's original speech.[137]

Real enforcement was the third (and crucial) part of the FCC's three-part strategy to regulate the Internet.[138] Adjudication and enforcement require sustained effort.[139] In comparison, the FDA has brought intermittent

130. *Id.* at 561 ("By the late 1990s, technological and market conditions had outpaced the premises that underpinned the Telecommunications Act of 1996.").

131. *See, e.g.*, Christopher S. Yoo, *Beyond Network Neutrality*, 19 HARV. J.L. & TECH. 1, 67 (2005); Weiser, *supra* note 39, at 563–64.

132. *See* Weiser, *supra* note 39, at 531 (arguing that early non-regulation of the Internet by the FCC was appropriate, but now it is not).

133. Wu, *supra* note 11, at 1844–45 (citing Michael K. Powell, Chairman, FCC, Preserving Internet Freedom: Guiding Principles for the Industry, Remarks at the Silicon Flatirons Symposium on "The Digital Broadbrand Migration: Toward a Regulatory Regime for the Internet Age" (Feb. 8, 2004), *available at* http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-243556A1.pdf).

134. *See* Wu, *supra* note 11, at 1845 (discussing Madison River Communications, LLC, 20 FCC Rcd. 4295 (2005), *available at* http://digital.library.unt.edu/ark:/67531/metadc4102/m1/371.

135. *Madison River*, 20 FCC Rcd. 4295, 4296 (citing the Communications Act of 1934, 47 U.S.C. § 201(b)); Weiser, *supra* note 39, at 563.

136. Appropriate Framework for Broadband Access to the Internet over Wireline Facilities, Policy Statement, 20 FCC Rcd. 14,986 (Aug. 5, 2005), *available at* http://digital.library.unt.edu/ark:/67531/metadc4091/m1/708.

137. Wu, *supra* note 11, at 1852; Preserving the Open Internet, 25 FCC Rcd. 17,905 (Dec. 21, 2010), *available at* http://digital.library.unt.edu/ark:/67531/metadc39308/m1/451.

138. *See* Weiser, *supra* note 39, at 584–89.

139. *See id.* at 585–86 (noting various deficiencies and underuse of the FCC's Enforcement Bureau and administrative law judges ("ALJs")).

enforcement actions against software products,[140] but it has not brought enough high-profile cases to offset its primary reliance on nonbinding (and since withdrawn) guidance.

The FCC deployed threats as a short-term precursor to more traditional, binding regulation; the FDA relied on threats as a long-term crutch. And unlike the FDA, the FCC did not make empty threats. These disparate experiences reveal important limitations on agency threats.

IV. THE REGULATORY TOOLKIT FOR DISRUPTIVE INNOVATIONS

By its nature, disruptive innovation challenges regulators. U.S. culture values innovation, sometimes for its own sake, and does not want regulation to be technologically regressive.[141] At the same time, new technologies can disturb existing regulatory frameworks, triggering intense public debates about whether they are compatible with accepted social and legal norms.[142] Instinctively, most assume that regulating novel technologies requires more information and longer decision-making processes than familiar, established technologies do.[143]

Compounding the challenge, regulators like the FDA and FCC must make four different types of decisions when confronting novel technologies or business practices. The decision to rely on agency "threats" is not only a decision about the form of intervention, but also about the timing of intervention, how durable that intervention should be, and how rigorously

140. For example, the FDA has an online database of warning letters issued to suspected violators. It categorizes these warning letters by subject, with three relating to device software: "Devices/Computer Software," listing only two warning letters (one in 1997 and one in 1998); "Devices/Patient Monitors Software," listing one warning letter from 1999; and "QSR for Medical Devices/Picture Archiving & Communication Systems Software/Adulterated," listing one warning letter in 2001. FDA's Electronic Reading Room–Warning Letters, FDA, http://www.accessdata.fda.gov/scripts/warningletters/wlFilterBySubject.cfm (last visited Nov. 9, 2013) (follow hyperlinks from subject index to individual subjects). Of course, these database categories may not be comprehensive, and a search through the FDA's warning letter database for "software" generates many more results (although some of these do not concern software devices, but software used for manufacturing or other reasons).

141. *See* Huber, *supra* note 10, at 1028.

142. Gaia Bernstein, *The Socio-Legal Acceptance of New Technologies: A Close Look at Artificial Insemination*, 77 WASH. L. REV. 1035, 1035–36 (2002).

143. *Cf.* Super, *supra* note 68, at 1411, 1416–17 (noting that "commentators assume that more information is an unalloyed good" and that "delaying decision making is necessary to allow procedures consistent with careful deliberation or broad participation," while new technologies can "complicate" and "chang[e] the optimal timing of" decision-making).

the agency should monitor and sanction noncompliance. The regulator thus makes four related types of decisions:

1. Timing: When should the agency intervene, if at all? Does waiting necessarily generate a better informational basis on which to regulate? What are the drawbacks of waiting?
2. Form: Should the agency regulate via rule, adjudication, guidance, or some alternate form? Given the costs and benefits of each, which best accommodates the uncertainties of the innovation? Does form even matter?
3. Durability: Should the agency's intervention be permanent, or temporary, or conditional? How long should it endure? And are there ways to better calibrate regulatory interventions to the innovation?
4. Enforcement: How rigorously should the agency monitor and sanction noncompliance? How much should agencies temper enforcement against novel products, firms, or industries?

With each decision, an agency must prioritize competing claims on these questions by different constituents, including Congress, the President, industry, and regulatory beneficiaries (those who benefit from the regulation of others).[144]

This Part considers how agencies might respond to disruptive innovation along these four dimensions. Informal agency "threats" seem, at first glance, like a nice middle ground—allowing agencies to say something early, but in a malleable and nonbinding way, while preserving maximum flexibility to gather more information. However, this Article questions the emerging sentiment that regulators should use "threats" to maintain a flexible stance, drawing lessons from the FDA's approach to software and the FCC's approach to the Internet. As explained above, the FCC disciplined its threats with timely enforcement and rulemaking. But the FDA's strategy compounded its problems along each dimension: on timing, it was late; on form, it was casual; on durability, it was transient; and on enforcement, it was spotty. These cumulative defects have undermined the FDA's long-term authority over software and raise serious questions about the prevailing orthodoxy on regulating new technologies.

144. Nina A. Mendelson, *Regulatory Beneficiaries and Informal Agency Policymaking*, 92 CORNELL L. REV. 397, 401–02 (2007).

A. TIMING

For regulators confronting disruptive technologies, timing is paramount. Premature regulation might suffocate new technologies.[145] Regulating with alacrity risks regulating in error.[146] Yet, deferring regulation to some future date forgoes the benefits that would accrue to the public while the technology is being regulated.[147] As the FDA's posture on software reveals, initial reticence to regulate can easily harden into a long-term, laissez faire default. Introductory stances often remain unexamined.

With novel technologies, agencies also risk paralysis by analysis. New technologies often present unforeseen risks if under-regulated[148] and dramatic opportunity costs if overregulated.[149] The inclination of most regulators is to avoid both extremes. Thus, the timing calculus is highly variable for new technologies. Agencies understandably agonize over it.

Despite the stakes, legal scholarship is oddly ambivalent about timing.[150] Administrative law scholars in particular are far more preoccupied with the *form* regulation takes rather than *when* agencies take it. But the two, no doubt, are connected. As Jacob Gersen and Eric Posner emphasize, decisions about the timing of laws can be just as important as decisions about their content.[151]

David Super argues convincingly that regulatory decisions are often postponed counterproductively, usually to wait for more information and

145. Wu, *supra* note 11, at 1849.

146. Jacob E. Gersen & Eric A. Posner, *Timing Rules and Legal Institutions*, 121 HARV. L. REV. 544, 558 (2007). Gervais, in particular, is concerned with regulating in error. Gervais, *supra* note 11, at 674.

147. Gersen & Posner, *supra* note 146, at 558.

148. *See, e.g.*, Lars Noah, *Assisted Reproductive Technologies and the Pitfalls of Unregulated Biomedical Innovation*, 55 FLA. L. REV. 603, 614–24 (2003) (discussing mounting evidence of risks associated with the relatively unregulated field of assisted reproductive technology); Rabin, *supra* note 35, at 1304–05 (noting that, in the realm of health and safety regulation, "inaccurate, insufficiently protective, administrative decisions might lead to irreversible long-term risks to society of devastating magnitude"); Nicholas Bagley & Richard L. Revesz, *Centralized Oversight of the Regulatory State*, 106 COLUM. L. REV. 1260, 1296–1300 (2006) (suggesting reasons why regulators might err on the side of under-regulation).

149. Huber, *supra* note 10, at 1027.

150. Super, *supra* note 68, at 1379 (arguing that scholarship on timing tends to "confound the questions of *when* a decision should be made with *who* should make it"). A recent article that does focus sustained attention to timing focuses on legislative timing, giving only passing thought to administrative agencies. *See* Gersen & Posner, *supra* note 146. Gervais addresses the timing of regulation for new technologies, but relies on the axiom that "technological change happens faster than social change" to argue that regulation of new technologies may be premature and fraught with error. Gervais, *supra* note 11, at 683–84.

151. Gersen & Posner, *supra* note 146.

preserve maximum flexibility.[152] This observation describes the FDA's posture towards software. Regulatory flexibility and discretion sound good in theory. A flexible stance allows regulators to revisit and retreat to earlier positions if necessary, and tailor their approaches as they learn more. Indeed, American culture generally values flexibility and condemns inflexibility.[153] Our legal culture valorizes executive discretion.[154] And legal scholarship in particular embraces regulatory flexibility, usually as a reaction to the ossification of agency rulemaking.[155]

But flexibility often leads to what Super calls "legal procrastination" and a resulting regulatory inertia.[156] Administrative law scholarship generally assumes that deferring action will necessarily lead agencies to use more robust, deliberative procedures that will generate more fair and accurate rules.[157] But as the FDA's policy towards software demonstrates, this assumption can be wrong. Flexibility and lengthy deliberation are only worthwhile if they will significantly improve the quality of the agency's decision.[158] Bureaucratic delay is widely recognized and condemned in American culture, but its costs to regulatory beneficiaries are infrequently studied.[159]

Regulators also tend to overvalue gathering information about new technologies, which further distorts decisions on regulatory timing. The need for information has long been a key concern when addressing new risks.[160] Wu channels Freidrich Hayek in arguing that traditional regulation is

152. Super, *supra* note 68, at 1380, 1409.

153. *Id.* at 1409 ("In social life, calling someone 'flexible' is generally a compliment; inflexibility is characteristic of bullies, dinosaurs, and control freaks.").

154. *Id.* at 1409.

155. *Id.* at 1410 n.148 (citing a string of very recent articles in a range of subjects touting flexibility); Thomas O. McGarity, *The Expanded Debate over the Future of the Regulatory State*, 63 U. CHI. L. REV. 1463, 1523 (1996).

156. Super, *supra* note 68, at 1382.

157. *Id.* at 1398.

158. *Id.* at 1406.

159. One exception is the literature on delaying environmental regulation. *See, e.g.*, Lea-Rachel D. Kosnik, *Sources of Bureaucratic Delay: A Case Study of FERC Dam Relicensing*, 22 J.L. ECON. & ORG. 258 (2005); Hilary Sigman, *The Pace of Progress at Superfund Sites: Policy Goals and Interest Group Influence*, 44 J.L. & ECON. 315 (2001); Amy Whritenour Ando, *Waiting to be Protected Under the Endangered Species Act: The Political Economy of Regulatory Delay*, 42 J.L. & ECON. 29 (1999).

160. Gervais, *supra* note 11, at 684–87 (addressing the "unintended consequences" that generate from early regulatory intervention, even questioning the efficacy of regulatory requirements like car seatbelts). *But see* Huber, *supra* note 10, at 1051–52 (acknowledging the concern, but questioning its salience).

"impressive in a world of perfect information, but terrible in this world."[161] However, for regulators, perfect information is often the enemy of good information.[162] Culturally, we are situated firmly in the Information Age, and thus tend to overvalue information.[163] We obsess about gathering all available information and recoil from making any big decisions without it.[164] Legal culture in particular values information over other inputs into legal decisions, such as the capacity to implement and enforce laws.[165] And the penetration of economics into legal thought, particularly the well-known market failure of information asymmetries, further aggravates this overemphasis on gathering information.[166]

But more information does not necessarily equate to better information. Ironically, the more information we gather, the more effort it takes to make a decision based on it.[167] Nor does information itself necessarily make agencies more efficient long-term. The more the FDA learned about device software, the more overwhelmed it became. Indeed, the FDA explained that it withdrew its 1989 Draft Software Policy because of the volume, variety, and complexity of medical device software.[168] In fact, waiting for better information can be counterproductive. Agency flexibility can be overrated.[169] Too much of it can paralyze agencies,[170] a notion that might offend our intuitions about how government should work.

Equally important, early interventions can benefit both regulated industry and regulatory beneficiaries. As Super demonstrates, "[d]ecisions rarely become more valuable to society as a whole when rendered later."[171] Indeed, regulated parties may appreciate the added certainty and reduced cost of compliance with early regulatory decisions.[172] The conventional wisdom that

161. Wu, *supra* note 11, at 1849 (citing F.A. Hayek, *The Use of Knowledge in Society*, 35 AM. ECON. REV. 519, 519 (1945)).

162. Super, *supra* note 68, at 1380.

163. *Id.* at 1379.

164. *Id.* at 1409–10.

165. *Id.* at 1401 ("We avoid discussing deficiencies in decisional or enforcement capacity as they embarrass the law . . . .").

166. *Id.* at 1401–02.

167. *Id.* (citing BARRY SCHWARTZ, THE PARADOX OF CHOICE: WHY MORE IS LESS 23 (2004) (noting that information overload can paralyze decision making)).

168. 2011 Mobile Medical Applications Draft, *supra* note 97, at 5.

169. Super, *supra* note 68, at 1381 (using as case studies the Federal Emergency Management Agency's response to Hurricane Katrina in 2005).

170. *Id.* at 1381.

171. *Id.* at 1405.

172. *Id.*

regulation necessarily stifles innovation does not always hold.[173] Indeed, an industry coalition of app developers and device makers opposed an effort in Congress to prohibit the FDA from finalizing a draft guidance on mobile software for eighteen months.[174] Industry can benefit from early, clear regulation.

The obvious exceptions are when the early regulatory intervention is deeply flawed, or when no intervention is needed, either because the problem is not severe enough or because it occurs with low frequency.[175] But from the agency's perspective, early decisions can capitalize on a motivated regulator's early passion for and attention to the issue,[176] which may wane over time absent an external shock. Typically, the shock is some tragedy followed by Congressional intervention forcing the agency to respond.[177] And such reactionary regulation is likely to be more severe and potentially miscalibrated.[178]

Early interventions may also benefit from a more objective regulatory atmosphere, before parties become entrenched and adversarial. In contrast, deferring action (usually in the name of preserving discretion and gathering information), often leads to incremental decision making, which is more susceptible to interest group influence.[179] Agencies that precommit to policies earlier can better avoid bias and capture, emulating (very crudely) a Rawlsian "veil of ignorance."[180] Although regulated firms generally prefer certainty,

173. This sentiment was echoed during a recent public workshop on FDA regulation of the latest generation of device software. *FDA Public Workshop*, *supra* note 80 (statement of Dean Kross, M.D.) (citing a decade of experience with unregulated electronic health records (EHRs) and prescribing software, which produced "ancient, hostile" applications with very little transparency as to how they operate). Gervais seems particularly wedded to the idea that there is an inverse correlation between regulation and beneficial innovation. Gervais, *supra* note 11, at 668–69.

174. Letter from mHealth Regulatory Coalition to Senators Enzi and Harkin, *supra* note 82.

175. Super, *supra* note 68, at 1405–06 (citing Louis Kaplow, *An Economic Analysis of Legal Transitions*, 99 HARV. L. REV. 509, 593 (1986) (favoring ex post rather than ex ante resolution with low frequency problems)).

176. Gersen & Posner, *supra* note 150, at 566.

177. Efthimios Parasidis, *Patients over Politics: Addressing Legislative Failure in the Regulation of Medical Products*, 2011 WISC. L. REV. 929, 937–49 (2011) (detailing the history of congressional reactions granting FDA more authority over drugs in the wake of tragedies).

178. *See* Super, *supra* note 68, at 1451.

179. Colin S. Diver, *Policymaking Paradigms in Administrative Law*, 95 HARV. L. REV. 393, 424–25 (1981) (noting that incremental policymaking is susceptible to interest groups and can sacrifice the agency's objectivity); Super, *supra* note 68, at 1418–19.

180. Gersen & Posner, *supra* note 146, at 572–73; Super, *supra* note 68, at 1378 (citing JOHN RAWLS, A THEORY OF JUSTICE 136–37 (1971)). A "veil of ignorance" generally "suppresses self-interested behavior on the part of decisionmakers . . . by subjecting the

and thus should prefer definite rules, current efforts tend to be aimed primarily at weakening agency oversight.[181] Delay may invite obstruction.

Of course, early intervention imposes costs on industry—costs that may be too heavy for fledgling industries to bear. And regulators should be aware of legal "transitions" that upset settled expectations.[182] The government should remain sensitive (as it has been under the Obama administration)[183] to imposing new costs on firms.

That said, recent experience with other new technologies illustrates the consequences of unregulated medical innovation.[184] For example, assisted reproductive technologies ("ART") like artificial embryo implantation proliferated in the 1980s and 1990s without any real FDA oversight, and despite the lack of clinical evidence that these interventions were safe and effective.[185] In the meantime, unproven reproductive technologies gave rise to a significant industry, and "questions quickly arose about the accuracy of promotional claims."[186] Arguably, neither tort law, state oversight, nor professional self-regulation filled in.[187] A similar story describes the explosive growth of the dietary supplement industry, which also advanced with very

decisionmakers to uncertainty about the distribution of benefits and burdens that will result from the decision." Adrian Vermeule, *Veil of Ignorance Rules in Constitutional Law*, 111 YALE L.J. 399, 399 (2001).

181. Super, *supra* note 68, at 1428.

182. Barbara H. Fried, *Ex Ante/Ex Post*, 13 J. CONTEMP. LEGAL ISSUES 128 (2003); Huber, *supra* note 10, at 1064 (noting that "transition costs are largely absent when new products are regulated"); Kaplow, *supra* note 175.

183. *See, e.g.*, Exec. Order No. 13,563, Improving Regulation and Regulatory Review, 76 Fed. Reg. 3821, 3822 (Jan. 21, 2011) (calling for federal agencies to identify ways to achieve regulatory goals "that are designed to promote innovation," promote regulatory flexibility, and reduce unnecessary regulatory burdens).

184. *See, e.g.*, Noah, *supra* note 148.

185. *See id.*

186. *Id.* at 614.

187. *Id.* at 648. Eventually, the FDA asserted jurisdiction over some ART technologies by proposing rules, and sent warning letters with notices of violations. *See* Current Good Tissue Practice for Manufacturers of Human Cellular and Tissue-Based Products, 66 Fed. Reg. 1508 (2001); Establishment Registration and Listing for Manufacturers of Human Cellular and Tissue-Based Products, 63 Fed. Reg. 26,744 (1998); Noah, *supra* note 148, at 650–51. It is also contested whether tort law and self-regulation by the American Society for Reproductive Medicine ("ASRM") did indeed fill the regulatory void. *See, e.g.*, Kimberly M. Mutcherson, *Welcome to the Wild West: Protecting Access to Cross Border Fertility Care in the United States*, 22 CORNELL J.L. & PUB. POL'Y 349, 361–62 (2012) (discussing the United States' "not completely deserved reputation as the Wild West of fertility treatment").

little meaningful oversight[188] and with very little evidence to support its claims.[189] As noted, waiting does not necessarily generate superior regulation.

B. FORM

A parallel decision for regulators confronting new technologies is form. Agencies are unique among government institutions in having several choices of policymaking form.[190] For years, scholars have debated whether agencies should make policy ex ante via rulemaking or ex post via adjudication.[191] The debate used to weigh these two primary modes, until rulemaking began to predominate in the 1970s.[192] By the early 1990s, the debate had shifted to rulemaking versus guidance, just as agencies began to rely on guidances even more than rules.[193]

Like most agencies, the FDA is statutorily authorized to choose among these forms.[194] Agencies enjoy significant discretion, as this choice is virtually immune from judicial review.[195] This Section evaluates the three primary forms of policymaking (rulemaking, adjudication, and guidance), as well as newer forms, and whether these forms matter when confronting new

188. The lack of oversight of dietary supplements was not due to the FDA dragging its feet, but to Congress tying the FDA's hands. *See Symposium: The Dietary Supplement Health and Education Act: Regulation at a Crossroads*, 31 AM. J.L. & MED. 147–364 (2005) (providing a comprehensive overview of how Congress prevented FDA action).

189. There have been many takedowns of the dietary supplement industry. For a recent survey of evidence (especially the lack thereof), see PAUL A. OFFIT, DO YOU BELIEVE IN MAGIC? THE SENSE AND NONSENSE OF ALTERNATIVE MEDICINE (2013) (finding that only four of the 51,000 supplements on the U.S. market are supported by valid scientific evidence).

190. Magill, *supra* note 55.

191. *See, e.g.*, Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557, 562–63 (1992) (wrestling with whether decision making via enforcement and adjudication are more efficient than rulemaking ex ante); Magill, *supra* note 55, at 1403 n.69 (citing two decades of literature on the subject); Super, *supra* note 68, at 1411–12.

192. Magill, *supra* note 55, at 1398–99 (noting that after Congress passed the APA in 1946, most agencies carried out their statutory obligations by adjudication in the 1950s and 1960s, shifting decidedly towards broad-based rulemaking in the 1970s).

193. Mendelson, *supra* note 144, at 398 (noting that the volume of guidances is "massive" compared to rules). Aware of this trend and suspicious of abuse, the House Committee on Oversight and Government Reform held hearings and issued a report in 2000. *See* H.R. REP. NO. 106-1009 (2000). In response, the EPA revealed that it had issued over 2000 guidances between 1996 and 1999, and OSHA revealed over 3,000 during the same period. In comparison, the EPA issued only 100 "significant" rules and OSHA twenty during that same period. Mendelson, *supra* note 144, at 399.

194. Magill, *supra* note 55, at 1388 n.11 (citing, for example, 21 U.S.C. §§ 334, 371(a) (2000)).

195. *Id.* at 1405–42 (discussing the germinal case, *SEC v. Chenery Corp.*, 332 U.S. 194 (1947)).

technologies. I argue that, in retrospect, the FDA's reliance on guidance weakened its oversight of software, and that it now risks repeating its mistake with the next generation of software.

1. Rulemaking

The FDA could have proceeded by rule. The benefits of rules are many,[196] although enthusiasm for them long ago waned. Rules can set clear, authoritative, and durable requirements. The rulemaking process uses robust, transparent procedures that accommodate broad public participation.[197] The rulemaking process is designed to ventilate issues of broad public concern and is particularly suitable when the problem is widespread and foreseeable.[198] As such, rulemaking can be more efficient than making policy via individualized adjudication, as federal agencies discovered in the 1960s and 1970s.[199] Rules are widely published and are more easily accessible than adjudications, although this advantage is slighter in an era of agency websites.[200] Rules are also subject to more oversight by politically accountable institutions, like Congress and the President. But above all, rules bind. Agencies can enforce them. This was the chief limitation of the FDA's reliance on guidance.

Of course, the disadvantages of rulemaking are also abundant, as well canvassed by scholars. The most familiar is that promulgating rules requires a long, lumbering process[201]—one that is almost universally condemned as being ossified and prolonged by cumulative procedural burdens,[202] including executive review and various statutory hurdles.[203] Studies confirm this

196. *See generally* Arthur Earl Bonfield, *State Administrative Policy Formulation and the Choice of Lawmaking Methodology*, 42 ADMIN. L. REV. 121, 122–36 (1990) (discussing the benefits of rulemaking vis-à-vis adjudication).

197. Rulemaking by federal agencies is governed by § 553 of the Administrative Procedure Act (APA), 5 U.S.C. § 553 (2012).

198. NLRB v. Bell Aerospace Co., 416 U.S. 267, 292–94 (1974) (discussing when a legislative rule would be appropriate in reviewing agency choices for an abuse of discretion); *cf.* Magill, *supra* note 55, at 1408.

199. Antonin Scalia, *Vermont Yankee: The APA, the D.C. Circuit, and the Supreme Court*, 1978 SUP. CT. REV. 345, 376 (1978).

200. Cortez, *Adverse Publicity by Administrative Agencies in the Internet Era*, *supra* note 60, at 1374.

201. *See* 5 U.S.C. § 553.

202. Magill, *supra* note 55, at 1390–91, 1391 n.17 (citing the "large literature on the 'ossification' of the rulemaking process"). Mark Seidenfeld once counted 109 discrete rulemaking requirements. Mark Seidenfeld, *A Table of Requirements for Federal Administrative Rulemaking*, 27 FLA. ST. U. L. REV. 533 (2000).

203. There are well-known executive burdens that apply to most rulemaking, the most recognized being review by the Office of Information and Regulatory Affairs ("OIRA") in the Office of Management and Budget ("OMB"). *See, e.g.*, Exec. Order No. 12,291, Federal

critique, finding that rulemaking at some agencies can take upwards of two years to complete.[204] At the Department of Health and Human Services ("HHS"), for example, the average duration of rulemaking without a statutorily imposed deadline was a staggering 817 days.[205]

Rules are also critiqued for being inflexible. They are hard to update, as doing so requires another round of notice and comment.[206] Rules also can be overinclusive or underinclusive, and thus unjust.[207] And like statutes, rules are prospective, so they must be general enough to accommodate unforeseen conduct.[208] These factors undoubtedly weigh on agency decisions to forgo rulemaking.

2. Adjudication

The FDA also could have established policy via adjudication, which has its own pros and cons. Among the pros, adjudication allows agencies to address problems in discrete, concrete circumstances, rather than prospectively and abstractly, as with rulemaking.[209] Agencies can use

Regulation, 3 C.F.R. 127 (1982); Exec. Order No. 12,866, Regulatory Planning and Review, 3 C.F.R. 638, 638, 642, 644 (1994) (§ 4 "Planning Mechanism" and § 6 "Centralized Review of Regulations"). And as many authors observe, there are lesser-known rulemaking requirements imposed by statute. Erica Seiguer & John J. Smith, *Perception and Process at the Food and Drug Administration: Obligations and Trade-Offs in Rules and Guidances*, 60 FOOD & DRUG L.J. 17, 19 (2005) ("In addition to the APA, there are a number of laws that impact the development of rules . . . . [A] variety of Executive Orders . . . also apply."); Connor N. Raso, Note, *Strategic or Sincere? Analyzing Agency Use of Guidance Documents*, 119 YALE L.J. 782, 788, 788 n.16 (2010) ("Agencies must also complete a number of lesser-known procedural requirements before issuing a legislative rule."); *e.g.*, Unfunded Mandates Reform Act of 1995, 2 U.S.C. §§ 1501–71, 1532–35 (2012); Small Business Regulatory Enforcement Fairness Act of 1996, Pub. L. No. 104-121, 110 Stat. 857 (codified as amended in scattered sections of 5 U.S.C. and 15 U.S.C.); Regulatory Flexibility Act, 5 U.S.C. §§ 601–612 (2012); Congressional Review Act, 5 U.S.C. §§ 801–808 (2012); Federal Register Act, 44 U.S.C. § 1505 (2012); Paperwork Reduction Act of 1995, 44 U.S.C. §§ 3501–3549 (2012); Small Business Paperwork Relief Act of 2002, 44 U.S.C. §§ 3504(c), 3506(i), 3520 (2012).

204. Jacob E. Gersen & Anne Joseph O'Connell, *Deadlines in Administrative Law*, 156 U. PA. L. REV. 923, 988 tbl.12 (2008) (studying the period between 1988 and 2003, and noting rulemaking actions without congressionally imposed deadlines).

205. *Id.*

206. *See* Mendelson, *supra* note 144, at 410.

207. Mark Seidenfeld, *Bending the Rules: Flexible Regulation and Constraints on Agency Discretion*, 51 ADMIN. L. REV. 429, 440–41 (1999); Richard B. Stewart, *Administrative Law in the Twenty-First Century*, 78 N.Y.U. L. REV. 437, 446 (2003) (noting that centralized, command-and-control regulation "suffers from the inherent problems involved in attempting to dictate the conduct of millions of actors in a quickly changing and very complex economy and society throughout a large and diverse nation."); *see also* Mendelson, *supra* note 144, at 438.

208. Mendelson, *supra* note 144, at 410.

209. Bonfield, *supra* note 196, at 122–36.

adjudication to respond quickly to problematic conduct as it arises.[210] Adjudication also allows agencies to create policy incrementally, in the absence of consensus.[211] Another benefit is that agencies can also target egregious, outlier conduct that does not occur often enough to require broad-based rulemaking.[212] And if the conduct of individual firms differs in key ways, adjudication can account for that.[213]

But like rulemaking, adjudication has well-known deficiencies. Adjudication does not invite broad public participation, so it can be awkward for establishing broader policies.[214] Agencies can choose sympathetic targets for adjudication, whose conduct may not fairly represent industry practices.[215] Any precedents established in adjudication can be limited by facts peculiar to the case.[216] Moreover, precedents technically bind only the parties.[217] Thus, agencies pushing new legal theories or interpretations face problems with fair notice, and may find it difficult to enforce the new position.[218] Formal administrative adjudications resemble judicial trials, with all their inefficiencies.[219] And agencies wishing to enforce their adjudications in court must convince the Department of Justice to take up their cause.[220] Adjudication may also be inefficient vis-à-vis rulemaking when the agency has to relitigate the issue repeatedly.

Yet the familiar dichotomy between rulemaking and adjudication can be a false one. The two must coexist if agencies wish to enforce rules and clarify them in individualized applications.[221] Rules can be toothless without enforcement, and enforcement requires something to enforce. Guidances can depart from this symbiosis.

3. Guidances

When the FDA first announced a software policy in 1987, it chose to make policy via guidance—a pattern it continues today. For agencies, the major attraction of guidances is that guidances need not comply with the

210. Magill, *supra* note 55, at 1396–97.
211. *Id.*
212. *Id.*
213. *Id.* at 1444–45.
214. Magill, *supra* note 55, at 1396.
215. *Id.*
216. *Id.*
217. *Id.*
218. Mendelson, *supra* note 144, at 436.
219. 5 U.S.C. §§ 556–557 (2012) (including the APA's procedures for formal adjudication and rulemaking); Magill, *supra* note 55, at 1391.
220. Magill, *supra* note 55, at 1391, 1393.
221. Bonfield, *supra* note 196, at 122–36.

notice-and-comment rulemaking procedures required by the APA[222] or any other statute.[223] Publishing guidances is quicker than even most forms of adjudication.[224] Compared to rulemaking, guidances are a very economical way to signal agency preferences.[225] They give agencies greater flexibility to update or retreat from the policy when necessary.[226] Guidances are useful for coordinating lower level agency personnel.[227] Agencies may also prefer guidances because they are virtually immune from pre-enforcement judicial review and are subject to less congressional oversight.[228] Finally, guidances are generally credible,[229] and they can be "sticky" like regulations.[230]

Agencies are known to use guidances to clarify highly technical or scientific requirements that would be difficult to update via rulemaking or

222. 5 U.S.C. § 553 (2012). In the early 1980s, Congress considered bills that would have required agencies to use notice-and-comment procedures for many guidances, but none of these bills passed. *See* H.R. 2327, 98th Cong. (1983); H.R. 220, 98th Cong. (1983); S. 1080, 97th Cong. (1982); H.R. 746, 97th Cong. (1981); Mendelson, *supra* note 144, at 401 & n.25.

223. Note, however, that even the guidance document process is becoming burdened by procedure. In 2007, the Bush administration published Executive Order 13,422, which allows the OMB to require agencies to consult with it before issuing a "significant" guidance document, defined as one that has an economic impact of more than $100 million in any single year, or raises important legal issues (among other criteria). Exec. Order No. 13,422, 3 C.F.R. 191–92 (2007) (rescinded by President Obama's Exec. Order. No. 13,497, 3 C.F.R. 218 (2013)).

224. *See* Magill, *supra* note 55, at 1391–92 (describing developing guidances as "relatively cheap[]" for agencies in comparison to adjudication and formal rulemaking).

225. Michael Asimow, *Nonlegislative Rulemaking and Regulatory Reform*, 1985 DUKE L.J. 381, 404–08 (1985) (arguing that agencies rely on guidance documents when budgetary pressures require it); Mendelson, *supra* note 144, at 408. In the words of one FDA official, "In an ideal world, we would always do rulemaking, but it is not as responsive and there is a lot of process involved, not only internal to the agency, but outside FDA." Seiguer & Smith, *supra* note 203, at 24. Another FDA official said, "To do a rule, it's a huge ordeal . . . there are economic analyses of the impact, notice and comment, involvement of the OMB, etc." *Id.*; *see also* Gersen & Posner, *supra* note 59, at 588–89, 594–95 (describing the similar allure of soft law statutes by Congress).

226. Mendelson, *supra* note 144, at 408.

227. Peter L. Strauss, *Publication Rules in the Rulemaking Spectrum: Assuring Proper Respect for an Essential Element*, 53 ADMIN. L. REV. 803, 808 (2001).

228. Magill, *supra* note 55, at 1395; Mendelson, *supra* note 144, at 408.

229. *See* Gersen & Posner, *supra* note 59, at 589 (noting that soft law can be credible even when it is not costly to speak).

230. The rulemaking process has been criticized as being inefficient and "sticky." *See, e.g.*, Jody Freeman, *Collaborative Governance in the Administrative State*, 45 U.C.L.A. L. REV. 1, 3 (1997); Jason M. Solomon, *Law and Governance in the 21st Century Regulatory State*, 86 TEX. L. REV. 819, 822 (2008) (book review) (discussing the "stickiness of the rulemaking process"). But stickiness may be good, if the agency wants the guidance to endure.

adjudication.[231] These issues may require "a scalpel, not a sledgehammer."[232] Indeed, there is almost a consensus in the literature that guidances are more appropriate than rulemaking to address rapid technological and scientific advancement, as Wu argues.[233] In a recent study, representatives from both the FDA and regulated industry agreed on this use of guidances.[234]

But for all the glowing evaluations, relying on guidance does pose problems. Many scholars suspect that agencies abuse guidance as a way to skirt procedure,[235] though a recent study disputes this.[236] Agency consideration may not be disciplined the same way as when subject to the procedures governing rulemaking and adjudication[237]—although in FDA's case its consideration is tempered by its Good Guidance Practices, and for most agencies by executive review protocols.[238] Guidances are hailed for their flexibility, yet one study found that they are updated with even less frequency than rules.[239] And if agencies use guidances to address scientific and technical issues, they might be doing so at the expense of public input.

The major shortcoming is that guidance is not enforceable. Agencies cannot allege violations of guidance documents or initiate enforcement actions based solely on them.[240] Defenders of guidance tend to emphasize its

231. Lisa Schultz Bressman, *Beyond Accountability: Arbitrariness and Legitimacy in the Administrative State*, 78 N.Y.U. L. REV. 461, 545–56 (2003); Raso, *supra* note 203, at 815 (citing as examples the Department of Transportation, EPA, and OSHA).

232. Philip J. Weiser, *The Next Frontier for Network Neutrality*, 60 ADMIN. L. REV. 273, 277 (2008).

233. *See* Wu, *supra* note 11, at 1842, 1848–54.

234. *See, e.g.*, Seiguer & Smith, *supra* note 203, at 23. Seiguer and Smith interviewed several representatives from the FDA and industry, and found that "[t]hose interviewed agreed that due to the rapid pace of scientific advancement, guidances . . . provide the best means of providing information to assist industry in understanding and complying with regulatory requirements." In fact, they noted that a common theme among interviewees was "[t]he importance of guidances in an era of complex science." *Id.*

235. Raso, *supra* note 203, at 785–87.

236. *Id.* at 809–23 (reviewing five agencies (EPA, FCC, FDA, IRS, and OSHA) over ten years (1996–2006) and finding that agencies generally do not use guidance documents to avoid rulemaking, at least for "significant" policies). Jacob Gersen and Eric Posner discuss some of the same "rule of law" objections to soft law by Congress. Gersen & Posner, *supra* note 59, at 597–98. Others have proposed procedural reforms to legitimate guidances. *See, e.g.*, Jessica Mantel, *Procedural Safeguards for Agency Guidance: A Source of Legitimacy for the Administrative State*, 61 ADMIN. L. REV. 343 (2009).

237. Magill, *supra* note 55, at 1397, 1446.

238. *See* 21 U.S.C. § 371(h) (2012); FDA Good Guidance Practices, 21 C.F.R. § 10.115 (2013).

239. Raso, *supra* note 203, at 818–19 (noting, however, the limitations of the available data).

240. William Funk, *When is a "Rule" a Regulation? Marking a Clear Line Between Nonlegislative Rules and Legislative Rules*, 54 ADMIN. L. REV. 659, 661–62 (2002); Magill, *supra* note 55, at

benefits to industry without contemplating how it can disadvantage regulatory beneficiaries.[241] Even an open, transparent guidance process like the FDA's can skew heavily towards industry, per Nina Mendelson's critique that agencies generally solicit comments on guidance from industry, but not from regulatory beneficiaries.[242]

Notwithstanding these criticisms, most scholars agree that some guidance is better than none.[243] But it is far from clear that doing nothing is the only counterfactual, or that guidance is always preferable to rulemaking, particularly when enforcement matters.[244] Agencies like the FDA still choose to promulgate rules when they desire enforceability.[245]

In many ways, the FDA's reliance on guidance is notable. The FDA is a serial user, dating back to 1902, when its precursor first used them.[246] Today, the FDA issues roughly twice as many guidances as it does rules,[247] and the ratio has increased steadily since the 1970s.[248] By the 1990s, the FDA was generating proportionally fewer regulations and more guidances, such that guidances became the agency's primary mode of policymaking.[249] By the late

1394; Mendelson, *supra* note 144, at 406–07, 410. Indeed, courts have prohibited agencies from trying to bind regulated firms via guidance documents. *See, e.g.*, Appalachian Power Co. v. EPA, 208 F.3d 1015 (D.C. Cir. 2000); McLouth Steel Prods. Corp. v. Thomas, 838 F.2d 1317 (D.C. Cir. 1988); Am. Hosp. Ass'n v. Bowen, 834 F.2d 1037, 1045 (D.C. Cir. 1987); Cmty. Nutrition Inst. v. Young, 818 F.2d 943 (D.C. Cir. 1987).

241. *See generally* Mendelson, *supra* note 144, at 402, 413–14 (noting that commentators tend to defend guidances by arguing that they are better than nothing at signaling regulators' intent and help keep the conduct of agencies' employees consistent).

242. *Id.* at 427–29 (noting the general tendency of agencies to "float" draft guidances to industry contacts and solicit comments from them but not the public, and describing the FAA's quite disturbing practice of accepting comments on guidance documents only from a list of seventeen industry organizations, which does not include passenger or consumer safety groups).

243. *See, e.g.*, Michael Asimow, *Guidance Documents in the States: Toward a Safe Harbor*, 54 ADMIN. L. REV. 631, 647 (2002); Asimow, *supra* note 225, at 381; Mendelson, *supra* note 144, at 413.

244. Seiguer & Smith, *supra* note 203, at 28.

245. *Id.*

246. K.M. Lewis, *Informal Guidance and the FDA*, 66 FOOD & DRUG L.J. 507, 509 (2011).

247. Seiguer & Smith, *supra* note 203, at 25–26 (finding that during 2001, 2002, and most of 2003, the FDA issued well more than twice as many guidances (298) as rules (129)).

248. Todd D. Rakoff, *The Choice Between Formal and Informal Modes of Administrative Regulation*, 52 ADMIN. L. REV. 159, 168 (2000) (noting the marked increase in the ratio of guidance documents to rules from the 1970s through the 1990s).

249. Lewis, *supra* note 246, at 520. In fact, the FDA center responsible for regulating medical devices, the Center for Devices and Radiological Health ("CDRH"), has relied far more on guidances than other centers in the agency. Between 1975 and 2009, CDRH published 659 guidances, compared to 507 by the drug center ("CDER"), 294 by the biologics center ("CBER"), and 190 by the food center ("CFSCAN"). *Id.* at 549–50 fig.5.

1990s, observers were already wondering whether rulemaking at the FDA had become an artifact from earlier ages.[250]

The FDA is also notable for being the only agency whose guidance practices are governed by statute and regulation.[251] The FDA initially proposed its own "Good Guidance Practices" by regulation in 1992.[252] Congress then codified the regulation by statute in 1997.[253] Together the two documents require the FDA to solicit public input and use notice-and-comment-like procedures when creating guidances.[254] Importantly, the statute requires that FDA staff generally follow the agency's own guidances.[255]

Yet, at the end of the day, even FDA guidances cannot bind as a matter of law[256] (though some report that individual FDA staff sometimes say otherwise).[257] When push comes to shove in court, the FDA must point to statutes and regulations, not guidance. Any paths blazed via guidance remain legally quarantined until the agency takes action to "codify" them.

4. New Forms?

We are undoubtedly in a new era of policymaking pluralism, so the menu for agencies like the FDA is more varied. The traditional substrata of rulemaking, adjudication, and guidance now support variations like negotiated rulemaking, waiver systems, and other more flexible, collaborative forms that are often called regulatory "experimentalism" or "new governance."[258] Agencies that struggle with limited, stagnant resources are

250. John C. Carey, *Is Rulemaking Old Medicine at the FDA?*, *in* FOOD AND DRUG LAW: AN ELECTRONIC BOOK OF STUDENT PAPERS 3–6 (Peter Barton Hutt ed., 2012), http://www.law.harvard.edu/faculty/hutt/book_index.html (last updated Jan. 10, 2013).

251. *See* 21 U.S.C. § 371(h) (2012); FDA Good Guidance Practices, 21 C.F.R. § 10.115 (2013).

252. 57 Fed. Reg. 47,314 (Oct. 15, 1992). The FDA finalized its Good Guidance Practices in 1997. 62 Fed. Reg. 8961 (Feb. 27, 1997).

253. Food and Drug Administration Modernization Act, Pub. L. No. 105-115 § 405, 111 Stat. 2296, 2368–69 (codified at 21 U.S.C. § 371(h)); *see also* Lars Noah, *FDA's New Policy on Guidelines: Having Your Cake and Eating it Too*, 47 CATH. U. L. REV. 113 (1997) (discussing the FDA's good guidance practices).

254. 21 U.S.C. § 371(h); 21 C.F.R. § 10.115; *see also* 65 Fed. Reg. 56,468, 56,473 (Sept. 19, 2000) (addressing public concerns regarding the respective roles of guidances versus notice-and-comment rulemaking); *id.* at 56,477–78 (describing comment procedures under 21 C.F.R. § 10.115 resembling the notice-and-comment procedures involved in rulemaking under the APA). Some doubt that public comments on guidance documents mean much to the FDA. Seiguer and Smith, *supra* note 203, at 30.

255. 21 U.S.C. § 371(h).

256. 21 U.S.C. §§ 371(h)(1)(A), (h)(2).

257. Seiguer & Smith, *supra* note 203, at 30.

258. Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COLUM. L. REV. 267 (1998); Freeman, *supra* note 230; Magill, *supra* note 55, at 1398–99;

increasingly looking to alternatives.[259] The recent scholarly trend is to champion nimble agencies that practice regulatory experimentalism in response to changing circumstances.[260] Even the hoary FDA drug approval process promises to benefit from experimentalism, argue proponents of new governance like Charles Sabel and Bill Simon.[261]

Although the FDA did not explicitly choose an experimental method with software, one cannot help but sense from the agency a collaborative mood in some of its public hearings (during which FDA comments were more than solicitous of industry and, somewhat asymmetrically, the industry was not shy about voicing displeasure with the FDA's approach).[262] This highlights how regulatory experimentalism can take for granted the "adversarial legalism" of our regulatory system—our tendency to formally contest policymaking.[263] Regardless of the form policymaking takes, industry will find ways to contest it.

5. Does Form Matter?

It is possible that the differences among policymaking forms are overstated. For example, as Nina Mendelson notes, guidances frequently

David Zaring, *Best Practices*, 81 N.Y.U. L. REV. 294 (2006). Sabel and Simon, in fact, point to quality regulation of food manufacturing as a paragon of regulatory experimentalism. Charles F. Sabel & William H. Simon, *Minimalism and Experimentalism in the Administrative State*, 100 GEO. L.J. 53, 55, 78 (2011). The Food Safety and Modernization Act of 2010 requires the FDA to require processors to use a flexible, customizable cycle of hazard detection and corrective and preventative actions. The process is familiarly known in FDA circles as Hazard Analysis and Critical Control Points ("HACCP"), but was relabeled in the 2010 legislation as Hazard Analysis and Preventive Controls. FDA Food Safety Modernization Act of 2010, Pub. L. No. 111-353, §§ 103, 201, 307, 124 Stat. 3885, 3889–90, 3923, 3962 (codified in scattered sections of 21 U.S.C.).

259. *See* Weiser, *The Future of Internet Regulation*, *supra* note 39, at 557 n.106 (quoting former FTC Chairman Robert Pitofsky).

260. *See, e.g.*, Sabel & Simon, *supra* note 258, at 56, 78 ("Experimentalism takes its name from John Dewey's political philosophy, which aims to precisely accommodate the continuous change and variation that we see as the most pervasive challenge of current public problems.").

261. *Id.* at 87–88 (discussing variations on the binary approve/reject decision for FDA reviewers of new drug approvals ("NDAs")).

262. *See FDA Public Workshop*, *supra* note 80 (containing transcripts and webcasts of the proceedings).

263. *See* Solomon, *supra* note 230, at 821, 847–51 (reviewing two examples of how scholars are rethinking the U.S. regulatory state and concluding that both pay inadequate attention to the "adversarial legalism" described by Robert Kagan). *See generally* ROBERT A. KAGAN, ADVERSARIAL LEGALISM: THE AMERICAN WAY OF LAW (2001). Interestingly, moving beyond adversarial legalism would probably require agencies to have more discretion, more finality, and less legal friction. Solomon, *supra* note 230, at 848.

resemble rules, which is why they are often called "nonlegislative rules."[264] Indeed, courts review whether agencies treat nonlegislative rules as having the force of law.[265] Scholars have long recognized that finding meaningful doctrinal distinctions between legislative and nonlegislative rules is a task "enshrouded in considerable smog."[266]

Moreover, regulated parties may not appreciate the limited legal status of guidances,[267] as they generally comply with them anyway.[268] Most observers note that industries choose to treat guidances as binding for practical purposes,[269] even after Congress pressured agencies to more clearly label their guidances as such and disclaim any legally binding effect.[270] The FTC even declares (in the Code of Federal Regulations, no less) that it may initiate enforcement actions against parties that do not comply with its guidances.[271]

264. Mendelson, *supra* note 144, at 399.

265. *See, e.g.*, Am. Mining Cong. v. Mine Safety & Health Admin., 995 F.2d 1106, 1109 (D.C. Cir. 1993).

266. David L. Franklin, *Legislative Rules, Nonlegislative Rules, and the Perils of the Short Cut*, 120 YALE L.J. 276, 286–87 (2010) (citing both courts and scholars that have struggled with the distinction).

267. In fairness to regulated industry, the legal status of FDA guidance documents has always been confusing. For example, in 1977, the FDA promulgated a regulation binding itself with officially published advisory opinions and guidelines. 42 Fed. Reg. 4680, 4708–10 (1977). Although this stance was ostensibly superseded by the agency's Good Guidance Practices, 21 C.F.R. § 10.115 (declaring that guidances are nonbinding at § 10.115(i)(i), (i)(iv)), there remains some confusion. *See, e.g.*, Lewis, *supra* note 246.

268. Mendelson argues that regulated parties comply with guidance documents rather than bear the expense, uncertainty, and acrimony of challenging them in court. Mendelson, *supra* note 144, at 407–08. This does not address guidance documents that are un- or under-enforced by the agency, which parties would have little incentive to challenge as they age and become stale. Nevertheless, Seiguer and Smith found that "in practice, most of those interviewed [from the FDA and regulated industries] said that industry treats guidances no differently than rules." Seiguer & Smith, *supra* note 203, at 29–30 (noting that this response reflected "the majority of those interviewed," and emphasizing that "industry is loathe to diverge from the agency's current thinking embodied in the guidance").

269. Mendelson, *supra* note 144, at 407–08; Seiguer & Smith, *supra* note 203, at 29–30.

270. *See* COMM. ON GOV'T REFORM, NON-BINDING LEGAL EFFECT OF AGENCY GUIDANCE DOCUMENTS, H.R. REP. NO. 106-1009, at 8–9 (2000); Mendelson, *supra* note 144, at 400 & n.17 (noting that the Committee's critiques "probably led the agencies to add disclaimers more systematically").

271. Of course, the FTC is careful to note that such enforcement would rely on statutory violations. 16 C.F.R. § 1.5 (2013) ("Failure to comply with the guides may result in corrective action by the Commission under applicable statutory provisions. Guides may relate to a practice common to many industries or to specific practices of a particular industry.").

As with rules, the APA requires agencies to publish most guidance documents in the Federal Register or in some other accessible source.[272] Sometimes Congress even requires agencies by statute to publish guidances.[273] As a result, many agencies, particularly the FDA, find themselves dedicating substantial time and resources to guidances.[274]

As Elizabeth Magill suggests, we might be more concerned with the outcome of agency action than "the package in which it is wrapped."[275] But Magill also notes that policymaking form dictates important things like its scope, the parties' procedural rights, enforceability, and judicial review.[276]

Moreover, the intuition that guidances give industry timely notice of new requirements while rulemaking does not is probably overwrought. Draft guidances are sometimes never finalized and agencies propose rules far before they are finalized.[277] The regulatory uncertainty argument thus rings hollow as to form. In fact, both the FDA and industry representatives observed that the agency often takes a long time to articulate a concrete position on new issues, regardless of whether it chooses rulemaking or guidance.[278] The FDA's draft guidances are known to linger for years before being finalized.[279] Many morph into de facto final guidances,[280] as the 1989 Draft Software Policy did.

Finally, differences in form may be overstated because they are not mutually exclusive. The reality is that "[m]any agencies regularly employ a mix of policymaking tools on a given issue—sometimes promulgating or amending a rule, sometimes bringing an enforcement action, and sometimes issuing a guidance document."[281] Threats can start informally before

272. *See* 5 U.S.C. § 552(a)(1)(D) (2012) (requiring publication of interpretive rules and policy statements in the Federal Register); § 552(a)(2)(b) (requiring public accessibility of important guidance documents that are not in the Federal Register).

273. *See, e.g.*, Raso, *supra* note 203, at 814 n.144 (citing the Small Business Regulatory Enforcement Fairness Act of 1996 § 212, 5 U.S.C. § 601 (2012), which requires agencies to issue compliance guides for small businesses on important rules).

274. *See* Seiguer & Smith, *supra* note 203, at 24 (quoting an anonymous interviewee as stating, "Ideally, [guidances] should be faster and more flexible, but in practice, they may take as long as rules to develop.").

275. Magill, *supra* note 55, at 1419 (explaining why courts might be reluctant to review the vehicles agencies choose to use for policymaking, ultimately deciding that form does matter).

276. *Id.* at 1420.

277. A recent commentary finds that, on balance, guidances benefit regulated industry, generalizing this point on regulatory certainty. *See* Lewis, *supra* note 246, at 541.

278. *See* Seiguer & Smith, *supra* note 203, at 31.

279. *See id.*

280. *See id.*

281. Magill, *supra* note 55, at 1410.

percolating into adjudication. For example, the FDA issues frequent "warning letters" to alleged regulatory violators. Warning letters are informal and are not considered official enforcement action.[282] But the industry treats FDA warning letters as precedent, and lawyers advise clients based on them. The FDA's warning letters all cite specific statutory or regulatory violations, which then form the basis of any subsequent judicial enforcement.

In the end, the choice of policymaking form may not be as stark as it once was. And one form does not preclude another. But the FDA's decision to address software by guidance, to the exclusion of rulemaking, necessarily weakened its enforcement posture.

C. DURABILITY

The third question for agencies confronting disruptive innovation is how durable to make their policies. A concern with rulemaking is that it might be *too* durable for evolving technologies, hence the intuitive appeal of informal "threats." Agencies can temper concerns over the timing and form of policies by manipulating how long the policy will endure.

Of course, a policy's durability will be dictated in large part by its form.[283] For example, we might be particularly concerned about the durability of guidances given the transience of agency leadership. Political appointees, who often set regulatory priorities, spend an average of only eighteen to twenty-four months with their agencies.[284] Transient appointees, not to mention transient administrations, can undermine agency guidances by not adhering to them or by simply deprioritizing them. Although rules are also subject to amendment and rescission, they tend to be more resilient.[285]

The trick is to craft enduring policy under high uncertainty.[286] Even large federal bureaucracies like the FDA cannot make perfectly calibrated decisions about new products or conduct.[287] So-called "timing rules," contemplated by Gersen and Posner in the legislative context,[288] might help.

---

282. The FDA includes "warning letters" in its definition of informal enforcement actions. 21 C.F.R. § 100.2(j)(1) (2013).

283. *See* Raso, *supra* note 203, at 803.

284. *Id.* at 803 (citing Paul C. Light, *Our Tottering Confirmation Process*, BROOKINGS (Spring 2002), http://www.brookings.edu/articles/2002/spring_governance_light.aspx).

285. *Id.*

286. *See generally* Fried, *supra* note 182, at 146–51 (discussing economic and empirical literature of decision making under uncertainty, weighing ex ante versus ex post decision making).

287. *See* Wu, *supra* note 11.

288. *See generally* Gersen & Posner, *supra* note 146.

Like legislators, regulators can manipulate the timing and duration of their interventions.[289] Options include mandatory deadlines, waiting periods, interim periods, phases, and sunsets.[290] These can control the timing of how laws are considered, enacted, and implemented, with the purpose of achieving some optimum result in the public interest.[291] For example, agencies frequently use interim final rules to impose new requirements immediately while extending public comment on them. Such tools can be most useful when operating under uncertainty, as they allow lawmakers to "economize on legislative costs, address problems quickly or enable citizens to adjust, and handle uncertainty about the effects of a legislative proposal."[292] The more inchoate the technology, the less "durable" the regulator might want its policy to be.[293]

Here, two particularly promising possibilities are sunsets and deadlines. Other possibilities include deferred or conditional rulemaking,[294] although these might only encourage further delay.[295]

1. Sunsets

Regulators can ease their anxiety over regulating new technologies prematurely or incorrectly by using sunsets—regulating temporarily. Congress frequently uses sunset clauses and other forms of temporary legislation that specify a finite duration.[296] Sunsets are useful when the benefits of intervening are uncertain, as they reduce the costs of errors and allow regulators to gather better information.[297] They can also counterbalance

289. *See id.* at 545 (defining a "timing rule" as "a rule that substantially affects the timing of government action, including legislation and executive action," and typically includes deadlines and mandatory delays).

290. *Id.*

291. *Id.* at 546.

292. *Id.* at 563.

293. *See* Gervais, *supra* note 11, at 701.

294. *See* Gersen & Posner, *supra* note 146, at 562 (describing deferred and conditional legislation).

295. Gersen and Posner detail, somewhat optimistically, four benefits of delay rules for administrative agencies: (i) they allow agencies to gather more information, (ii) they "might reduce the effect of deliberative pathologies" like polarization over an issue, (iii) they might allow agency executives to better oversee major initiatives, and (iv) they could mitigate interest group incentives to weaken the rule. *Id.* at 588. These benefits are suspect, as noted above.

296. *See* Jacob E. Gersen, *Temporary Legislation*, 74 U. CHI. L. REV. 247, 250 (2007) (refuting "the notion that temporary legislation is a new, peculiar, or particularly suspect legislative tool"); Gersen & Posner, *supra* note 146, at 562 n.78 (noting that sunset clauses date back to the First Congress).

297. *See* Gersen, *Temporary Legislation*, *supra* note 296, at 248, 266–78 (noting that "when initial decisions are likely to be wrong, staged decision procedures facilitate the correction of

cognitive biases in assessing uncertainty and overreacting to new risks.[298] Sunset clauses proliferated among states during the regulatory reform movement of the 1970s and 1980s, on the theory that they would encourage more legislative oversight and thus make agencies more efficient and more responsive.[299]

Despite their deregulatory roots, sunset clauses might be useful for addressing disruptive innovation. Temporary regulation would allow agencies like the FDA to impose *some* standards as the technology gestates and the agency learns more about it. Sunset regulation would impose binding, enforceable requirements in the meantime. As such, it could serve as a stopgap,[300] allowing us to capture the public benefits that we would otherwise forego by deferring regulation entirely.

Sunsets certainly are not a panacea. Like other variations on the regulatory process, they are susceptible to strategic behavior[301] and, as emphasized above, the wait for more information is not always worthwhile. But sunsets decrease the costs of premature or incorrect regulation by time-limiting the damage they can inflict. Moreover, the argument that temporary regulation would generate too much uncertainty for industry is tempered by the reality that sunset laws can always be reauthorized or extended, and that "permanent" regulation can always be amended or repealed.[302] Finally, wrapping the temporary policy in a regulation rather than a guidance would make it more durable.[303] Had the FDA taken this approach with software, it could have provided a much-needed regulatory backbone—one that would be enforceable but also amenable to occasional updates as computing advanced.

errors, and this is particularly likely to be the case in policy contexts dominated by uncertainty").

298. *See id.* at 269–72 (noting that "new policy initiatives are often enacted in the immediate aftermath of realized or recognized risks" and that "staged decision procedures are utilized as compensation mechanisms for conditions of uncertainty in many other fields," citing examples).

299. *Id.* at 259–60.

300. *See id.* at 273–74.

301. *See id.* at 275; LAWRENCE LESSIG, REPUBLIC, LOST: HOW MONEY CORRUPTS CONGRESS—AND A PLAN TO STOP IT 203–06 (2011) (critiquing sunset legislation in the tax code).

302. Gersen, *Temporary Legislation*, *supra* note 296, at 261.

303. Gersen addresses durability in light of public choice theory and the political calculations it can generate. *See id.* at 279–86.

2. Deadlines

To address the concern that an agency's initial restraint will calcify into a long-term laissez faire posture, Congress might set a regulatory deadline for the agency. Congress frequently does.[304]

The scant scholarship scrutinizing regulatory deadlines suggests that they have both advantages and limitations, particularly on the procedures agencies use, the extent of public participation, and the duration of the agency's action.[305] Their most obvious advantage is to confront persistent agency delay.[306] Deadlines can be desirable when "it is more important that a rule exist than it be right."[307] Congress has imposed deadlines on numerous agencies, particularly the EPA.[308] The most familiar FDA examples are the various statutory deadlines for approving new drug approval applications ("NDAs"),[309] which demonstrate the dynamics between the agency and the pharmaceutical industry.[310]

Congressional deadlines can counterbalance barriers to challenging agency delay or inaction.[311] Jacob Gersen and Anne O'Connell found that although deadlines speed up agency action only modestly, the impact on HHS rulemaking is remarkable, reducing the average length by more than forty percent (from 817 days to 445 days).[312] Procedurally, deadlines tend to correlate with interim final rules, which legally bind while the agency gathers public comments.[313] Deadlines also correlate with *more* comment periods,

304. *See* Gersen & Posner, *supra* note 146, at 545; Gersen & O'Connell, *supra* note 204.

305. Gersen & O'Connell, *supra* note 204, at 942–49.

306. *Id.* at 927.

307. *Id.* at 971 (noting that such cases can be rare).

308. *Id.* at 939, 981 tbl.2 (tallying 611 statutory deadlines for the EPA and 940 for the Department of Commerce).

309. Prescription Drug User Fee Act of 1992, Pub. L. No. 102-571, 106 Stat. 4491 (1992) (codified in scattered sections of 21 U.S.C.).

310. Daniel P. Carpenter, *Groups, the Media, Agency Waiting Costs, and FDA Drug Approval*, 46 AM. J. POL. SCI. 490 (2002). Of course, the power dynamic is complicated by the (now quite hefty) user fees required by the Prescription Drug User Fee Act (PDUFA) and its reauthorizations. *See* 21 U.S.C. § 379h(c)(4) (2012); FDA, Prescription Drug User Fee Rates for Fiscal Year 2013, 77 Fed. Reg. 45,639, 45,642 (Aug. 1, 2012) (setting the fee for filing a New Drug Approval ("NDA") application at $1,958,000 for fiscal year 2013).

311. *See* Gersen & O'Connell, *supra* note 204, at 927–28, 951–54 (noting that deadlines sometimes "provide a rare opportunity for parties to successfully sue for agency inaction under section 706(1) of the APA").

312. *Id.* at 945–46, 988 tbl.12. Gersen and O'Connell gathered data between April 1988 to October 2003 from agency reports in the *Unified Agenda of Federal Regulatory and Deregulatory Actions*, published semiannually in the Federal Register. *Id.* at 938 n.57.

313. *Id.* at 943–44.

probably because Congress tends to use deadlines for more significant actions.[314]

On the downside, deadlines can exacerbate the cognitive bias to overestimate newer, higher-profile risks over older, established ones.[315] Yet, in the case of medical device software, the newer risks demonstrated by mobile technologies highlight longstanding risks with software that the FDA historically under-regulates.[316]

Another downside is that deadlines might prompt industry efforts to defeat or weaken the regulation preemptively. Again, this concern is not overwhelming, as agencies already publish Notices of Proposed Rulemaking ("NPRM") and even Advanced Notices of Proposed Rulemaking ("ANPRM") that forecast their interventions.[317] Interim final rules also serve this purpose by binding the public (hence the "final"), but only for a limited time (hence the "interim").[318]

In short, agencies can manipulate the durability of their policies. The FDA's 1989 software policy was not durable enough, and in fact, it did not endure.

D. ENFORCEMENT

Once an agency has settled on when and how to regulate a new technology, it must decide how rigorously to enforce its policy. Executive agencies have unmistakable discretion here.[319] Indeed, agency discretion may reach its apex versus judicial interference in matters of enforcement.[320] As a practical matter, agency discretion is limited only by political and resource constraints.

---

314. *Id.* at 943.

315. *Id.* at 974.

316. Cortez, *supra* note 78 (detailing some of the risks of mobile health technologies and how these echo some of the early concerns with primitive device software).

317. For a discussion of why agencies sometimes use ANPRM, see Barbara H. Brandon & Robert D. Carlitz, *Online Rulemaking and Other Tools for Strengthening Our Civil Infrastructure*, 54 ADMIN. L. REV. 1421, 1465–68 (2002); Alan B. Morrison, *OMB Interference with Agency Rulemaking: The Wrong Way to Write a Regulation*, 99 HARV. L. REV. 1059, 1065 (1986) (criticizing the use of ANPRM as defensive and dilatory). For a discussion of the dynamics of NPRM, see Gersen & Posner, *supra* note 146, at 587.

318. *See* Adoption of Recommendations, 60 Fed. Reg. 43,108, 43,110–12 (Aug. 18, 1995) (discussing interim final rulemaking by several agencies and endorsing its use in certain circumstances); Michael Asimow, *Interim-Final Rules: Making Haste Slowly*, 51 ADMIN. L. REV. 703 (1999); Gersen & Posner, *supra* note 146, at 587.

319. *See* Heckler v. Chaney, 470 U.S. 821, 828–35 (1985); Lisa Schultz Bressman, *Judicial Review of Agency Inaction: An Arbitrariness Approach*, 79 N.Y.U. L. REV. 1657 (2004).

320. *See* Bressman, *supra* note 319, at 1705–10 (analyzing the Supreme Court's decision in Norton v. S. Utah Wilderness Alliance, 542 U.S. 55 (2004)).

But enforcement discretion still demands fidelity to the agency's statutory responsibilities, and, above all, to the public interest. Agencies may be perfectly justified in tempering early enforcement against new technologies. Yet, as the innovation matures and as its risks become more evident, these initial justifications can become threadbare, as demonstrated by the FDA's history with software.

This argument counters several deeply rooted cultural norms about government regulation and technology. The overzealous regulator has long been an archetype in American political discourse. The idea that federal agencies systematically over-regulate is an old one, reaching a crescendo during the Reagan era,[321] with a renewed vigor today. But, as Nicholas Bagley and Richard Revesz argue, this idea is generally wrong.[322] Still, the intuition endures. In fact, efforts to reform federal regulation often do so primarily with regulated industry in mind, rather than regulatory beneficiaries.[323] Some even argue that new technologies will generally regulate themselves.[324]

It is hard not to notice the cognitive dissonance here. There is widespread recognition that compliance with regulations is spotty,[325] and that agencies generally lack the resources and personnel to enforce the law optimally.[326] Scholars have examined regulatory underenforcement in a number of contexts.[327] Recent studies show that even when agencies do make

321. *See* Nicholas Bagley & Richard L. Revesz, *Centralized Oversight of the Regulatory State*, 106 COLUM. L. REV. 1260, 1263–66 (2006); McGarity, *supra* note 155, at 1484–91, 1491–98 (describing the arguments of those advocating for less (or no) regulation).

322. *See generally* Bagley & Revesz, *supra* note 321 (strongly refuting the claim of rampant overregulation and arguing that agencies routinely under-regulate).

323. *See* Mendelson, *supra* note 144, at 429–31 (suggesting that, for a number of reasons, informal rulemaking and reform tend to disenfranchise beneficiaries); McGarity, *supra* note 155, at 1467, 1476–77, 1478–79 (giving examples of industry-oriented reform efforts).

324. *See, e.g.*, Gervais, *supra* note 11, at 689–90 (citing LAWRENCE LESSIG: CODE AND OTHER LAWS OF CYBERSPACE (1999) and considering copyright problems on the Internet). Gervais's analysis, however, is inapposite to many health and safety problems that require regulation.

325. *See* Freeman, *supra* note 230, at 3, 14–17 (noting "that implementation is inconsistent, and that enforcement is at best sporadic are by now uncontroversial claims."); Solomon, *supra* note 230, at 822.

326. *See* RENA STEINZOR & SIDNEY SHAPIRO, THE PEOPLE'S AGENTS AND THE BATTLE TO PROTECT THE AMERICAN PUBLIC: SPECIAL INTERESTS, GOVERNMENT, AND THREATS TO HEALTH, SAFETY, AND THE ENVIRONMENT 54–71 (2010) (discussing the "hollow government" problem, in which agency funding stagnates while statutory responsibilities grow); Rena Steinzor, *The Truth About Regulation in America*, 5 HARV. L. & POL'Y REV. 323, 327–28 (2011).

327. Frequent targets include the EPA, FDA, FEC, OSHA, and SEC. But federal underenforcement of immigration law probably draws the most attention. *See, e.g.*, Adam Cox & Cristina M. Rodríguez, *The President and Immigration Law*, 119 YALE L.J. 458, 534–45

the effort to pursue violations, they frequently fail to enforce judgments and settlements.[328] Underenforcement is often attributed to agencies being underfunded and understaffed.[329] But agencies may be reluctant to ask for more resources, lest they signal their own inefficiency or create new pressures to produce results commensurate with new resources.[330] Thus, we lament the overzealous regulator despite actual evidence that most are not zealous enough.[331]

This narrative fits the FDA well. The agency has been criticized for lax oversight over virtually every product category in its jurisdiction, including food,[332] drugs,[333] and medical devices.[334] In response, former FDA officials

(2009); Stephen Lee, *Private Immigration Screening in the Workplace*, 61 STAN. L. REV. 1103, 1126–29 (2009). For works that address regulatory underenforcement more generally, see Steinzor, *The Truth About Regulation in America*, *supra* note 326; Matthew C. Stephenson, *Public Regulation of Private Enforcement: The Case for Expanding the Role of Administrative Agencies*, 91 VA. L. REV. 93, 116–17 (2005).

328. A remarkable recent study found that regulators systematically undercollect corporate fines and penalties. *See* Ezra Ross & Martin Pritikin, *The Collection Gap: Underenforcement of Corporate and White-Collar Fines and Penalties*, 29 YALE L. & POL'Y REV. 453, 456, 473–74 (2011) (finding that regulatory agencies "never collect the vast brunt of regulatory and criminal penalties," pointing to the Justice Department's collection rate of only 4% of its criminal penalties and fines). In 2006, the Justice Department collected $35 billion in penalties and fines. Martha Mendoza & Christopher Sullivan, *More Corporations Stiffing Government on Fines*, USA TODAY, Mar. 19, 2006, http://usatoday30.usatoday.com/news/washington/2006-03-19-unpaid-fines_x.htm. Federal regulators leave a lot of money on the table, due to a combination of insufficient resources, insufficient incentives, institutional malaise, agency capture, and, perhaps, simply confusion. *See* Ross & Pritikin, *supra*, at 496–507.

329. *See, e.g.*, Freeman, *supra* note 230, at 17 (describing how the EPA's insufficient funding and personnel trace back to Congress).

330. *See* Ross & Pritikin, *supra* note 328, at 502–03. Ross and Pritikin note perceptively that for agencies, it may be "[b]etter to do poorly with little than risk doing poorly with a lot," and that agency leaders may be reluctant "to admit existing deficiencies." *Id.* at 502–03. Conversely, they "may risk funding if they establish that they are doing *too* well" at collecting fines. *Id.* at 503 n.325.

331. This dynamic has been highlighted in other areas, notably for street and violent crimes. *See* Alexandra Natapoff, *Underenforcement*, 75 FORDHAM L. REV. 1715 (2006).

332. *See, e.g.*, INST. OF MED., ENHANCING FOOD SAFETY: THE ROLE OF THE FOOD AND DRUG ADMINISTRATION (2010) (criticizing the agency's piecemeal authority and oversight over food safety); Diana R.H. Winters, *Not Sick Yet: Food-Safety-Impact Litigation and Barriers to Justiciability*, 77 BROOKLYN L. REV. 905, 913–15 (2012).

333. *See, e.g.*, INST. OF MED., THE FUTURE OF DRUG SAFETY: PROMOTING AND PROTECTING THE HEALTH OF THE PUBLIC (2007) (finding major, sometimes structural deficiencies with the FDA's approach for regulating the safety of prescription and over-the-counter drugs).

334. *See, e.g.*, INST. OF MED., MEDICAL DEVICES AND THE PUBLIC'S HEALTH: THE FDA 510(K) CLEARANCE PROCESS AT 35 YEARS (2011) (finding major flaws with the primary process by which most medical devices are introduced to the U.S. market).

argue that the agency suffers from "the hollow government syndrome," whereby it is simultaneously saddled not only with expanding jurisdiction, but also with "stagnant resources and the consequent inability to implement or enforce its statutory mandates."[335] It can be easy to criticize an overwhelmed agency.

At the same time, these excuses dissolve the longer a problem lingers. Underenforcement can negate the value of even optimally timed regulation.[336] It can also erode careful decisions about the proper form and durability of interventions. The decision to proceed by guidance or even by more collaborative forms of soft regulation is easier to justify if the agency provides a traditional hard law backstop via binding rules or enforcement. As new governance proponents like Ian Ayres and John Braithwaite observe, "[r]egulatory agencies will be able to speak more softly when they are perceived as carrying big sticks."[337] Thus, while the FDA's decision to address software by guidance may have been defensible in 1987 or even 1989, it became less so over time, as software products proliferated and enforcement continued to lag.

How do these lessons apply to disruptive innovation? Again, this Article tempers the suggestion that agencies should rely on "threats" as new technologies mature. The primary example Wu uses is the FCC's early Internet policy, which warned Internet service providers to maintain net neutrality.[338] Chairman Powell's speech warned the broadband industry to preserve net neutrality.[339] But the FCC did not make an empty threat. Just one year later, the FCC investigated a telephone company for violating one of these principles,[340] citing violations of the Communications Act of 1934.[341] Shortly after the settlement, the FCC published a policy statement explaining

335. Peter Barton Hutt, *Recent Developments: The State of Science at the Food and Drug Administration*, 60 ADMIN. L. REV. 431, 432 (2008).

336. *See* Super, *supra* note 68, at 1404.

337. IAN AYRES & JOHN BRAITHWAITE, RESPONSIVE REGULATION: TRANSCENDING THE DEREGULATION DEBATE 6 (1992); *see also* Freeman, *supra* note 230, at 14–17 (critiquing the EPA's process of enforcing standards through permits).

338. *See* Wu, *supra* note 11. "Net neutrality," of course, is a phrase Wu coined. Tim Wu, *Network Neutrality, Broadband Discrimination*, 2 J. TELECOM. & HIGH TECH. L. 141 (2003).

339. Wu, *supra* note 11, at 1844–45 (citing Michael K. Powell, Chairman, FCC, Preserving Internet Freedom: Guiding Principles for Industry, Remarks at the Silicon Flatirons Symposium on "The Digital Broadbrand Migration: Toward a Regulatory Regime for the Internet Age" (Feb. 8, 2004), *available at* http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-243556A1.pdf).

340. *See* Wu, *supra* note 11, at 1845; Madison River Commc'ns, LLC, 20 FCC Rcd. 4295 (2005), *available at* http://digital.library.unt.edu/ark:/67531/metadc4102/m1/371.

341. *Madison River*, 20 FCC Rcd. 4295, 4296 (citing the Communications Act of 1934, 47 U.S.C. § 201(b)); Weiser, *supra* note 39, at 563.

four Internet freedoms,[342] which preceded a legally binding rule just six years later.[343] The FCC buttressed the rule by enforcing it.[344]

Adjudication and enforcement require sustained effort.[345] The FDA has brought intermittent enforcement actions against software products, particularly via warning letters,[346] but it has not pursued enough high-profile cases to offset its primary reliance on nonbinding (and since withdrawn) guidance.

Underenforcement can undermine an agency's credibility, and failure to enforce can compound other problems for agencies.[347] A caution voiced during a recent FDA public workshop on software products was that if too many people do not follow the FDA's rules, they would undermine the entire system.[348] As noted in Part III, this is close to what happened with medical device software. And history risks repeating itself with the newest generation of software products for mobile platforms.[349]

Of course, it is hard to discuss FDA enforcement without appreciating its gatekeeping authority over new technologies. Agencies, like the FDA, that enjoy preapproval power can leverage it to better enforce their policies whether those policies are technically binding or not.[350] In the case of software products, the FDA's statutory authority to pre-approve and pre-clear devices gives it gatekeeping authority.[351] FDA guidances thus demand respect. Yet this leverage dissolves when the FDA exempts certain products from premarket review, as it has with many lower risk categories of devices,

342. Appropriate Framework for Broadband Access to the Internet over Wireline Facilities, Policy Statement, 20 FCC Rcd. 14,986 (Aug. 5, 2005), *available at* http://digital.library.unt.edu/ark:/67531/metadc4091/m1/708.

343. *See* Wu, *supra* note 11, at 1852; Preserving the Open Internet, 25 FCC Rcd. 17,905 (Dec. 21, 2010), *available at* http://digital.library.unt.edu/ark:/67531/metadc39308/m1/451.

344. *See* Weiser, *supra* note 39, at 584–89.

345. *See id.* at 585–86.

346. *See, e.g.*, *supra* note 140.

347. *See* Weiser, *supra* note 39, at 586 (discussing the FCC's underenforcement).

348. *FDA Public Workshop*, *supra* note 80 (Statement of Grant Elliott, Voxiva).

349. Cortez, *supra* note 78 (drawing parallels between the FDA's initial approach to device software and its approach to the latest generation, embodied in mobile medical applications for smartphones, tablets, and other wireless devices).

350. *See* Raso, *supra* note 203, at 803 (noting that the FDA and FCC both enjoy gatekeeping authority that encourages regulated industries to comply with the agencies' guidances). *See generally* CARPENTER, *supra* note 19. For criticisms of FDA's gatekeeping authority over drugs, see C. Frederick Beckner, III, *The FDA's War on Drugs*, 82 GEO. L.J. 529 (1993); Michael P. VanHuysen, Note, *Reform of the New Drug Approval Process*, 49 ADMIN. L. REV. 477 (1997).

351. Federal Food, Drug, and Cosmetic Act § 510(k), 21 U.S.C. §§ 360e (2012).

like the medical device data systems mentioned above.[352] Gatekeeping leverage is less useful when the product has already passed through the gate, or when the gates are left wide open.[353]

So, what is the best way to address suboptimal enforcement by resource-strapped agencies like the FDA? Some commentators suggest relying on citizen enforcement and private rights of action.[354] Nina Mendelson argues that expanding such mechanisms could not only increase overall enforcement, but also redress some of the procedural deficiencies of agencies relying on guidance documents.[355] In effect, we could deputize citizens and interest groups to use private resources to enforce regulatory obligations that might otherwise go un- or under-enforced.[356]

Currently, citizen suits are somewhat rare in administrative enforcement schemes, outside of environmental statutes.[357] Private rights of action are also rare.[358] However, Mendelson argues that "[r]egulatory beneficiaries' concerns over inadequate agency enforcement of statutory provisions could be addressed by expanding citizen enforcement provisions to encompass obligations under a broader array of health, safety, and environmental laws."[359] These vehicles would remind agencies that they have constituents other than industry. They could also combat agency "slack."[360] Of course, it might be difficult for regulatory beneficiaries to "detect and document" regulatory violations.[361] And there are other problems,[362] including the reality

352. Medical Device Data Systems, 76 Fed. Reg. 8637 (Feb. 15, 2011) (codified at 21 C.F.R. § 880.6310 (2011)).

353. *See* Seiguer & Smith, *supra* note 203, at 28 (noting that when FDA wants enforceability, it chooses rulemaking over guidances, particularly when regulation slants towards postmarket rather than premarket requirements).

354. *See, e.g.*, Stephenson, *supra* note 327.

355. *See* Mendelson, *supra* note 144, at 450–52.

356. *See* Stephenson, *supra* note 327, at 107–08.

357. *See* Barton H. Thompson, Jr., *The Continuing Innovation of Citizen Enforcement*, 2000 U. ILL. L. REV. 185, 191–92 (2000) (discussing private enforcement of environmental laws); Mendelson, *supra* note 144, at 450–51 (citing examples of citizen suits).

358. *See* Mendelson, *supra* note 144, at 451 (noting that the major exceptions occur in securities and civil rights statutes).

359. *Id.*

360. Agency "slack" is "the tendency of government regulators to underenforce certain statutory requirements because of political pressure, lobbying by regulated entities, or the laziness or self-interest of the regulators themselves." Stephenson, *supra* note 327, at 110 (internal citations omitted).

361. *See* Mendelson, *supra* note 144, at 451.

362. *See id.* at 452 (noting that citizen suits are expensive, may face constitutional difficulties, and can interfere with agencies' ability to set their enforcement priorities and allocate scarce regulatory resources); Stephenson, *supra* note 327, at 114–20 (listing problems with private enforcement).

that private enforcement would probably need to be authorized by Congress.[363] But for consumer products and markets, consumers might serve as early sentinels of problems.[364]

In short, agencies need not be so deliberate and tentative with regulating innovations—even disruptive ones. If agencies are concerned with regulating prematurely or incorrectly, then they can experiment with timing rules, push Congress to authorize private enforcement, or try other adaptations of traditional interventions that reduce the costs of errors. If agencies do choose to proceed by making threats, then they should use them as a short-term precursor to more decisive action, as the FCC did, and guard against relying on them as a long-term crutch, as the FDA did.

V. CONCLUSION

Periodically, regulators are confronted by novel products, technologies, or business practices that fall within their jurisdiction but do not fit comfortably within their regulatory frameworks. Agencies face "regulatory disruption." Many scholars and policymakers intuit that the appropriate response is for regulators to be cautious, not decisive. As Tim Wu recently argued, agencies should make "threats" to encourage desired behavior or discourage undesired behavior.

Though intuitive, the idea of using "threats" as a tentative, flexible initial posture can easily calcify, creating a weak default position that leads to suboptimal regulation over longer periods. Regulatory inertia can be hard to break without an external shock, usually a tragedy or massive failure that reignites interest in regulation.

This discussion implicates not just the FDA's or the FCC's consideration of new technologies, but other agencies' as well. If agencies are concerned about regulating novel products or practices prematurely or erroneously, they can experiment with timing and enforcement methods to reduce the risks of both. Agencies can use experimental rules, regulatory sunsets, or rulemaking deadlines to calibrate their approach to novel technologies or business practices. Also, citizen suits or private rights of action might help correct systematic underenforcement by resource-strapped regulators, who might find themselves even more strapped for resources when new products, firms, or even industries materialize within their jurisdictions.

363. *See generally* Stephenson, *supra* note 327 (arguing that Congress should grant such delegations, and if not, courts should infer it).

364. *See id.* at 108 (noting that private parties directly affected by regulatory violators' conduct may be better suited to detection violations).

The plight of the modern regulator is to adapt old regulatory frameworks to new technologies and practices. This should not be as paralyzing as it seems.

BEYOND EFFICIENCY: CONSEQUENCE-SENSITIVE THEORIES OF COPYRIGHT

Oren Bracha[†] *& Talha Syed*[††]

ABSTRACT

A major development in the analysis of copyright in the last two decades has been the emergence of "democratic" theories of cultural expression, challenging the long-standing dominance in this field by normative theories of natural rights and economic efficiency. The central thrust of democratic theories is to emphasize the significance of the expressive activities regulated by copyright law not only for political life, but also for individual self-authorship and robust engagement by persons in their surrounding culture, to take an active part in social processes of meaning-making. However, a major stumbling block facing democratic theories has been a lack of understanding, by both advocates and skeptics, of how democratic theories relate to rival views in terms of concrete implications for specific questions of copyright law and policy. In particular, democratic (and, as we show, distributive-equity) theories seem to share with economic analysis an attractive attentiveness to the consequential effects of copyright, in contrast to so-called "deontological" natural-rights views. Yet, by that same token, it has been unclear precisely when—and on what grounds—democratic theories offer distinct prescriptions from the one commended by economic analysis, which is to achieve an optimal balance between the maximal production and wide dissemination of expressive works valued by consumers.

This Article contributes to the further development and reception of this emerging normative camp by clarifying its relationship to dominant rival views. Specifically, we seek to advance the sophistication and precision of democratic and distributive analyses of copyright in three respects. First, we offer a characterization of democratic and distributive theories as being, in their most attractive form, robustly "consequence-sensitive," and we bring out a crucial feature of these theories when conceived this way. So conceived, democratic and distributive theories share an important structural affinity with economic efficiency, which is that they must confront the possibility that an adjustment to copyright rules along one dimension may have countervailing repercussions along other dimensions.

As a result of this structural affinity, democratic and distributive theories—surprisingly, given their invocation of different values—seem at first blush to align closely with economic analysis and its familiar incentive-access tradeoff. This sharply poses the challenge of identifying the stakes in the choice between these alternative outlooks. Our second aim is to respond to this challenge by specifying with greater precision than has yet been done how

---

© 2014 Oren Bracha & Talha Syed.

† Howrey LLP and Arnold, White & Durkee Centennial Professor of Law, The University of Texas School of Law.

†† Assistant Professor, University of California, Berkeley School of Law. We thank David Adelman, Bob Bone, Yochai Benkler, Kevin Collins, Karen Engle, Niva Elkin-Koren, Terry Fisher, Peter Gerhart, John Golden, Wendy Gordon, Amy Kapczynski, Chris Kutz, Larissa Mann, Roni Mann, Peter Menell, Rob Merges, Jennifer Rothman, Pam Samuelson, Molly Van Houweling and participants at the Penn Law Center for Technology, Innovation and Competition workshop, the Loyola IP Colloquium, and the IP Scholars Conference for invaluable discussions and comments on an earlier version of this Article.

democratic and distributive values make a concrete difference in the context of particular policy questions. Third and finally, with the normative differences with efficiency so clarified, we are then faced with two tasks: justifying such departures from efficiency's beacon of maximal satisfaction of existing preferences and specifying how a consequence-sensitive approach that adopts plural values is able to decide between conflicting values and outcomes, in the absence of reliance on the efficiency calculus. We explore these questions in the context of six concrete issues in copyright policy.

TABLE OF CONTENTS

I. **INTRODUCTION** 231

II. **ECONOMIC ANALYSIS AND ALTERNATIVE CONSEQUENCE-SENSITIVE THEORIES** 237

A. ECONOMIC ANALYSIS OF COPYRIGHT: THE STATE OF THE ART 237
1. *The Incentive-Access Framework* 237
2. *Supramarginal and Inframarginal Markets* 240
3. *Product Differentiation* 241
4. *Global Distortions* 243

B. CONSEQUENCE-SENSITIVE THEORIES OF COPYRIGHT 244

III. **SELF-DETERMINATION, DEMOCRACY AND FLOURISHING THEORIES** 248

A. MAPPING THE THEORIES 249
1. *Self-Determination* 251
2. *Political Democracy* 252
3. *Cultural Democracy* 253
4. *Human Flourishing* 256

B. THE THEORIES IN ACTION 258
1. *Critical Secondary Uses* 258
2. *Derivative Works* 266
3. *Fan Fiction* 274
4. *Personal Uses* 281

IV. **DISTRIBUTIVE EQUITY IN COPYRIGHT POLICY** 287

A. VARIETIES OF DISTRIBUTIVE CONCERNS 289
B. KEY PARAMETERS OF POSITIVE ANALYSIS 296
C. APPLICATIONS 299
1. *Special-Needs Individuals* 301
2. *Lower-Income Individuals* 305

D. OBJECTIONS 311
1. *Futile/Perverse* 311
2. *Unwieldy/Distortionary* 312

V. **CONCLUSION** 313

I. INTRODUCTION

The recent decision of a New York federal district court in *Authors Guild v. HathiTrust*[1] is a powerful reminder that copyright law is about more than just efficiency. HathiTrust, a collaborative of major research institutions and libraries dedicated to preserving and providing access to printed knowledge, used digital copies of copyrighted books provided to its members by Google as part of Google's mass book digitization project.[2] In its opinion, the court firmly rejected the claim that HathiTrust, by using digital copies, infringed copyright.[3] Specifically, the court found that the use of the unauthorized digital copies for purposes of non-display full text searches, and access for people with print-disabilities, constituted fair use (with the latter use also being shielded by a specific statutory exemption).[4] To be sure, this outcome could be plausibly justified on efficiency grounds. The social value of the uses that, were they to be found infringing, may be frustrated by various "market failures" in licensing might well outweigh any cost in terms of undermining creative incentives.

Reading the opinion, however, clearly reveals a different tenor. The court's references to the "invaluable contribution to the progress of science and cultivation of the arts"[5] of the alleged infringing uses, and to the "ideals espoused by" the Americans with Disabilities Act of securing for those disadvantaged by disability "the opportunity to compete on an equal basis and to pursue those opportunities for which our free society is justifiably famous,"[6] leave little room for doubt. Although not explicitly articulated, the decision is based on normative assumptions quite distinct from economic efficiency. Promoting "the progress" and fair opportunities seem clearly not reducible in the court's analysis to the maximization of the total dollar value of producer and consumer preferences.

On one level, there is nothing new about this normative pluralism beyond efficiency. The Constitution's original "promote the progress" mandate for the intellectual property power[7] was associated with Enlightenment ideas of human advancement that were by no means

1. Authors Guild, Inc. v. HathiTrust, 902 F. Supp. 2d 445 (S.D.N.Y. 2012).
2. *Id.* at 448.
3. *Id.* at 464–65.
4. *Id.*
5. *Id.* at 464.
6. *Id.* at 465 (quoting 42 U.S.C. § 12101(a)(8)).
7. U.S. CONST. art. I, § 8, cl. 8.

reducible to economic development.[8] And in various subsequent periods, many Americans have read broad normative visions into the mandate and the copyright regimes instituted under its aegis.[9]

Yet, the modern era of American copyright, beginning in the late nineteenth century, has been dominated by utilitarian-economic ideas. To be sure, natural rights theories, based on the principle of just deserts for authors' intellectual labor, have also been present from the start. And continental notions of property rights as existing to protect an author's personality or will as embodied in his creation, did at some point percolate into American thinking on copyright, even if they have never gained much traction.[10] Nevertheless, the framework of economics has taken a front seat in American debates on copyright law and policy. Thus, in his seminal 1970 article *The Uneasy Case for Copyright*, Justice Breyer observed that "none of the noneconomic goals served by copyright law seems an adequate justification for a copyright system."[11] The trend intensified with the rise to preeminence of the law-and-economics school in American legal culture, especially in fields of private law. Economics, either in the form of formalized analyses of efficiency or in a looser guise under the notion of "innovation," became the predominant normative framework for American copyright.

But the preeminence of economic analysis has gradually begun to diminish over the past two decades. New contenders have appeared in the normative arena of copyright. An assortment of theories and ideas about free speech, democratic self-governance, human flourishing, and, to a lesser extent, distributive justice has emerged.[12] For purposes of convenience we loosely designate these theories "democratic and distributive theories." In academic writing these approaches have gathered force, likely surpassing natural rights theories in influence and perhaps even challenging the previously unrivaled preeminence of economics. A shift is also apparent in public discourse about copyright policy. The phrases "freedom of speech" and "access to knowledge" have become familiar sound bites in public debates. One rarely comes across a mission statement of any public-interest organization in the field without encountering some version of the

8. *See* Michael Birnhack, *The Idea of Progress in Copyright Law*, 1 BUFF. INTELL. PROP. L. J. 3, 17–21 (2001).

9. *See generally* MEREDITH MCGILL, AMERICAN LITERATURE AND THE CULTURE OF REPRINTING, 1834–53 (2003).

10. *See* William Fisher, *Theories of Intellectual Property*, *in* NEW ESSAYS IN THE LEGAL AND POLITICAL THEORY OF PROPERTY 168, 174 (Stephen R. Munzer ed., 2001).

11. Stephen Breyer, *The Uneasy Case for Copyright: A Study of Copyright in Books, Photocopies, and Computer Programs*, 84 HARV. L. REV. 281, 291 (1970).

12. *See* discussion *infra* Parts III, IV.

commitment to defend the "public's access to knowledge."[13] And although the change is less apparent in litigation and judicial reasoning, we can discern some initial signs of it there as well, as evinced by *Authors Guild v. HathiTrust*.[14]

And so a question arises: what is the relationship between economic analysis and these new normative approaches? Unfortunately, exploration of that topic has thus far been plagued by considerable confusion and miscommunication. Many scholars working within the framework of economics simply adopt efficiency as the only plausible normative criterion for the field, remaining content to ignore any alternative. This attitude is buttressed by two widespread and related beliefs. The first is a belief in the inherent implausibility of any alternative criteria to the economists' one beacon of caring solely about the welfare effects of policy choices.[15] Second, and more specifically, the democratic and distributive normative theories of copyright are often believed to exhibit an unattractive "deontological" disinterest in the consequential effects of policy or absolutist disregard for other considerations.[16]

Occasionally, however, proponents of economics advance a very different critique of these alternative theories. They point out that many of the new democratic and distributive theories are actually quite interested in the various positive and negative ways that adjustments to copyright may affect different individuals' interests.[17] But in that case, then, these theories are seen basically to collapse into efficiency: their ultimate normative calculus dovetails with economic cost-benefit,[18] only doing so in a less explicit fashion and with less sophisticated tools.

From the other side of the debate, proponents of democratic and distributive approaches frequently tend to assume too quickly that simply adopting their alternative framework will lead to normative conclusions significantly different from those dictated by efficiency.[19] Too often, the mere

13. *See, e.g.*, *About Public Knowledge: Our Work*, PUBLIC KNOWLEDGE, http://www.publicknowledge.org/about (last visited Sept. 18, 2013).

14. Authors Guild, Inc. v. HathiTrust, 902 F. Supp. 2d 445, 464–65 (S.D.N.Y. 2012).

15. *See generally* LOUIS KAPLOW & STEVEN SHAVELL, FAIRNESS VERSUS WELFARE (2002).

16. *See infra* text accompanying notes 45–47.

17. *See* Michael Abramowicz, *An Industrial Organization Approach to Copyright Law*, 46 WM. & MARY L. REV. 33, 89 (2004).

18. *Id.* (positing that "the democratic and economic interests underlying copyright are, for the most part, likely to be aligned on issues of copyright policy").

19. This seems, for example, to be the premise of Madhavi Sunder's observation that increasingly "intellectual property laws come to bear on giant-sized values, from democracy and development to freedom and equality" in the opening chapter of her book entitled

invocation of "freedom of speech" or "distributive justice" in regards to a question of copyright law or policy, is thought to point toward a different result than the one commended by efficiency. Two reasons seem to lie behind this tendency. The first is that in some cases, an unduly thin or simplistic version of the efficiency analysis is assumed, which fails to see that efficiency can and should take into account many of the effects highlighted by competing normative frameworks. For example, copyright barriers to access for print-disabled readers have, alongside their distributive effects, negative efficiency implications as well. And it is hardly self-evident that efficiency's prescriptions for how to remedy these effects will differ significantly from a more distribution-sensitive analysis. A second and more important reason is that proponents of non-efficiency normative frameworks tend to miss the relevance of the tradeoffs highlighted by efficiency analysis for their own concerns. For example, it is often ignored that any reduction in creative incentives from reduced copyright protection is not just an efficiency cost, but also a potential negative effect from the point of view of democratic theory or even distributive justice. As a result, proponents of alternative theories seldom grapple with prioritizing between various countervailing effects similar to those traded off in an efficiency analysis.

The purpose of this Article is to begin transcending this state of affairs, in which economic analysis and democratic and distributive approaches talk past one another. The new theories that are the focus of our inquiry here are distinguished by the quality of being "consequence sensitive." As such, they differ both from traditional natural rights and personality theories of copyright on the one hand, and on the other, from utilitarian or welfarist theories (and their most common legal-policy variant of economic efficiency). Consequence-sensitive theories are distinct from traditional labor-desert and personality justifications of copyright, as well as any negative-liberty-based assertions of users' rights, because of their *comprehensive* interest in the full array of consequences attributable to a particular rule, including effects on individuals not privy to the regulated behavior, and the forward-looking effects of general incentives created by the rule. At the same time, unlike efficiency, which is a full-blown consequentialist theory, consequence-sensitive theories need not make effects in the world—conceived

Beyond Incentives. *See* MADHAVI SUNDER, FROM GOODS TO A GOOD LIFE: INTELLECTUAL PROPERTY AND GLOBAL JUSTICE 32 (2012). Similarly, Neil Netanel argues that "a copyright law driven by neoclassical economic property theory would give copyright owners such far-reaching control over productive uses of existing creative works that it would unduly constrain robust debate upon which democratic self-rule depends." Neil Weinstock Netanel, *Copyright and a Democratic Civil Society*, 106 YALE L.J. 283, 288 (1996).

independently of the character of decisions or processes leading to them—as the sole element relevant to normative analysis.

The central analytic point we develop here follows from the consequence-sensitive character of the new theories we discuss. These consequence-sensitive normative views focus on values other than wealth maximization. As a result, they diverge from efficiency in how they evaluate the effects of copyright protection, for example, in terms of impacts on the creation of new works, access to existing works, or any other outcomes (such as rent dissipation or diversion of resources from other sectors). Nevertheless consequence-sensitive theories must attend, like efficiency analysis, to the *multiple* (and often countervailing) impacts that different legal rules will have. Just as efficiency analysis seeks to balance incentives with access or, more comprehensively, various, countervailing economic effects, so too must non-efficiency views also factor in the full range of effects. Once stated, the point may seem obvious. Too often, however, advocates of non-efficiency values have seemed satisfied simply to point out how their chosen value may weigh one particular outcome—typically increased access—differently from efficiency, without considering the other effects that also must be assessed in light of the deviation between efficiency and their value(s).[20]

This fundamental structural similarity between efficiency and its consequence-sensitive alternatives poses a three-part challenge to proponents of democratic and distributive theories of copyright. First, one must explain clearly the value or values that the theory seeks to promote and how they differ from the maximum satisfaction of individuals' revealed preferences (i.e., efficiency). Second, each theory must identify when and how it leads to results significantly different from those dictated by efficiency regarding specific copyright policy questions. The difference in values alone does not necessarily produce such a divergence. Given the consequence-sensitive nature of the theory, its prized value—say individual self-determination or distributive priority to a specific group—is likely to be subject to an incentive/access tradeoff that closely tracks the efficiency analysis.[21] Third, in

20. *See* Daniel A Farber, *Conflicting Visions and Contested Baselines: Intellectual Property and Free Speech in the Digital Millennium* 89 MINN. L. REV. 1318, 1321 (2005) (describing advocates of non-efficiency values in copyright as "Jeffersonians" who seek to defend the public domain and fair use rights "against the incursion of stronger IP rights" and reject the "Hamiltonian" argument for stronger incentive).

21. If, for example, both efficiency and a particular brand of democratic theory care about the incentive-to-create and the impediment-to-access sides of the copyright tradeoff, would they reach different results on such issues as the proper treatment by copyright law of fan fiction or the application of fair use to critical uses of copyrighted materials? Simply

a case where a theory does support a result materially different from the efficiency balance, we need a convincing justification for not yielding to subjective individual preferences.[22] To put it bluntly, a consequence-sensitive theory that rejects the efficiency calculus overrides the satisfaction of subjective preferences in favor of some other value.[23] Such an outcome may be acceptable given the theory's values, but an explicit recognition of its prima facie troubling character, and some justification for it, are in order.

In this Article we respond to these three challenges and demonstrate how the positive-analytical tools of economics can be used heretically by those who do not accept efficiency as a normative criterion. Our goal is to make a case for these new normative perspectives that is both analytically precise and defensible in the face of the serious criticisms and challenges just adduced. By doing so, we seek to clarify the stakes in the choice between these theories and efficiency (as well as among the theories themselves). Part II lays the necessary foundation for the discussion by briefly explicating the positive-analytical framework of economic analysis of copyright and explaining the structural similarity between efficiency and many democratic and distributive theories. Part III focuses on our so-called "democratic theories" of copyright. We map the main variants of the different normative perspectives grouped together, somewhat imperfectly, under this title. We then proceed to demonstrate how, despite their structural similarity to efficiency, some of these theories yield significantly different conclusions with respect to some central issues of contemporary copyright law and policy, and we explain how these alternative outcomes may be justified. We also highlight how differences in normative commitment between the diverse

pointing out the high value that democratic theory places on diverse and critical expression is insufficient to tilt the balance in favor of access when such a rule could severely reduce the incentive to create, thereby ultimately reducing the supply of expressive works that provide opportunities for diverse and critical expression.

22. Democratic and distributive theories relate differently to this aspect. Democratic theories sometimes override subjective preferences for reasons other than the gap between preferences per se and efficiency's wealth-maximization metric of "effective preferences" (i.e., preferences as revealed by willingness and ability to pay). In other words, these theories may sometimes override not just effective preferences but subjective preferences per se. Distributive theories, by contrast, diverge from wealth maximization due, at least in part, precisely to the gap between effective preferences and preferences per se.

23. Consider a case where efficiency militates against exempting a particular use of a copyrighted work under the fair use doctrine because the social value of the works that would not be created due to the exemption clearly outweighs the social value of the increased access generated by exempting the use. A finding of fair use, in the name of say the value of self-determination, is tantamount to forcing upon the public (and possibly the individual user of the copyright work) a less preferable state of affairs by their own subjective lights.

variants of democratic theories can matter for resolving particular policy questions. Among the issues we take up are critical uses, transformative uses, derivative-work rights, fan fiction, and personal uses. Part IV is devoted to a similar analysis of distributive considerations in copyright. Following a brief sketch of distributive justice theory, we identify the main types of distributive concerns that are implicated by copyright law. We then canvass some central cases meriting distributive consideration (namely, priority to individuals with disabilities and at lower-income levels), specify how a distributively-informed analysis may diverge from efficiency, and evaluate the merits of different ways of factoring in such distributive concerns (ranging from precisely-targeted provisions to generally-applicable rules). Part V concludes.

II. ECONOMIC ANALYSIS AND ALTERNATIVE CONSEQUENCE-SENSITIVE THEORIES

In this Part, we lay the foundation for discussing democratic and distributive theories of copyright by summarizing the fundamentals of copyright's economic analysis and explaining the basic structural features that those theories share with efficiency. The individual elements in our summary of economic analysis of copyright should be familiar to those with a background in the field. Our account emphasizes, however, significant features that are often neglected in standard accounts, and we provide an integrated picture by bringing together several distinct levels of analysis.

A. ECONOMIC ANALYSIS OF COPYRIGHT: THE STATE OF THE ART

1. The Incentive-Access Framework

From an economic perspective, the policy effects of copyright, like those of some other intellectual property rights, are analyzed against the backdrop of the public-goods character of informational goods.[24] Works of authorship—the sort of informational works protected by copyright law—

24. On public goods, see Thomas E. Borcherding, *Competition, Exclusion, and the Optimal Supply of Public Goods*, 21 J.L. & ECON. 111, 111–12 (1978); J.G. Head, *Public Goods and Public Policy*, 17 PUB. FIN. 197 (1962); Paul A. Samuelson, *The Pure Theory of Public Expenditure*, 36 REV. ECON. & STAT. 387 (1954). For examples of the analysis of informational works as public goods, see, for example, Kenneth Arrow, *Economic Welfare and the Allocation of Resources for Invention*, *in* THE RATE AND DIRECTION OF INVENTIVE ACTIVITY: ECONOMIC AND SOCIAL FACTORS 609 (Nat'l Bureau of Econ. Research ed., 1962); William W. Fisher III, *Reconstructing the Fair Use Doctrine*, 101 HARV. L. REV. 1659, 1700–04 (1988); William R. Johnson, *The Economics of Copying*, 93 J. POL. ECON. 158, 161 (1985); William M. Landes & Richard A. Posner, *An Economic Analysis of Copyright Law*, 18 J. LEGAL STUD. 325, 326–27 (1989); Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEX. L. REV. 989, 994–95 (1997).

exhibit the two defining features of public goods. One, they are nonexcludable, in the sense that once created and published it is very difficult to prevent others from using and consuming them. Two, they are also nonrivalrous, meaning that the use and consumption of the work by one person does not reduce the ability of others to use and consume it. The source of the innovation policy problem is the nonexcludable character of the work, which allows others to reproduce and/or consume it for a fraction of the cost incurred by the creator. To the extent that such access by others does not allow the creator to appropriate enough of the social value of the work to cover its development cost, the work (or similar ones in the future) will not be created in the first place. The result: an insufficient or suboptimal level of compensation or incentives to produce and publish works of authorship.[25] Copyright solves this problem by conferring upon the creator a limited set of legal entitlements to exclude others from certain uses of the work.[26] Copyright converts, in other words, a nonexcludable resource into a partially excludable one in order to allow the copyright owner to internalize a substantial part of the social value of the work, thereby boosting the financial resources or incentives for creation and publication.

This solution results, however, in a new problem or cost stemming from the other public-goods aspect of informational works, their nonrivalrousness. As a result of copyright protection, some potential users will be excluded from using the work, which—because the work is nonrivalrous—is inefficient from a static point of view. Market transactions cannot fully

25. Breyer, *supra* note 11, at 294. There is an important strand of arguments in copyright scholarship that challenges the need for, or the efficacy of, copyright-generated incentives for creation. One set of arguments points out that under some circumstances alternative mechanisms such as first-mover advantages or contractual arrangements may guarantee sufficient incentives to create even in the absence of copyright. *See id.* at 299–306. A more fundamental critique questions the extent to which makers of creative works are motivated at all by extrinsic, monetary considerations and consequently the need for copyright or its ability to stimulate creation. *See* Julie E. Cohen, *Creativity and Culture in Copyright Theory*, 40 UC DAVIS L. REV. 1151, 1192–93 (2007) (advocating a "much more modest conception of the role that copyright plays in stimulating creative processes and practices"). In our analysis here we bracket such considerations. To be sure, if it could be demonstrated that extrinsic incentives have, across the board, little or no effect on creativity, the analysis of all consequentialist and consequence-sensitive approaches would have to be amended to give little weight to considerations of creating incentives through exclusionary entitlements. To date, this strong proposition has not been established. A more nuanced version of the argument may try to assess and quantify the extent to which non-monetary intrinsic and social motivations play a role in specific contexts. Such assessment, if and when it becomes available, should be incorporated into the incentive/access analysis, presumably by adjusting the need for incentive side of the tradeoff downward in specific contexts.

26. *See* 17 U.S.C. § 106 (2012) (conferring upon the copyright owner the exclusive rights, inter alia, to reproduce and distribute copies of the copyrighted work).

remedy this allocative inefficiency. In the absence of the unrealistic possibility of costless and perfect price discrimination,[27] the copyright owner will use the pricing power conferred by his exclusive entitlement to charge a uniform price for the work that will exceed the marginal cost of its dissemination. Some consumers who are willing to pay more than this marginal cost but less than the copyright holder's price will be denied access to the work, an allocative inefficiency commonly referred to as deadweight loss.[28]

The incentive function of copyright and its access cost are correlatives. Providing incentives to create through copyright requires some pricing power conferred on the owner. Pricing power, in the absence of costless and perfect price discrimination, necessarily means some deadweight loss. The goal of copyright policy within this incentive-access framework is optimization—incurring only the minimal access cost necessary for incentivizing the creation of works. There are two ways in which copyright law can be optimized to achieve this goal. First, the exclusive entitlements should be so crafted that the pricing power conferred on the copyright owner is not stronger than necessary to incentivize the creation of the work.[29] Any pricing power beyond this level means additional access cost without the corresponding incentive benefit. Second and more subtly, in choosing between alternative ways to generate the necessary incentive, the method that generates the least access cost should be preferred. This reflects the point that sometimes, using various levers of copyright law, the same level of incentive may be secured by several distinct methods, each generating

27. Perfect price discrimination is the ability to charge each consumer exactly the price s/he is willing and able to pay for the relevant good. The economic and other social effects of various partial price discrimination schemes, their desirability as a matter of policy, and the extent to which legal doctrine should encourage them is the subject of continuing debate in legal scholarship. *See generally* Yochai Benkler, *An Unhurried View of Private Ordering in Information Transactions*, 53 VAND. L. REV. 2063 (2000); James Boyle, *Cruel, Mean, or Lavish? Economic Analysis, Price Discrimination and Digital Intellectual Property*, 53 VAND. L. REV. 2007 (2000); Julie E. Cohen, *Copyright and the Perfect Curve*, 53 VAND. L. REV. 1799 (2000); William W. Fisher III, *When Should We Permit Differential Pricing of Information?*, 55 UCLA L. REV. 1 (2007); Wendy J. Gordon, *Intellectual Property as Price Discrimination: Implications for Contract*, 73 CHI.-KENT L. REV. 1367 (1998); Glynn S. Lunney Jr., *Copyright's Price Discrimination Panacea*, 21 HARV. J.L. & TECH. 387 (2008); Michael J. Meurer, *Copyright Law and Price Discrimination*, 23 CARDOZO L. REV. 55 (2001).

28. *See* Fisher, *supra* note 24, at 1702 (stating "consumers who value the work at more than its marginal cost but less than its monopoly price will not buy it," resulting in deadweight loss); Glynn S. Lunney, Jr., *Reexamining Copyright's Incentives-Access Paradigm*, 49 VAND. L. REV. 483, 497–98 (1996) (arguing that copyright imposes "deadweight loss, measured by the combined loss in consumer and producer surplus associated with the sales lost as a result of the higher more monopolistic price").

29. This means that the producer should expect, ex ante, to be able to recoup the capitalized costs of developing the work.

different amounts of access cost.[30] From the point of view of efficiency, each of the numerous details of copyright law should be assessed and adjusted within this framework so as to provide only the necessary amount of incentive and to do so at the minimal cost to access.

2. Supramarginal and Inframarginal Markets

Were the terms of copyright tailored for each work, determining their contours would be theoretically straightforward. Each work would receive only as much protection as necessary to incentivize its creation, and among alternative possible configurations of exclusive entitlements that generate this level of incentive, the one with the least access cost would be preferred. Weaker protection would fail to incentivize the creation of the work. Stronger protection would add access cost with no additional benefit.

In a copyright regime with at least some degree of uniformity—meaning a regime of rules that apply uniformly to entire categories of works, irrespective of different specific circumstances—things work differently.[31] In a uniform copyright regime there is an incentive/access tradeoff between different works. At any level of protection some works will be marginal in the sense that copyright suffices exactly to incentivize the creation of the work. Other works will be inframarginal: copyright protection will be stronger than necessary to incentivize the creation of these works. Others still will be supramarginal: copyright protection will provide insufficient incentive, and the work will not be created. The incentive/access tradeoff at the heart of copyright's economic policy is as between inframarginal works and supramarginal works. At any baseline level of copyright, any additional increment of copyright protection will add social value by incentivizing the creation of previously supramarginal works. At the same time, any such increment will further unnecessarily limit access in regard to inframarginal works, thereby adding deadweight loss.

Thus, in a modern copyright regime whose rules have some degree of uniformity, the basic policy analysis of any legal arrangement focuses on its

30. *See* Fisher, *supra* note 24, at 1703–17 (developing an "incentive/loss ratio" analysis for evaluating alternative packages of entitlements to confer on copyright owners); *cf.* Louis Kaplow, *The Patent-Antitrust Intersection: A Reappraisal*, 97 HARV. L. REV. 1813, 1829–34 (1984) (developing an "incentive/loss ratio" analysis in the patent context).

31. For a discussion of the policy tradeoffs involved in setting the level of the intellectual property regime uniformly, see, e.g., Dan L. Burk & Mark A. Lemley, *Is Patent Law Technology-Specific?*, 17 BERKELEY TECH. L.J. 1155, 1159 (2002). *See generally* Michael W. Carroll, *One for All: The Problem of Uniformity Cost in Intellectual Property Law*, 55 AM. UNIV. L. REV. 845 (2006); Michael W. Carroll, *One Size Does Not Fit All: A Framework for Tailoring Intellectual Property Rights*, 70 OHIO. ST. L.J. 1361 (2009).

countervailing effects on works in different markets. Strengthening copyright protection in some respect (say by lengthening its duration) may allow the copyright owner to appropriate a larger portion of the social value of the work, thereby incenting the creation of some previously supramarginal works. The social value of these new works must be compared, however, to the deadweight loss added to previously marginal and inframarginal works. Similarly, when considering the effects of a certain copyright exemption, the benefit of reducing deadweight loss in regard to marginal works must be assessed against the possible cost of causing some works to become supramarginal due to declining profits of the copyright owner.

3. Product Differentiation

The economic effects of copyright are traditionally analyzed under the assumption that copyright's exclusionary power creates a monopoly.[32] More recent scholarship, however, suggests the more plausible model of monopolistic competition between differentiated products.[33] The basic insight is that non-identical works can be partial market substitutes for each other. For example, while a copyright owner may hold the exclusive right of reproducing and selling music album A, consumers may see music albums B, C, and D as partial substitutes for A. It follows that the copyright owner does not enjoy a monopoly in regard to her copyrighted work, as it is subject to competitive pressures from partial substitute products. Contrary to perfect competition, however, each copyright owner still enjoys some pricing power over marginal cost for her work. Indeed, in the absence of such pricing

32. *See, e.g.*, Fisher, *supra* note 24, at 1700–03; Michael O'Hare, *Copyright: When Is Monopoly Efficient?*, 4 J. POL'Y ANALYSIS & MGMT. 407, 411 (1985); Arnold Plant, *The Economic Aspects of Copyright in Books*, 1 ECONOMICA 167 (1934) (referencing the "copyright monopoly"). It is a standard practice by many writers to concede that the extent to which a copyright owner enjoys market power varies according to the availability of substitutes of its product. *See, e.g.*, Fisher, *supra* note 24, at 1702–03; Gordon, *supra* note 27, at 1367 n.76; Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEX. L. REV. 989, 1041 (1997); Meurer, *supra* note 27, at 82–83. It is also a standard practice, however, to proceed to analyze the economic effects of copyright as monopoly pricing following this concession. At the other extreme one finds assertions that since copyrighted works usually have close substitutes and the owner enjoys no monopoly, copyright involves no pricing power and should be analyzed under a competitive market model. *See* Edmund W. Kitch, *Elementary and Persistent Errors in the Economic Analysis of Intellectual Property*, 53 VAND. L. REV. 1727, 1734–38 (2000).

33. *See generally* Michael Abramowicz, *A New Uneasy Case for Copyright*, 79 GEO. WASH. L. REV. 1644 (2011); Michael Abramowicz, *A Theory of Copyright's Derivative Right and Related Doctrines*, 90 MINN. L. REV. 317 (2005) [hereinafter Abramowicz, *A Theory of Copyright's Derivative Right and Related Doctrines*]; Abramowicz, *supra* note 17; Christopher S. Yoo, *Copyright and Product Differentiation*, 79 N.Y.U. L. REV. 212 (2004).

power, copyright would have no incentive function.[34] The number of differentiated products and the extent of competition in this mixed model of monopolistic competition depends mainly on the size of the market, meaning the size of consumer demand available to producers relative to the cost of developing the work.[35] The larger the demand available to producers, the more entrants with differentiated products will be able to recoup their development cost and will enter the market.

The implications of the differentiated product competition model for the economic analysis of copyright are complex, and the conclusions that should be drawn from it about the optimal shape of copyright law are fiercely debated.[36] A full survey of this debate is beyond the scope of this Article. Here we briefly distill its core analytic insights and central, countervailing policy implications.

First, the product differentiation framework highlights economic effects of copyright different from the incentive and access effects at the heart of the traditional model, and it identifies new ways that all these effects interact with each other.[37] To see this, assume that at a certain base level of copyright protection a particular work—a novel, for example—is marginal in the sense that the profit generated by the copyright owner suffices exactly for incentivizing the creation of the work. Now assume that copyright protection is increased significantly. Under the traditional framework the result is an increase in the copyright owner's returns who now will enjoy "rents," that is, profits beyond what is necessary for inducing the creation of the work, and an increase in the deadweight loss suffered by consumers in regard to the work. On balance, the change may or may not be desirable depending on its total effects on other supramarginal, inframarginal, and marginal works. Product differentiation theory, however, predicts a different result. To the extent that the additional profits or rents made available by the increase in copyright protection suffice to cover the development cost of rival works, new entrants will develop partial substitutes for the work—in our example, other novels targeting roughly the same audience—and will enter the market.[38]

34. *See* Oren Bracha & Talha Syed, *Beyond the Incentive-Access Paradigm? Product Differentiation and Copyright Revisited*, TEX. L. REV. (forthcoming 2014) (manuscript at 8–10) (on file with authors).

35. Yoo, *supra* note 33, at 239.

36. *See id.*; *see also* Abramowicz, *supra* note 17; Bracha & Syed, *supra* note 34.

37. *See* Bracha & Syed, *supra* note 34, at 15–20.

38. Yoo, *supra* note 33, at 238 ("[T]he presence of supracompetitive profits attracts other authors selling close substitutes.").

What will be the result of such differentiated product competition? One negative effect will be the new development cost incurred by each new entrant. To a large extent, this cost is duplicative and wasteful because it is incurred simply in order to divert consumer demand that was already being satisfied by the original novel. But it is not necessarily completely duplicative. To the extent that the different variants are better tailored for the tastes of some consumers, demand is better satisfied. Thus a possible, positive effect of differentiated product competition is better demand satisfaction by the new product variants. Finally, the competition triggered by the increase in copyright protection is likely to have an effect on price and deadweight loss. The competition from new variants will likely reduce the price compared to what the copyright owner in the original work could charge under the new level of protection were he a monopolist, as assumed in the traditional framework. Accordingly, the levels of deadweight loss in this market will be lower than predicted by traditional analysis. The effect on deadweight loss in this market compared to the prior lower level of copyright is a more debatable question.[39] Depending on circumstances and assumptions, the increase in protection may result in decreased, increased, or unchanged levels of deadweight loss.[40]

4. Global Distortions

Assume that through a Herculean effort we managed to obtain all the relevant information and optimized every detail of copyright law under the framework of incentive/access in inframarginal and supramarginal markets modified by insights from product differentiation theory. Unfortunately, a major efficiency concern remains. Production of copyrightable works takes place against the backdrop of other actual and potential economic activities. A general-equilibrium efficiency perspective requires shaping copyright law in a way that minimizes the distortions in the global patterns of economic activity. Specifically, the copyright regime must not generate a rate of return on investment for copyrightable works that is significantly different from the rate of return for alternative economic uses of the same resources.[41]

39. *Compare* Yoo, *supra* note 33, at 254 ("[A]ccess should be promoted indirectly by stimulating entry and allowing the ensuing competition to drive price closer towards marginal cost."), *with* Bracha & Syed, *supra* note 34, at 28 ("[T]he deadweight loss effect of increased entry through increased copyright protection is by no means necessarily positive.").

40. Bracha & Syed, *supra* note 34, at 49 ("Under conditions of differentiated competition, strengthening copyright has an indeterminate effect on overall deadweight loss . . . .").

41. Lunney, *supra* note 28, at 490.

Assuming limited available capital, a rate of return on investment that is systematically higher in one area of economic activity compared to others will inefficiently attract investment in that area at the expense of others. This would be an inefficient distortion since optimally resources should be invested according to the net total social value generated by the investment, not according to the private profits to the producer. Yet in many realms—ranging from education to infrastructure to pioneering new markets—economic actors internalize far less than the full social value of their activities or products.[42] If the rate of return created by copyright—even locally optimized copyright—is significantly higher than the rate of return in many of the alternative fields where producers of copyrighted works could have invested their energy and capital, the result will be a distortive over-production in the field of copyright at the expense of those other fields. There will be too many books produced and not enough teaching, or too much investment in producing entertaining films and not enough investment in other forms of recreational services.[43] It is not enough, then, to optimize the ratio between copyright's incentive benefit and access cost and to avoid inefficient, excessive duplicative creation of similar works. The copyright regime must also be crafted so it does not inefficiently attract investment away from other economic activities by creating a significantly higher rate of return in the field.

B. CONSEQUENCE-SENSITIVE THEORIES OF COPYRIGHT

At this point a skeptic may be tempted to question the usefulness of the entire economic analysis project. The intricacies of the theoretical apparatus, difficulties in reconciling different parts of it, debates about how it should be applied to concrete copyright questions, and, above all, the sheer amount of unavailable but necessary empirical data seem to make the generation of clear, determinate answers a staggeringly difficult task.[44] Meanwhile, when proponents of the economic method take notice of theoretical rivals, those

42. *See* Amy Kapczynski & Talha Syed, *The Continuum of Nonexcludability and the Limits of Patents*, 122 YALE L.J. 1900, 1923–43, 1945–46 (2013).

43. *See* Fisher, *supra* note 10.

44. ROBERT P. MERGES, JUSTIFYING INTELLECTUAL PROPERTY 3 (2012) ("[T]he data are maddeningly inconclusive"); Fisher, *supra* note 10, at 180 ("With respect to incentive theory, the primary problem is lack of the information necessary to apply the analytic."). George Priest's pessimistic observation of almost three decades ago—"[I]n the current state of knowledge, economists know almost nothing about the effect on social welfare of the patent system or other systems of intellectual property"—only gathers force when we factor in the subsequent refinements of the basic incentive-access framework summarized here. George Priest, *What Economics Can Tell Lawyers About Intellectual Property*, 8 RES. L. & ECON. 19, 21 (1986).

rivals are often dismissed precisely because they both seem indifferent to empirical consequences and lack any theoretical apparatus for analyzing or predicting outcomes. Non-economic theories, that is, are frequently labeled "deontological" and correspondingly assumed to involve no consequential implications relevant to their normative dictates.[45]

Where are consequence-sensitive theories of copyright located in this divide? At a position closer to that of efficiency but not identical to it. Traditional rights-based justifications of copyright, such as labor-desert and personality theories, do not completely deserve the censure for being detached from consequences. Many variants of these theories actually incorporate elements that are consequential in nature at critical junctures.[46] The distinguishing feature of these traditional rights-based theories is not that they never care about consequences, but that their interest in consequences tends to be narrow and restricted. The range of consequences that are considered within these normative frameworks is limited by various stipulations, including: considering only consequences for the parties directly regulated by the rule; drawing categorical distinctions between actions (or "doing") and omissions ("allowing"); narrowly construing what counts as a "harm"; prescinding from forward-looking, or ex ante, effects of the rule; or similar devices.

By contrast, efficiency and the democratic and distributive theories with which we are concerned are *comprehensive* in their treatment of consequences.

---

45. *See, e.g.*, John Duffy, *Merges and Descartes*, PRAWFSBLOG (Jan. 30, 2013, 3:09 PM), http://prawfsblawg.blogs.com/prawfsblawg/2013/01/merges-and-descartes.html ("[I]f we are frustrated with the complexities of economic theories and are searching for a more solid foundation for justifying the rules of intellectual property, is Kant (or Locke or Rawls or Nozick) really going to help lead us out of the wilderness?"); Jonathan Masur, *The New Institutional Philosophy of Rob Merges*, PRAWFSBLOG (Jan. 29, 2013, 12:53 PM), http://prawfsblawg.blogs.com/prawfsblawg/2013/01/the-new-institutional-philosophy-of-rob-merges.html ("What would it mean for IP doctrine in practice not to have properly advanced Lockean or Kantian ethics? How could anyone tell? The problem—or, more accurately, the advantage for Kant and Locke—is that those approaches are purely theoretical and do not generate testable predictions.").

46. A good example is the interpretation of Lockean labor-desert theory as applied to copyright as elaborated and popularized by Wendy Gordon. *See* Wendy J. Gordon, *A Property Right in Self-Expression: Equality and Individualism in the Natural Law of Intellectual Property*, 102 YALE L.J. 1533, 1544–45, 1561 (1993). Gordon construes both the principle of ownership by a person who exerted his labor on the commons and the most significant limitation on it, known as the sufficiency proviso, as directly based on the no-harm principle under which people should be free to act as they wish as long as they do not inflict unjustifiable harm on others. The question of inflicting harm, especially as it applies to the sufficiency proviso, under which a property right is justified only as long as it leaves as much and as good in the commons so others are not worse off, is an inherently empirical question about real-world consequences.

These theories consider and assess (each theory according to its own distinct normative criteria) the broad range of consequences generated by a specific rule, including effects on parties not privy to the regulated behavior and general incentive effects of the rule. Labor-desert theory, for instance, may ignore the fact that a particular feature of copyright generates restrictions on access whose negative aggregate social value is greater than the value of the works incentivized by that feature, as a "harm" irrelevant to the theory.[47] But that is an effect that any theory comprehensive in its regard of consequences would take into account.

Despite their shared comprehensive regard for consequences, however, we can still draw a crucial distinction between efficiency and the democratic and distributive theories. Efficiency is a consequentialist theory, meaning that consequences are all that matter for purposes of its normative analysis.[48] By contrast, democratic and distributive theories are "consequence sensitive."[49] This means that consequences, comprehensively considered, constitute an important element of the normative analysis of these theories, but they are

47. *See* Gordon, *supra* note 27, at 1549.

48. "Consequentialist" theories may be broadly identified as those for which consequences are all that matter for normative evaluation. Or, more restrictively, as those that prescribe maximization of net-beneficial consequences. "Welfarist" theories comprise a subset of consequentialist views, which define the relevant consequences only in terms of their impact on individuals' well-being, which in turn may be understood in terms of hedonic experiences, preference satisfaction, or other criteria. Welfarist theories that care about the distribution of welfare across individuals are not consequentialist in the restrictive sense, but remain so in the broader sense. Efficiency's directive to maximize wealth or effective preference satisfaction is hence a welfarist theory that qualifies as consequentialist in both senses.

49. Somewhat more precisely, by "consequence-sensitive" (or, more technically, "consequence-supervenient") we mean to denote theories for which, simultaneously, (a) "consequences always matter" (pace views for which a normative difference between two states of affairs can be present without any difference in the relevant consequences), yet (b) "consequences are not all that matters" (pace consequentialist theories), such that normative evaluation may not always be reducible to information about the relevant consequences. For an elaboration of these distinctions and their significance for normative legal theory more generally, see Talha Syed, Is Welfare the Only Value? (Nov. 14, 2013) (unpublished manuscript) (on file with author). We note that some of the democratic and distributive theories we take up here may in fact not only be consequence-sensitive but full-blown consequentialist theories, and thus share with efficiency/welfarist views the commitment not only of always caring about consequences but also of caring only about consequences, while still diverging from such views in sometimes caring about consequences in terms other than their impact on individuals' welfare (especially when that is understood in the common economic sense of revealed subjective preferences). But other variants of such theories will not be consequence-reductive in this stronger sense, and share with economic analysis only the weaker commitment to consequence-sensitivity.

not its exclusive element. From this perspective, consequences matter a great deal, but they are not all that matters.[50]

Because of their shared comprehensive interest in effects of rules, consequence-sensitive theories and the full-blooded consequentialist theory of efficiency bear a strong structural similarity. Although the theories may normatively evaluate these effects differently, each of them cares about the entire range of multiple, potentially countervailing, effects on all individuals and must trade them off, or otherwise make value judgments across them, in some way. It follows that the same incentive/access tradeoff at the heart of efficiency analysis is also highly relevant for the application of all the consequence-sensitive theories. To be sure, each theory seeks to promote different goods and notions of fairness than wealth maximization. But the various economic effects produced by copyright are likely to be as relevant to any of these alternative commitments. If, for example, a particular democratic theory prizes public access to various creative works (for whatever reason), it would be short-sighted to recommend lowering copyright barriers to access while ignoring any possible incentive effects of such a measure. If the likely result of such a recommendation is a sharp drop in the creation of new works, the resulting decline in overall access to works will make the policy undesirable from a normative perspective internal to the relevant democratic theory. Thus there seems to be no a priori reason to assume that the basic incentive/access tradeoff is different from that underlined by efficiency. And mere incantations of "free speech" or "democracy" cannot change that.

The Supreme Court expressed this logic decades ago when it responded to an argument that access privileges should be interpreted capaciously with regard to works involving strong public interest, with the observation that: "It is fundamentally at odds with the scheme of copyright to accord lesser rights in those works that are of greatest importance to the public. Such a notion ignores the major premise of copyright, and injures author and public alike."[51] How then, if at all, do any of the democratic and distributive theories differ from efficiency (or from each other) when applied to concrete questions of copyright law and policy? The remainder of this Article responds to this question.

50. *See supra* note 49.

51. Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 559 (1985).

III. SELF-DETERMINATION, DEMOCRACY, AND FLOURISHING THEORIES

The striking structural affinity just adduced between efficiency and consequence-sensitive democratic theories presents a challenge, then, for the various normative views often lumped together under the labels of "free speech" or "democracy." Interestingly, scholars hailing from very disparate normative orientations have voiced similar versions of this challenge. Speaking from the vantage of economic analysis, Michael Abramowicz argues that democratic theories are virtually indistinguishable from efficiency, suggesting that democratic proponents have not managed to show that in any particular case "the democratic balance would be any different from the economic one."[52] Jennifer Rothman deploys similar logic to advocate a very different cause: a shift from free speech copyright policy arguments to arguments based on users' individual rights grounded in liberty and due process.[53] Since most free speech theories of copyright are cast in a way that incorporates the incentive/access framework, she argues, commentators and courts are led to "engage in broad utilitarian calculations of the overall speech markets rather than considering the legitimacy of individual users' speech claims."[54] This reduces the analysis to the utilitarian framework and typically results in the rejection of claims for restricting copyright protection due to a belief in copyright's function as "an engine of free expression"[55] and in its internal built-in balancing mechanisms.[56]

Are democratic theories of copyright really no different from efficiency? Are their proponents faced with the inevitable dilemma of either embracing their true character and coming into the fold of efficiency or, realizing this character, disdainfully switching allegiances? Given the fundamental, structural similarity to efficiency, the onus indeed lies with proponents of any specific variant of these theories to show why the democratic balance would be different from the economic one.

---

52. Abramowicz, *supra* note 17, at 92. Abramowicz refers specifically to Neil Netanel's theory of copyright and a democratic civil society. *See* Netanel, *supra* note 19. The challenge applies, however, more broadly to most variants of democratic, self-determination, and flourishing theories.

53. Professor Rothman's argument is made mainly in the context of First Amendment-based constitutional arguments in copyright law, but it appears to apply more broadly to the use of free speech or democratic theory in copyright policy analysis. *See generally* Jennifer E. Rothman, *Liberating Copyright: Thinking Beyond Free Speech*, 95 CORNELL L. REV. 463 (2010).

54. *Id.* at 485.

55. *Id.* at 484–85.

56. *Id.* at 481–84.

Three steps are necessary to demonstrate this difference. First, one must elaborate the specific ways in which democratic theories differ from efficiency and from each other in their normative commitments regarding crucial elements, such as the reason(s) why expressive activity is valuable and what kinds of expressive activities are valuable. Second, one has to explain why, in view of these different commitments, the same mechanisms that will maximize economic efficiency will not always also be the ones best suited to serve the different values or activities prized by a specific democratic theory. Put differently, each theory needs to explain why its prized desideratum requires an incentive/access tradeoff or prioritization different from the one mandated by efficiency. Third, each theory must justify this gap. It must explain why it is prepared to override, at least in some instances, individual subjective evaluations by preferring outcomes that do not maximize subjective preferences. It is only when elaborating these three elements and specifying how they apply to particular features of copyright that the stakes of the choice between efficiency and its democratic rivals, as well as of the choice among the rivals themselves, are made clear.

This Part responds to these three challenges in two moves. First, we provide a general taxonomy of the very different normative frameworks often lumped together under the banner of democratic or freedom-of-speech theories of copyright, isolating the key distinctive features that separate them. Second, we examine four copyright policy questions from the distinct perspectives of these theories, identifying how the various democratic frameworks support outcomes that diverge from efficiency in particular contexts, and how the divergence with maximizing satisfaction of subjective preferences may be justified.

A. MAPPING THE THEORIES

The various frameworks for normative analysis of copyright typically referred to together as "democratic" theories or under similar labels (such as "free speech") often differ significantly from one another in their core philosophical or normative tenets, as well as in their concrete policy prescriptions. Things are further complicated by the fact that the democratic theories circulating in existing scholarship frequently mix and match normative elements from different philosophical positions and sometimes remain unclear with respect to their central underlying commitments. What follows here is not a full survey of the various positions actually advanced in democratic copyright literature, or the way they map onto existing philosophical traditions. We provide, instead, an analytic taxonomy of the primary alternative normative views available regarding the role, in advancing

individual freedom and a desirable polity or society, of the expressive works and activities regulated by copyright.

A first possible approach would be to treat expression from the perspective of "negative liberty." Although a negative liberty view is sometimes implicitly assumed, it is rarely explicitly articulated in debates about copyright and free speech. From this standpoint, acts of expression—both in terms of accessing existing expressive materials and generating and disseminating new materials—have moral significance as fundamental acts of human individual liberty. As such, expression should be protected against coercive interference by others, whether private individuals or governments. Individuals, in other words, have the "negative" right to be left alone, free from any active, coercive interference by others that would limit their choices and actions with respect to expressive works or activity.[57] The hallmark of this position is an exclusive focus on coercive interference by others and a disregard for effective agency, meaning the endowment of individuals with the effective means and opportunities for actually engaging in the valued expressive activities. Whatever their merits, such positions are not consequence-sensitive in the sense we specify above and thus lie outside the scope of this Article. Leaving them aside, there remain four main normative frameworks with significant implications for the expressive works and activities regulated by copyright.

57. This is not a very pervasive position in copyright scholarship. For an early analysis of liberty claims for users' rights in copyright policy and some of their difficulties, see Jeremy Waldron, *From Authors to Copiers: Individual Rights and Social Values in Intellectual Property*, 68 CHI.-KENT L. REV. 841, 868–75 (1993). Two examples of thoughtful positions that fit this mold in the copyright policy debate are Edwin Baker's analysis of free speech limitation on copyright and Jennifer Rothman's due process and liberty framework for rights claims against copyright protection. Professor Baker argues that self-expression liberties create absolute entitlements to individuals to express and copy copyrighted expression by others (but not to commercially benefit from it). Consistent with his general theory of the First Amendment, however, Professor Baker distinguishes between freedom of speech and freedom of the press. The former protects expression as an individual human liberty and the latter is based on a more institutional vision of the role of speech in democracy. *See* C. Edwin Baker, *First Amendment Limits on Copyright*, 55. VAND. L. REV. 891 (2002). Jennifer Rothman similarly advocates an individual liberty framework but calls for abandoning the framework of free speech altogether in favor of a theory of individual rights grounded in liberty and due process. *See* Rothman, *supra* note 53. Professor Rothman's account remains somewhat unclear whether it is based on a negative liberty conception or incorporates a broader idea of effective agency.

1. Self-Determination

A theory quite distinct from speech as negative liberty is the theory of expression as an important component of individual self-determination.[58] Self-determination here refers to the ability of individuals reflectively to form and revise their own conception of the good, and effectively to pursue a life plan for realizing it.[59] The ideal is sometimes described as self-authorship: being able freely to choose or affirm one's own ends in life and to work toward effectively achieving them.[60] Two elements are central to this conception. First, although it also values negative liberty as the "freedom from" coercive interference by others, it places equivalent importance on "freedom to" effectively pursue one's ends. This emphasis on freedom as effective agency means a commitment to ensuring individuals have equitable access to effective means and robust opportunities for pursuing and realizing their goals, rather than merely to abstaining from coercive restrictions.[61] Second, the ultimate value in self-determination resides in persons' ends being truly their own, meaning reflectively adopted, chosen, and refined by themselves, determining their projects and preferences free from both coercion and manipulation.[62] This entails taking seriously the idea of free choice and the conditions it requires. These conditions include a meaningful range of options from which to choose, the capacity to understand and evaluate the options, and the opportunity and means to critically reflect upon and possibly revise one's choices and sense of attractive alternatives.

---

58. The most elaborate discussion of self-determination in the context of copyright is found in YOCHAI BENKLER, THE WEALTH OF NETWORKS 133–74 (2006) [hereinafter BENKLER, THE WEALTH OF NETWORKS]; *see also* Yochai Benkler, *Siren Songs and Amish Children: Autonomy, Information, and Law*, 76 N.Y.U. L. REV. 23 (2001) [hereinafter Benkler, *Siren Songs*]. For an early application of self-determination to copyright, see Waldron, *supra* note 57, at 875–77.

59. On self-determination in this sense, *see* GERALD DWORKIN, THE THEORY AND PRACTICE OF AUTONOMY (1988); JOEL FEINBERG, HARM TO SELF 33–35 (1986); JOSEPH RAZ, THE MORALITY OF FREEDOM 369 (1986); Benkler, *Siren Songs*, *supra* note 58, at 33–35; John Rawls, *Kantian Constructivism in Moral Theory*, 77 J. PHIL. 515, 525 (1980). For our purposes, "self-determination" so conceived is roughly synonymous with "autonomy," although to avoid confusion with a negative-liberty framework and some related difficulties associated with the term autonomy, we prefer self-determination. For a contrasting development of the implications of a somewhat different "Kantian" conception of autonomy for intellectual property rights, see MERGES, *supra* note 44, at 72–83.

60. RAZ, *supra* note 59, at 370. Note that choosing one's own ends in life does not mean creating ex nihilo values and conceptions of the goods but rather a reflective endorsement of largely predefined possibilities. *See* DWORKIN, *supra* note 59, at 20; FEINBERG, *supra* note 59, at 33–35 (discussing "self-creation").

61. *See* DWORKIN, *supra* note 59, at 18 (arguing that the idea of autonomy includes some ability to make one's preferences effective).

62. RAZ, *supra* note 59, at 377.

Within this framework, expression—meaning both access to speech and one's ability to speak—plays a critical role. Expressions of various kinds may simply be part of the preferences and goals individuals have, for whose realization they must be given meaningful opportunities. Expression, however, may also be constitutive of individual aims rather than simply instrumental to achieving them. As a communicative activity, expression is deeply entwined with the conditions for genuine free choice and formation of individuals' outlooks. Expression is the process through which individuals form their preferences. It exposes individuals to competing options, allows them to articulate their own views and subject them to critical evaluation by others. Expression supplies the process through which people compare, reflect upon, and revise their choices. Finally, systematic control over the expressive content that individuals can access and disseminate tramples on the autonomy of such individuals by constraining and manipulating the process through which they form their preferences.[63] Thus, expression is not just one among many individually valued goods whose realization should be effectively enabled.[64] A robust and diverse expressive environment, one that is relatively free from control and manipulation by others and offers many opportunities for critical reflection and revision of one's commitments, is a critical enabling condition for meaningful self-determination.

2. Political Democracy

The starting point for political democracy theories[65] is that certain decisions—those that happen in political space, as it is usually termed—are

63. DWORKIN, *supra* note 59, at 18 (discussing subverting "people's critical faculties"); RAZ, *supra* note 59, at 377–78; Benkler, *Siren Songs*, *supra* note 58, at 38.

64. The early analysis by Jeremy Waldron of autonomy as applied to copyright seems unduly focused on expression as the subject matter of some individual free choices and as the means for achieving them, while neglecting the role of the expressive environment in constituting the conditions for genuine self-determination. Arguably this is the reason why Waldron, at least as an initial matter, reaches the conclusion that "the value of autonomy settles nothing" as between conflicting autonomy claims of authors seeking to protect the integrity of their expression and copiers seeking to express themselves through secondary uses. *See* Waldron, *supra* note 57, at 877. Similarly Jed Rubenfeld argues that autonomy does not explain why freedom of speech protects "the reading of books by persons for whom such reading is not an act of self-expression or self-realization." Jed Rubenfeld, *The Freedom of Imagination: Copyright's Constitutionality*, 112 YALE L.J. 1, 35 (2002). These arguments miss the fact that expressive activity is not just part of the content of individual free choice, but a crucial precondition for its exercise.

65. The main applications of this theory to the copyright context are in NEIL NETANEL, COPYRIGHT'S PARADOX 19 (2008); Niva Elkin-Koren, *Cyberlaw and Social Change: A Democratic Approach to Copyright Law in Cyberspace*, 14 CARDOZO ARTS & ENT. L.J. 215, 224–32 (1996); Netanel, *supra* note 19.

inherently collective or irreducibly social in nature. A commitment to individual self-determination requires procedures for making collective decisions that account for and enable the autonomy of all members of society in such contexts.[66] To achieve this aim, democratic processes must meet two requirements. First, they must give equivalent weight to the input of all individuals by affording each an effective opportunity to participate on equal terms in the collective decision-making process.[67] Second, democratic processes must reflect, with respect to collective decisions, the same regard for meaningful reflection and deliberation familiar from the context of individual choice.

Just as in the individual context, expression is a crucial element for collective self-determination, only now the focus is on the particular subset of expression that takes place in the realm relevant for public decision-making, the public sphere. Expressive activity supplies individuals with alternative views and choices, as well as information and tools for evaluating the competing options. Public expressive fora are where interests are defined and agendas are formed and subjected to critical review.[68] Participation in the public sphere is also where individuals develop the skills and habits of self-determination necessary for agency as a citizen, including the propensity to seek information, capacities of critical discernment and reflection, and the ability to articulate and communicate one's own views.[69] Thus, fostering an open, diverse, and active expressive public sphere as the locale of "collective will formation"[70] is as essential to collective self-determination as a vibrant expressive environment is for individual autonomy.

3. Cultural Democracy

How deeply into the realm of culture, beyond what is plausibly considered political expression in the narrow sense, does political democracy theory extend? One prevalent view seems to be: quite deeply. The cultural

66. *See* Robert Post, *Equality and Autonomy in First Amendment Jurisprudence*, 95 MICH. L. REV. 1517, 1524–25 (1997).

67. As Robert Post explains, an adequate democratic process that manages to create "a linkage of individual and collective autonomy" may require much more than nominally equal vote with regard to the collective decision. *See* Post, *supra* note 66, at 1526. It may require equal opportunity for robust participation in the political and discursive process by all individuals.

68. *See, e.g.*, NETANEL, *supra* note 65, at 38.

69. *See* Netanel, *supra* note 19, at 343–44.

70. JURGEN HABERMAS, BETWEEN FACTS AND NORMS: CONTRIBUTIONS TO A DISCOURSE THEORY OF LAW AND DEMOCRACY (1992); Jurgen Habermas, *Towards a Communication-Concept of Rational Collective Will-Formation: A Thought Experiment*, 2 RATIO JURIS 144 (1989).

sphere broadly defined, this argument goes, is highly relevant for the process of individual and collective self-determination.[71] It is where agendas are set through the articulation of various alternative options and views with regards to both individual life plans and collective decisions, even when no core political speech is involved.[72] As Neil Netanel puts it, "Homer Simpson and Mickey Mouse . . . may be used to subvert (or buttress) prevailing cultural values and assumptions, and with greater social impact than the most carefully considered Habermasian dialogue."[73] As a result, Netanel argues, a "vast swath of copyright-impacted expression fits well within the free speech domain" even when it is "creative expression that is neither overtly political nor rationally apprehensible."[74] This claim, however, seems to reach too far. One can recognize the importance of the cultural sphere for collective self-determination, and the consequent extension of political-democratic notions deep into this domain, yet still conclude that at some point the argument runs out of steam. At some point the connection between cultural expression and the subject matter of the formal political process seems too tenuous or intractable.[75]

Does it follow that at this point any concern about collective self-determination dissipates? Not necessarily. *Cultural*, as opposed to *political*, democracy extends a concern with collective self-determination into the cultural sphere that remains viable long after the plausible claims of political democracy run out. The starting point of this theory is the assumption of deeply socially situated selves, or socially constructed subjects.[76] On this

71. *See* NETANEL, *supra* note 65, at 32–33.

72. *See id.* at 33. This is a version of Meiklejohn's "art in the ballot booth" argument. *See* Alexander Meiklejohn, *The First Amendment Is an Absolute*, 1961 SUP. CT. REV. 245, 255–57.

73. *See* NETANEL, *supra* note 65, at 33.

74. *Id.*

75. For a similar argument, see Rubenfeld, *supra* note 64, at 33 (stating that protecting art on the basis of its importance for forming political views is to exaggerate its "political significance while instrumentalizing it, making it carry democracy's water").

76. *See* BENKLER, THE WEALTH OF NETWORKS, *supra* note 58, at 274 (observing that "[a]s individuals and as political actors, we understand the world we occupy, evaluate it, and act in it from within a set of understandings and frames of meaning and reference that we share with others"); WILLIAM FISHER, PROMISES TO KEEP: TECHNOLOGY, LAW AND THE FUTURE OF ENTERTAINMENT 30 (2004) (referring to "[t]he increasingly dense cloud of images, sounds, and symbols through which we move"); Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 35 (2004) (arguing that "democratic culture is valuable because it gives ordinary people a say in the progress and development of the cultural forces that in turn produce them"); Rosemary J. Coombe, *Objects of Property and Subjects of Politics: Intellectual Property Laws and Democratic Dialogue*, 69 TEX. L. REV. 1853, 1858 (1991) ("What we experience as social reality is a constellation of cultural structures that we ourselves construct and transform in ongoing practice."); Elkin-Koren, *supra* note 65, at 233 (arguing that "social agents are engaged in an

perspective, the identities, beliefs, and world-views of individuals are all deeply shaped by the social relations they are enmeshed in and the widespread meanings circulating around them. Culture here refers to the irreducibly interactive or social processes through which meanings are forged, communicated, enacted, interpreted, adapted, challenged, revised, recombined, and so forth. In a word, it is the "semiotic" realm of meaning-making, the "fog of symbols in which we move and with which we define ourselves,"[77] shaping the subjectivity of the individuals participating in or exposed to it.[78] At the same time, far from being passive receptacles for fixed and constant cultural forms, individuals have the capacity for actively interacting with culture—to develop, change, and subvert it.[79] On this view, culture—as the arena for the dialectic of constructing subjectivity, where individuals are shaped by a surrounding symbolic environment that in turn they may contribute to shaping—is crucial for preference formation in a deep sense.[80] Rather than being an important adjunct to the formal political realm by providing the input for the political process, culture simply *is* political in a broader sense, being a realm of irreducible social interactions having powerful impact on individuals' sense of themselves.[81]

So conceived, cultural democracy carries two distinct prescriptive implications. The first is a robust extension of the concern with individual self-determination into the cultural realm, ensuring that individuals have meaningful agency vis-à-vis the dense fog of symbolic materials suffusing their environment to enable *effective participation in the semiotic shaping of one's own*

on-going process of constructing the meaning of symbols" and thereby "give meaning to the objective world and define their own identity").

77. William W. Fisher, *The Implications for Law of User Innovation*, 94 MINN. L. REV. 1417, 1458–59 (2010).

78. Balkin, *supra* note 76, at 35 ("Culture is a source of the self.").

79. FISHER, *supra* note 76, at 30; Coombe, *supra* note 76, at 1863–64.

80. Balkin, *supra* note 76, at 35 (explaining that "[a] democratic culture is valuable because it gives ordinary people a fair opportunity to participate in the creation and evolution of the processes of meaning-making that shape them and become part of them"). *But cf.* Robert P. Merges, *Locke Remixed ;-)*, 40 UC DAVIS L. REV. 1259, 1265, 1267–68 (2007) (arguing that mass media "content is not strictly necessary for 'identity formation' ").

81. MADHAVI SUNDER, FROM GOODS TO A GOOD LIFE: INTELLECTUAL PROPERTY AND GLOBAL JUSTICE 64 (2012) (positing that "[d]emocratic culture may be even more important than democratic politics"); Balkin, *supra* note 76, at 35 ("A 'democratic' culture means . . . much more than democracy as a form of self-governance. It means democracy as a form of social life . . . in which ordinary people gain a greater say over the institutions and practices that shape them and their futures."); Elkin-Koren, *supra* note 65, at 231 ("[Democratic] participation is no longer confined to a narrowly defined political realm, but is instead perceived as an activity that can be realized in the social and cultural spheres as well. . . . [P]olitical will-formation . . . ought to be governed by democratic principles.").

subjectivity. Second, and what truly distinguishes cultural democracy from cultural self-determination, is a commitment to decentralizing meaning-making power on grounds of equalizing individuals' effective opportunity to participate in the semiotic shaping of *others'* subjectivity.

4. Human Flourishing

A final family of views moves beyond the foregoing theories' focus on enabling the conditions for robust individual and social self-determination, and advances a substantive conception of the good life or of human flourishing.[82] While not going so far as to dictate the details of individual or collective choices, theories of human flourishing insist that there are some ways of living one's life that are more fulfilling and allow for better realization of our "truly human" capacities and needs.[83] A just and attractive society, in turn, is one that fosters the social conditions conducive to flourishing lives, and equitably provides opportunities for all its members to realize their human capacities.

Not surprisingly, the alternative approaches in this family vary in their precise specification of the central features of the good life and the components essential to human flourishing. Nevertheless, most theories in this vein do tend to converge on a number of similar or common dimensions.[84] And for our purposes, it is possible to distill three central elements of convergence (without claiming to capture the entire range or nuances of the various alternative views). The first dimension, which largely incorporates the self-determination approach, emphasizes the importance of reflectively and deliberatively forming one's own conception of valuable aspirations, projects, and preferences, and of having the effective means for their pursuit and realization.[85] The second is "meaningful activity"—activity in which one's physical or cognitive capacities are highly engaged and developed in a manner involving challenge and discipline, an engagement valuable both for its own sake and for the realization of various intrinsic and

82. *See* Oren Bracha, *Standing Copyright on Its Head? The Googlization of Everything and the Many Faces of Property*, 85 TEX. L. REV. 1799, 1842–49 (2007) (discussing human flourishing under the title of "cultural democracy"); Fisher, *supra* note 77, at 1463–72; Fisher, *supra* note 24, at 1744–66 (discussing a "[u]topian [a]nalysis" of copyright).

83. *See* MARTHA C. NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT: THE CAPABILITIES APPROACH 13 (2000) (anchoring the capabilities approach in "the Marxian/Aristotelian idea of truly human functioning"); Fisher, *supra* note 24, at 1746.

84. *See* Bracha, *supra* note 82, at 1844–48; Fisher, *supra* note 77, at 1468–71.

85. Bracha, *supra* note 82, at 1846–47; Fisher, *supra* note 77, at 1468; Fisher, *supra* note 24, at 1747–50.

social rewards "internal" to the activity.[86] The third is sociality, meaning involvement in relations and activities whose character and meaning is established communally through interaction with others—with such affiliations and collaborations being both intrinsically valuable and partially constitutive of one's sense of self.[87]

A question that may naturally arise is why, if the latter two goods are indeed valuable components of the good life, their pursuit cannot be left to individuals' private choices—which would allow policy decisions to remain neutral between competing individual visions of the good life?[88] The answer, in brief (with elaborations given below), is that—in certain contexts—sole reliance on individuals' free choice is, for various reasons, either insufficient or unpersuasive for resolving a policy question. In such cases, the best route is to forthrightly acknowledge and defend the adoption of some substantive conception of the good.

Expressive works and activities are significant on this flourishing view principally in two distinct ways. First, expressive works may be instrumental as means or preconditions for achieving many of the elements of human flourishing. Consider a familiar example, the relation between a vibrant and diverse sphere of expressive communication and self-determination. Second,

86. "Internal" rewards refer not only to the pleasures intrinsic to the process of undertaking an activity (such as the joys of physical exertion, mental stimulation, and collaborating with others), but also the satisfactions associated with realizing, and being recognized by others to realize, the activity's aspirational ends, those ends associated with the activity's particular value or "virtue" as a form of human undertaking (such virtues ranging from notions of performative or competitive excellence, to aesthetic and philosophical ideals of originality, knowledge, beauty/form, craftsmanship, and social contribution). For illuminating philosophical accounts of this Aristotelian-Marxian notion see ALASDAIR MACINTYRE, AFTER VIRTUE 181–204 (1981); Fisher, *supra* note 24, at 1747–50; Richard W. Miller, *Marx and Aristotle: A Kind of Consequentialism*, *in* MARX AND ARISTOTLE: NINETEENTH CENTURY GERMAN SOCIAL THEORY AND CLASSICAL ANTIQUITY 275–83 (George E. McCarthy ed., 1992).

87. *Cf.* Fisher, *supra* note 77, at 1470 (emphasizing the importance of "participation in freely chosen communities as crucial to human flourishing"); Fisher, *supra* note 24, at 1753 (identifying the importance for individuals to be grounded in "communities of memory" in order to possess the "secure sense of self" necessary for self-determination, as well as the importance of access to " 'constitutive' group affiliations" for persons to self-realize).

88. Such "neutrality about the good" is of course a hallmark of liberal theories, whether of welfarist or alternative stripe. *See, e.g.*, KAPLOW & SHAVELL, *supra* note 15, at 24–27 (arguing that judgments of social choice should be exclusively a function of individuals' well-being, understood in terms of subjective preference satisfaction); JOHN RAWLS, A THEORY OF JUSTICE § 60, at 347–50 (rev. ed. 1999) (stating that principles of liberal justice must rely on a "thin theory of the good"); RONALD DWORKIN, *Liberalism*, *in* A MATTER OF PRINCIPLE 181, 191 (1995) ("[P]olitical decisions must be, so far as is possible, independent of any particular conception of the good life, or of what gives value to life.").

expressive activity will also often be intrinsically valuable as a particular realm of human activity where key components of human flourishing are directly realized. For instance, expressive activity will often—perhaps typically—comprise one important subset of human practices that constitute meaningful activity. Similarly, sociability often unfolds within and through expressive activities; not only when such activities are undertaken in collaborative ways but also, significantly, when interaction with expressive works—their consumption, interpretation, sharing, etc.—happens in group settings.

Each of the theories just distilled focuses on a different set of values, prizing expression for somewhat different reasons. Do these differences, however, have implications for purposes of copyright policy? The question is twofold. First, given the basic structural similarity between all of these consequence-sensitive theories and efficiency, is it not still the case that "democratic benefits are roughly proportional to economic benefits"?[89] One still needs to demonstrate how each theory may produce different results than efficiency in specific cases and how, in such cases, the fact that they may override subjective individual preferences can be justified. Second, does the choice among the theories matter? That is, when, if ever, do these theories point in divergent directions from one another, or offer different justifications for overriding preferences in ways that matter for copyright policy?

B. THE THEORIES IN ACTION

To illustrate how the foregoing abstract commitments of democratic theories bear on concrete policy questions, we discuss four familiar issues of copyright law: the appropriate treatment of critical uses of copyrighted works, the entitlement to make derivative works, the application of copyright to fan fiction, and the status of personal uses of copyrighted materials. Our discussion is not meant exhaustively to survey how all the foregoing theories apply to each of these four topics. Rather, we use the theories selectively to demonstrate how, in different contexts, the outcomes they mandate may diverge from efficiency, as well as, at times, each other.

1. Critical Secondary Uses

To what extent, if at all, should a copyright holder's entitlements apply to uses of his or her work that are critical in character, drawing heavily on the copyrighted work but in a manner that significantly transforms its overall meaning? Or, conversely, to what extent should such critical uses be immune

89. Abramowicz, *supra* note 17, at 94.

from liability under the fair use doctrine?[90] One classic and much discussed example is the comic strip *Air Pirates Funnies*.[91] The short-lived series (consisting of only two issues) involved close reproduction of Walt Disney characters, which were presented in bawdy adult contexts, such as drug use or sexual activities. The makers of the comics claimed to be engaged in criticism of an American culture that they regarded as unduly conformist and hypocritical.[92] Disney sued the makers, who based their defense on a fair use argument but eventually lost.[93] Another more recent example is the novel *The Wind Done Gone*.[94] According to its author Alice Randall, the novel is a critique of the romanticized portrayal of slavery and the Civil War era depicted in Margaret Mitchell's *Gone with the Wind*.[95] To achieve this effect, Randall drew heavily on the original novel, using many of its characters, plot lines, and other creative elements. Despite this heavy borrowing, however, *The Wind Done Gone* tells a very different story in a very different voice. And, stretching the meaning of the term, the Eleventh Circuit held that Randall's use was parodic.[96] Partly as a result of this holding, the court ruled that the

90. Fair use is simply the most readily available mechanism for immunizing critical works from copyright liability under existing doctrine. However it may not, at least in its current form, be the optimal means for doing so. This doubt is traceable to an ongoing debate about the efficacy of the fair use doctrine as copyright's main "safety valve." Views skeptical of the efficacy of fair use as a reliable ex ante assurance to users emphasize the flexible, ad hoc, and unpredictable nature of the doctrine that often makes reliance on it risky and unappealing, especially to parties who lack substantial financial backing. *See, e.g.*, Michael W. Carroll, *Fixing Fair Use*, 85 N.C. L. REV. 1087, 1090 (2007) (arguing that fair use "is so case-specific that it offers precious little guidance about its scope"); Fisher, *supra* note 24, at 1693–94 (discussing the harmful effects of uncertainty in the fair use doctrine); John Tehranian, *Whither Copyright? Transformative Use, Free Speech, and an Intermediate Liability Proposal*, 2005 BYU L. REV. 1201, 1215–16. Other scholars take the position that properly understood and applied fair use can yield a sufficient level of predictability and avoid chilling effects. *See, e.g.*, Neil Weinstock Netanel, *Making Sense of Fair Use*, 15 LEWIS & CLARK L. REV. 715, 718 (2011) (arguing that recent empirical studies have found "patterns in fair use case law that give the doctrine some measure of coherence, direction, and predictability"); Pamela Samuelson, *Unbundling Fair Uses*, 77 FORDHAM L. REV. 2537, 2541 (2009) (arguing that once common patterns are identified, fair use case law is "both more coherent and more predictable than many commentators have perceived").

91. *See, e.g.*, BOB LEVIN, THE PIRATES AND THE MOUSE: DISNEY'S WAR AGAINST THE COUNTERCULTURE (2003); NETANEL, *supra* note 65, at 19; Gordon, *supra* note 46, at 1601–05.

92. *See* LEVIN, *supra* note 91, at 78 (describing counterculture's "anti-Disney line").

93. Walt Disney Prods. v. Air Pirates, 581 F.2d 751, 760 (9th Cir. 1978).

94. Suntrust Bank v. Houghton Mifflin Co., 268 F.3d 1257 (11th Cir. 2001).

95. *Id.* at 1259.

96. *See id.* at 1269.

use was allowed under the fair use doctrine as applied in light of First Amendment principles.[97]

Perhaps there are persuasive ways of distinguishing the two cases and reconciling their seemingly divergent results. And in any case, fair use doctrine did develop in the interim between them (and has continued to since).[98] Nevertheless, the basic policy question remains: how should copyright treat cases of close, extensive reproduction of copyrighted materials when said reproduction is done in the service of creating a new work that is materially different from the original and critical of it or of the views or values with which it is commonly associated?

From an efficiency perspective, the answer is highly uncertain. It depends on the net result of several distinct effects of treating critical uses as infringing. First, to the extent that unauthorized critical uses erode the incentive for producing originals, including them within the scope of copyright would provide the benefit of created works that otherwise would be supramarginal. An erosion of incentives may result from lost licensing fees, if there are any cases where a deal may have been struck, or, perhaps more plausibly, from any dampening of the demand for the original, resulting from its critical treatment by the secondary use. Of course, it seems very unlikely that being deprived of (the highly speculative) licensing revenue from critical uses or any tarnishing effect would make Mickey Mouse or *Gone with the Wind* unprofitable or prevent their creation. Nevertheless, it is possible that some other works, with lower profit margins, would be pushed from being marginal to supramarginal and hence, not be created. Perhaps there would be no Mighty Mouse, or perhaps some very expensive motion picture would not be produced. In addition, the inframarginal effects of possibly reduced profitability, even if by a small amount, should be considered. A lower surplus available to producers may result in fewer entries by producers of close substitutes for the original successful works. Perhaps there will be no Wile E. Coyote, or we will go without some other grand romantic novel set on a historical stage. The result would be that predicted by product differentiation theory: complex effects on duplicative cost,

97. *Id.* at 1268–70.

98. The most important developments are the rise in importance of transformative uses, and the especially favorable treatment accorded parodic uses. *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) ("[T]he goal of copyright, to promote science and the arts, is generally furthered by the creation of transformative works. Such works thus lie at the heart of the fair use doctrine's guarantee of breathing space within the confines of copyright . . . and the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use." (internal citation omitted)).

tailored demand satisfaction, and deadweight loss.[99] Finally, prohibiting critical uses would result in an inframarginal access cost regarding the secondary uses. Some efficient secondary uses may be frustrated by the transaction costs of obtaining a license or their inability to internalize positive externalities of the use, while for others, the added cost of the licensing fee may translate into a higher price and hence, an increased deadweight loss. The analysis is further complicated by competing arguments regarding likely levels of transaction costs and types of market failures relevant to the fair use analysis.[100]

Our present task is not to conclusively decide the efficiency debate; rather, it is to examine whether any of the democratic theories add something to the analysis. Recall that these theories should also be interested in the countervailing costs on expression with which the expressive benefit of an exemption to critical uses is bought. Should they not, then, apply the same calculus as efficiency to the tradeoffs among these effects to reach similar conclusions?

To see how, in fact, these theories might depart significantly from efficiency's prescriptions, consider self-determination. The emphasis here, again, is not only on individuals being free from coercive interference or possessing the means to effectively pursue their given projects and preferences, but also on ensuring sufficiently robust conditions for individuals to freely form and revise their outlooks. Self-determination emphasizes conditions that enable and foster the exercise of the capacities to evaluate, choose among, and reconsider a wide variety of projects and preferences, or distinct ways of living, thinking, feeling, etc. In this regard, a high normative premium is placed on critical expression due to its role in

99. *See supra* Section II.A.3.

100. *Compare* Richard A. Posner, *When Is Parody Fair Use?*, 21 J. LEGAL STUD. 67, 69 (1992) ("A use is fair . . . when the costs of transacting with the copyright owner over permission to use the copyrighted work would exceed the benefits of transacting."), *with* Tom W. Bell, *Fair Use vs. Fared Use: The Impact of Automated Rights Management on Copyright's Fair Use Doctrine*, 76 N.C. L. REV. 557, 581 (1998) ("[A]s a response to market failure, the fair use doctrine can and should give way in the face of the effective enforcement of authors' rights through automated rights management."), Wendy J. Gordon, *Market Failure and Intellectual Property: A Response to Professor Lunney*, 82 B.U. L. REV. 1031, 1034 (2002) ("Transaction cost barriers are neither the only kind of economic problem to which fair use responds, nor the only kind of problem to which fair use *should* respond."), *and* Lydia Pallas Loren, *Redefining The Market Failure Approach to Fair Use in an Era of Copyright Permission Systems*, 5 J. INTELL. PROP. L. 1, 6 (1997) ("A permission system only remedies the market failure that occurs because of high transaction costs. . . . [But it] does nothing to cure . . . the market failure that occurs when there are significant external benefits associated with a particular use that cannot be internalized in any bargained-for exchange.").

encouraging critical reflection on, and presenting alternatives to, existing options, especially those that are dominant or widely held. Moreover, ensuring steady exposure to such expression is also likely to cultivate the faculties needed to exercise meaningful reflective choice, since "[t]he mental and moral, like the muscular powers, are improved only by being used."[101] Thus, critical expression is particularly important to preserve because it is an essential component of the conditions conducive to individual self-determination.

To see the significance of this normative premium, consider the following hypothetical. Assume first that we could calculate, to the dollar, the various economic effects of exempting works such as *Air Pirates* or *The Wind Done Gone* from copyright liability. Further assume (perhaps implausibly) that total consumer demand for works not created because of the exemption significantly outweighs consumer demand for the critical works that, without the exemption, would either not be created or would be less accessible.[102] From an efficiency perspective this is a no-brainer: the exemption would clearly generate more cost than benefit and thus should be rejected. Not so for self-determination theory. From this perspective, the critical expression lost or to which access is curtailed is more valuable in comparison to other expression that enjoys higher consumer demand but does not offer the same benefits for ensuring conditions of genuine choice. Better a world with the *Air Pirates* and no Mighty Mouse than a world with Mighty Mouse but no *Air Pirates*, even if consumers are willing to pay more for Mighty Mouse.

This raises with force the issue of overriding or discounting subjective individual preferences. Michael Abramowicz puts the point sharply: "If . . . economists are correct in seeing property rights as 'putting existing works of authorship to their most socially valued uses,' " then insisting "that the works be placed in the public domain nonetheless is perverse."[103] In other words, why should we place greater value on critical uses than that indicated by actual individual preferences? The answer: because ensuring ample space for critical expression is crucial to nourishing the conditions of genuine free choice, and that, in turn, is a higher-order concern compared to satisfying existing preferences, being as it is a precondition for valuing or deferring to such preferences to begin with. Recall the grounding normative commitments of self-determination theory: the reason why this theory values

101. JOHN STUART MILL, ON LIBERTY 55 (Batoche Books 2001) (1859).

102. To simplify, ignore any effects of the exemption on inframarginal competition from substitutes, or assume they are insignificant by comparison to the effects on supramarginal and critical works.

103. Abramowicz, *supra* note 17, at 91.

individual preferences in the first place is to respect and enable individual autonomy, in the sense of persons authoring their own lives. And individual choices (along with the preferences revealed by them) are autonomous only to the extent that the conditions exist for reflective, discerning choice among a meaningful array of alternatives. Our interest, then, in the conditions for genuine free choice is a higher-order one than our interest in satisfying any given set of existing preferences, meaning that it is both qualitatively distinct from, and normatively prior to, the latter. The conditions that shape, and validate deference to, individual preferences cannot themselves be evaluated simply by a given set of existing preferences. Thus, when a specific feature of copyright has substantial effect on the higher-order conditions that validate individual preferences, there is nothing perverse in not subjecting this feature to the test of existing preferences.

To be sure, the conclusion would be different if it was plausibly predicted that exempting certain critical secondary uses would have devastating effects on expression. For example, if exempting the likes of *Air Pirates* would undermine the creation of a substantial share of those expressive works potentially subject to such critical uses (which is quite unlikely), then the self-determination evaluation might well shift. We would have to recalibrate our sense of the appropriate balance between protecting one of the conditions that validate deference to preferences (since availability of critical expression is not the only factor influencing whether free choice exists) and enabling the satisfaction of a given set of preferences. Striking that balance strictly in favor of protecting the preconditions might not only result in an unattractive mix of enabling sound formation versus effective satisfaction of preferences. Indeed, it may also be internally self-defeating: even from the sole point of view of improving the conditions of preference formation, it may be a net deficit to facilitate one set of critical uses at the cost of undermining a very large chunk of expressive output that may itself offer a variety of alternative perspectives (which, in turn, may subsequently serve as the foil for further secondary critical treatments).

There remains, however, a significant gap between a normative analysis under efficiency versus one under self-determination. In cases where self-determination plausibly predicts that a certain legal feature is likely to be significantly conducive to an environment closer to the ideal of free choice, it would support this feature even if a precise efficiency calculus is either unavailable or goes the other way. Moreover, even when the two normative perspectives point in the same direction there may be a difference in intensity. It is possible, for example, that in light of claims of new cost-

effective mechanisms for licensing transactions,[104] efficiency only supports the exemption by a small margin or simply lacks sufficient empirical data to make the close call. Nevertheless, self-determination theory will continue to weigh strongly and conclusively in favor of the exemption because of the large normative premium it places on critical uses.[105]

Consider now a related question: is there an additional normative value to enabling broader opportunity for all members of society to take part in critical uses of dominant cultural materials, distinct from the degree of critical output generated? Suppose for the sake of argument that, under certain conditions, facilitating creative opportunities only for a relatively narrow professional elite suffices for generating robust critical expression, and that extending such opportunities to the entire population, including poorly-financed amateurs, is unlikely to have additional significant effect on the variety and depth of critical output.[106] This assumption may have important doctrinal implications. Arguably, allowing critical opportunities for just a professional elite may require only a mild restriction of copyright, rather than a complete exemption. For this purpose, it may suffice to simply limit the remedy for infringement to actual damages, and thus remove the copyright holder's veto power by denying injunctions,[107] disgorgement of the user's

104. *See* Bell, *supra* note 100.

105. Rob Merges has recently suggested that autonomy considerations may play a tiebreaker role where efficiency is "blurry" or indeterminate, so as to put a "thumb on the scale" in tight cases. *See* Robert P. Merges, *Autonomy and Independence: The Normative Face of Transaction Costs*, 53 ARIZ. L. REV. 145, 162–63. By contrast, our argument for the higher-order status of certain interests in self-determination results in a considerably stronger departure from efficiency: not only may it provide firmer support for a rule when efficiency is uncertain, but it can continue that support even when efficiency points the other way, by virtue of the lexical priority it gives in certain circumstances to these higher-order interests, over and above efficiency's satisfaction of preferences.

106. To be sure, under actual economic and social realities of producing expressive works, broadening of expressive opportunities will tend to be structurally connected to the character of the expression produced. *See infra* Section III.B.2 (discussing "heterodox works"). We nevertheless make the stylized assumption that robust critical expression may be achieved absent wider opportunities, in order to pinpoint an important normative difference between theories of self-determination versus cultural democracy.

107. Under the Copyright Act, courts have a broad discretion with regard to injunctive relief. *See* 17 U.S.C. § 502(a) (2012). Following the Supreme Court patent decision in *eBay v. MercExchange*, courts have reversed a previous trend and started applying a more exacting and discretionary approach to the grant of injunctions in copyright cases in line with general injunctive relief principles. *See* eBay, Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006) (holding that courts may only grant injunctions in patent cases as discretionary equitable relief subject to the general four-factor test); Flava Works, Inc. v. Gunter, 689 F.3d 754, 755 (7th Cir. 2012) (applying *eBay* to copyright cases); Flexible Lifeline Sys. v. Precision Lift, Inc., 654 F.3d 989, 997–98 (9th Cir. 2011) (same); Salinger v. Colting, 607 F.3d 68, 77 (2d Cir. 2010) (same).

profits,[108] or the threat of excessive statutory damages.[109] Effectively opening up such opportunities for all individuals, however, would require considerably more. Individuals not backed by a commercial enterprise, or who do not expect substantial income from the use, would need to be free of both the prospect of substantial damages and the costly uncertainty created by fuzzy, hard-to-predict standards.[110] In other words, broadening opportunity for critical expression to all would require something approximating a complete exemption founded on a clear, predictable rule.

Self-determination as such provides little support for expanding opportunity for critical uses, independent from an increase in critical output. What is important on this view is the creation of an expressive environment conducive to reflective self-authorship, rather than the process of who contributes to it. Political democracy, by contrast, does require equal opportunity for critical uses, albeit only within a certain subdomain of expression. Recall political democracy's two distinguishing features: an extension of self-determination's ambit from individual to collective decision-making and, given the latter's inherently relational character, a concern with equalizing opportunities for participation in such decisions. From these two features follows the need to broaden opportunities for all individuals to engage in critical uses but only within the delimited domain of shaping the deliberative process relevant to political decisions. A clear case would be extensive citation from a political speech for purposes of a critical analysis. And, under a broad reading, the principle would also apply to uses such as *The Wind Done Gone* that, while not political tracts, have plausible bearing on the subject matter of politics. At some point, however, the connection between cultural works and the political process becomes too tenuous; it is doubtful, for example, that the political democracy rationale plausibly applies to the *Air Pirates*. It is at this point that the specific contribution of cultural democracy theory kicks in. Here too, the guiding principles are twofold: to extend the ambit of self-determination to the

108. Under existing law, a copyright owner is entitled as matter of right to recover both actual damages and any additional profits by the infringer. *See* 17 U.S.C. § 504(a)(1)–(2), (b).

109. *See* 17 U.S.C. § 504(c).

110. For similar reasons, Molly Van Houweling has suggested there be a presumption of fair use for noncommercial derivative uses. Molly Shaffer Van Houweling, *Distributive Values in Copyright*, 83 TEX. L. REV. 1535, 1547–49, 1565–68 (2005). Admittedly, Van Houweling offers her proposal as an administrable proxy for her actual underlying concern, which is to enable greater expressive opportunities for lower-income creators. However, although Van Houweling couches that latter concern in the language of "distributive values," the substance of her argument strikes us as straddling what we analyze in this Part under the rubric of "political democracy" and what we analyze in Part IV, *infra*, as "strictly distributive concerns."

cultural sphere (where a person's subjectivity is shaped) and, recognizing the irreducibly social character of meaning-making, to extend democratic concerns with equalizing collective power to the realm of shaping the subjectivity *of others*. Cultural democracy's commitment, then, to decentralizing meaning-making requires enabling all individuals to participate in shaping culture. Applied to critical uses this requires not just conditions conducive to ample critical materials, but also a broad opportunity for all individuals to engage in such uses. The doctrinal upshot: a strong, predictable rule that exempts critical uses from copyright liability.[111]

2. Derivative Works

Another fiercely debated topic in copyright policy is the entitlement to make derivative works.[112] Granting this exclusive entitlement to the copyright holder, alongside an expansive understanding of the reproduction requirement,[113] has the effect of bringing within the holder's control a broad

111. Interestingly, two developments in fair use case law in recent decades have brought copyright law closer to such a strong exemption for critical uses. The first is the heavy weight given to the transformative nature of the use under the fair use analysis. *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) ("The more transformative the new work, the less will be the significance of other factors . . . ."). Not all transformative uses are critical, but most strongly critical uses are likely to be considered highly transformative when the concept is properly applied, and therefore are likely to be found non-infringing. Second, a rising, although not uniformly accepted, rule in the case law is that any negative effect of a highly transformative use on a possible licensing market for such use is to be ignored when inquiring into market harm under the fourth fair-use factor. *See* Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 614–15 (2d Cir. 2006) (refusing to consider the impact of defendant's transformative use on the licensing market under the fair use analysis). The practical effect of this rule is to buttress the certainty and predictability of the fair use safe haven for transformative and critical uses, by replacing the highly malleable and uncertain inquiry concerning market harm with a clear, irrebuttable presumption of no market harm whenever the use is strongly transformative. The net result is that the two most important of the four fair-use factors cut strongly and clearly in favor of non-infringement in regards to highly transformative and critical uses. Incomplete acceptance of the second rule mentioned, different standards applied in determining what is a highly transformative use, and the general malleability of the fair use analysis all cause the current situation to fall short of the clear and predictable full exemption to critical uses described in the text. Nevertheless the combined effect of the two rules brings the current fair use doctrine significantly closer to such an exemption.

112. *See* 17 U.S.C. § 106(2).

113. 17 U.S.C. § 106(1). The borderline between the derivative-works entitlement and other entitlements, especially a broadly defined reproduction right, is hard to pin down. This has led some commentators to conclude that the derivative-works right is superfluous because wherever it is implicated other entitlements would also be infringed. *See* Twin Peaks Prods. v. Publ'ns Int'l, Ltd., 996 F.2d 1366, 1373 (2d Cir. 1993) ("[The] right to make derivative works is 'completely superfluous' . . . ." (quoting MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 8.09[A])); MELVILLE B. NIMMER & DAVID NIMMER,

array of secondary uses that go well beyond mere copying. Historically, copyright's scope was limited to verbatim reproduction with a narrow zone of minor variations.[114] Starting in the late nineteenth century, this scope was expanded through the addition of new entitlements and changes in the concept of, and legal tests for, infringement.[115] By the time of the 1976 Copyright Act,[116] the pendulum had swung to the opposite end. A copyright owner now holds exclusive control over any work based on hers, no matter how much the original is "recast, transformed or adapted."[117] This capacious control is almost unlimited by the terms of the derivative-works right itself[118]

NIMMER ON COPYRIGHT § 8.09[A] (2010) (explaining that any infringement of the derivative-works entitlement also infringes either the reproduction or public-performance entitlements). *But see* Daniel Gervais, *The Derivative Right: Or Why Copyright Law Protects Foxes Better than Hedgehogs*, 15 VAND. J. ENT. & TECH. L. 785, 800 (2013) ("[T]here are forms of derivation that copy and some that do not.").

114. *See* Oren Bracha, *The Ideology of Authorship Revisited: Authors, Markets, and Liberal Values in Early American Copyright*, 118 YALE L.J. 186, 224–26 (2008); Paul Goldstein, *Derivative Rights and Derivative Works in Copyright*, 30 J. COPYRIGHT SOC'Y U.S.A. 209, 211–14 (1983); Naomi Abe Voegtli, *Rethinking Derivative Rights*, 63 BROOK. L. REV. 1213, 1233–34 (1997).

115. Bracha, *supra* note 114, at 226–33; Goldstein, *supra* note 114, at 214–15; Voegtli, *supra* note 114, at 1234–39.

116. Pub. L. No. 94-553, 90 Stat. 2541 (codified in various sections of 17 U.S.C.)

117. *See* 17 U.S.C. § 101 (defining a "derivative work"). Pam Samuelson has recently argued that a proper understanding of the legislative history, constitutional purpose and textual structure of the 1976 Act shows that the derivative-works right is not nearly so broad as many have thought and, in particular, that the open-ended clause at the end of the statutory definition, purporting to cover "any other form in which [a copyrighted] work is recast, transformed, or adapted" should be narrowly construed so as to limit it to cover only close analogues of the preceding nine "exemplary derivatives" listed in the statutory definition (namely "a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgement, [and] condensation"). *See* Pamela Samuelson, *The Quest for a Sound Conception of Copyright's Derivative Work Right*, 101 GEO. L.J. 1505, 1509, 1511 (2013) (quoting 17 U.S.C. § 101). However, Professor Samuelson also points out that a number of courts have nevertheless "interpreted broadly" this ending clause, leading to "highly problematic" decisions that have increased uncertainty about the scope of the clause. *Id.* at 1509–10. This resulted in substantively "unsound" outcomes. *Id.* at 1547. Moreover, even if Professor Samuelson's narrower construal of the right as currently enacted were to prevail, our normative analysis may have more far-reaching implications, prescribing a rolling back of the existing categories as well.

118. Some courts have tried to delimit the entitlement by pointing to internal constraints imposed by its definition. In some cases the result has been to adduce very minimal internal constraints, while in others it has resulted in a plainly confusing and unpersuasive understanding of the concept of derivative works. For an example of the former, see Micro Star v. FormGen Inc., 154 F.3d 1107, 1110 (9th Cir. 1998) ("[A] derivative work must exist in a 'concrete or permanent form' . . . and must substantially incorporate protected material from the preexisting work . . . ." (quoting Lewis Galoob Toys, Inc. v. Nintendo of Am., Inc., 964 F.2d 965, 967 (9th Cir. 1992)). For an example of the latter, see Warner Bros. Entm't Inc. v. RDR Books, 575 F. Supp. 2d 513 (S.D.N.Y. 2008) (holding that a lexicon based on

and is only moderately limited by external checks, chief among which is the fair use doctrine.[119] Anything from a rap version of a pop song,[120] to a costume based on a cartoon character,[121] may be a derivative work that falls within the entitlement's scope.

Is such broad control by the copyright owner desirable? Under a standard efficiency analysis, the answer is probably no. The basic justification for the entitlement is that by including secondary markets, it allows the creator to internalize a larger share of the social value of her creation.[122] In many cases, however, the additional profit generated from licensing in secondary markets (or from development of those markets by the copyright owner) will be either unnecessary or inconsequential as an incentive. It will be unnecessary where exploitation of the work in its primary market is clearly sufficient to recoup development expenditures.[123] It will be inconsequential when the expected value of the secondary markets at the time of creating the original is so small relative to development costs that it is unlikely to make the difference between overall net loss and profit.[124] As a rule secondary markets tend to be significantly profitable when the original is very successful (think *Star Wars*)—in other words, in exactly the cases where the additional profit they yield is likely to be unnecessary.[125] Of course, these are rough generalizations. Sometimes the development of a work may be so costly that the pooling of profits from all markets, even in cases where secondary markets are small relative to the primary one, may be necessary to recoup

the Harry Potter novels was both too transformative to be considered derivative of the original and yet that it was nevertheless a reproduction of it). For another attempt to infer the scope of the derivative works entitlement from the concept, see Gervais, *supra* note 113, at 846 ("Derivation is best seen as taking what makes a preexisting work (or more than one) original for the purpose of transforming it, but not to the point of a fundamental alteration of the message." (emphasis omitted)).

119. 17 U.S.C. § 107. For a discussion on the limited restrictions imposed by general copyright principles on the derivative works entitlement, see Voegtli, *supra* note 114, at 1217–32.

120. *See* Campbell v. Acuff-Rose Music Inc., 510 U.S. 569 (1994).

121. *See* Entm't Research Grp., Inc. v. Genesis Creative Grp., Inc., 122 F.3d 1211 (9th Cir. 1997).

122. Goldstein, *supra* note 114, at 216.

123. *See* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 109–10 (2003); Landes & Posner, *supra* note 24, at 354; Stewart E. Sterk, *Rhetoric and Reality in Copyright Law*, 94 MICH. L. REV. 1197, 1216 (1996); Voegtli *supra* note 114, at 1241–42.

124. *See* Sterk, *supra* note 123, at 1216.

125. *See id.* (contrasting the first-time author who is unlikely to expect significant profits from secondary markets with the established author who is more likely to earn such profits but do not need them to recoup her investment); *see also* Abramowicz, *supra* note 17, at 68–77.

costs. At other times, a work may generate much more profit in secondary markets than in the primary one.[126] On balance, however, as long as it is plausible to expect the aforementioned generalizations to hold, the incentive value of the derivative work entitlement will be weak. On the other side of the copyright tradeoff, the entitlement is likely to create substantial access costs as a result of prohibitive transaction costs either barring efficient uses or driving up their price. The net result: a relatively low incentive benefit likely outweighed by access costs.

Factoring in product differentiation modifies the analysis somewhat, but the overall conclusion likely remains intact. The main element added by product differentiation is that the rents generated in inframarginal markets by the derivate-works entitlement are likely to attract new entrants with closely differentiated products. The picture is not limited to the one developer of the adventure movie of the season, enjoying large rents through control of both primary and secondary markets, along with correspondingly large deadweight losses. Rather, we are likely to see several adventure movies whose developers share the surplus in the primary and secondary markets, with correspondingly significant duplicative wastes and some degree of price competition and, somewhat less likely, perhaps some reduction in the wasteful and prohibitive effects associated with transacting over secondary uses. The net result is a similarly indeterminate conclusion as traditional economic analysis, albeit on somewhat different grounds.

On balance, then, both traditional and product-differentiation-based economic analyses tilt against a broad derivative-work right.[127] They favor either the abolition of the entitlement or its replacement with narrower rights over secondary markets, which target those contexts where efficiency may tilt in favor of such control, such as exclusive rights for translations or motion-picture adaptations.

Does the analysis change if we switch our normative lens to, say, self-determination? The starting point for the analysis is again the notion of a normative premium this view places on certain kinds of expressions. Specifically, a special value is imputed to what may be called "heterodox" works. Heterodox works challenge broadly held views, beliefs or tastes, and

126. *See* Derek E. Bambauer, *Faulty Math: The Economics of Legalizing* The Grey Album, 59 ALA. L. REV. 345, 381–82 (2008) (discussing the example of comic books but concentrating mainly on whether secondary markets' revenue is substantial rather than whether it is necessary); Sterk, *supra* note 123, at 1216 (giving the example of an "extraordinarily high-budget movie with the potential for sales of toys, t-shirts, and the like").

127. For additional considerations under efficiency analysis of the derivative work entitlement, see Landes & Posner, *supra* note 24, at 354–55; Voegtli, *supra* note 114, at 1241–47.

include content that is outside the mainstream, is avant-garde, or is aimed at preferences and tastes that are relatively marginal.[128] Self-determination theory places a normative premium on such content because it is essential for a distinct precondition of genuine free choice than opportunity for critical reflection and revision, namely, meaningful variety. As Joseph Raz explains, free choice requires a sufficient variety of options from which to choose, options that, unlike a group of identical cherries in a bowl, must be materially different from each other.[129] The choice between a hundred mainstream channels, each providing similar formulaic programing, does not constitute such meaningful variety.[130] The lack of challenge to entrenched and comfortable options also breeds attitudes of passivity and conformity—the antithesis of self-determination. In the words of Mill, "[p]recisely because the tyranny of opinion is such as to make eccentricity a reproach, it is desirable, in order to break through the tyranny, that people should be eccentric."[131] Heterodox works supply eccentricity and thereby a crucial part of the range of possibilities and alternatives that make individual choice meaningful. Such works provide real alternatives to dominant options and a foil against which to critically compare and assess those options. Heterodox works also create the potential for tomorrow's prevailing choices to be different from today's, and consciously or reflectively so rather than being determined by sheer inertia or drift. Yet by their very nature as relatively marginal and/or avant-garde, heterodox works run the risk of being in short supply if left to the forces of the market that reflect existing preferences. The conditions that allow heterodox expression to flourish are thus part of the preconditions for validating individual choices and their valuation cannot therefore be reduced to such choices.

Even with a normative premium placed upon heterodox works, however, the self-determination analysis of derivative works may seem to track that of efficiency. To the extent that heterodox expression is equally affected by the incentive and access sides of the entitlement's effects, the efficiency and autonomy tradeoffs will tend to converge. Upon closer reflection, however, possible distinctions emerge.

---

128. *See, e.g.*, NETANEL, *supra* note 65, at 40 (explaining the similar concept of "oppositional expression"); STEVEN H. SHIFFRIN, THE FIRST AMENDMENT, DEMOCRACY, AND ROMANCE 86–109 (1990) (discussing the value of dissent in free speech jurisprudence).

129. *See* RAZ, *supra* note 59, at 398.

130. On the difference between meaningful diversity of content and mere product differentiation, see NETANEL, *supra* note 65, at 39–40.

131. MILL, *supra* note 101, at 62.

Consider first the incentive effect of the derivative works entitlement on heterodox works. Control of secondary markets will tend to be much less consequential to the recoupment of investment in heterodox works. Licensing fees from the T-shirt or action-figure markets, or even a hope for a lucrative mainstream movie adaptation, are likely to represent a small portion of their expected value at the time of creating heterodox works. In many cases, non-trivial income from such secondary markets will only materialize exactly when the work ceases to be heterodox and shifts into the mainstream—an elusive and unpredictable process. As a result, relatively few heterodox works will be pushed from the supramarginal to creation as a result of the derivative works entitlement.

On the other side of the tradeoff, the access cost of the entitlement will take a disproportionately large toll on heterodox works. An important common strategy for creating heterodox works is what Neil Netanel calls "creative appropriation," meaning a use of dominant mainstream materials that is at once "highly derivative" and "powerfully subversive."[132] An example is Tom Forsythe's Food Chain Barbie—a series of photographs of Barbie dolls attacked, tortured, and mutilated by kitchen appliances.[133] Creative appropriation is not new. The digital age holds, however, the promise of a radical democratization and spread of this form of creation. A combination of pervasive, relatively affordable technology for creating and disseminating high quality digital content, alongside emerging social norms and practices, has produced a growing wave of bottom-up decentralized creativity. A very common element within this phenomenon is "remixing"[134] or "glomming on"[135]—the use of dominant mass media content while imbuing it with new, subversive meaning(s) or simply using content as the building blocks for new creativity.

The promise of this decentralized wave of user innovation in producing a vibrant and diverse expressive realm, with a steady supply of heterodox works in contrast to the tendency of mainstream mass media toward the

132. NETANEL, *supra* note 65, at 159; *see also* Voegtli, *supra* note 114, at 1221–26 (discussing "appropriative art").

133. The series is described by Tom Forsythe as a "pictorial antidote to the powerful cultural forces persuading us to buy the impossible beauty myth." *See Food Chain Barbie, R Prints, Edition of 450*, ARTSURDIST, http://www.tomforsythe.com/food-chain-barbie---r-prints.html (last visited Jan. 10, 2014). Forsythe endured a five-year legal battle after being sued by Mattel for copyright and trademark infringement, at the end of which he emerged victorious. *See* Mattel Inc. v. Walking Mountain Prods., 353 F.3d 792 (9th Cir. 2003).

134. *See generally* LAWRENCE LESSIG, REMIX: MAKING ART AND COMMERCE THRIVE IN THE HYBRID ECONOMY (2008).

135. *See* Balkin, *supra* note 76, at 10–12.

uniform and uncontroversial, is great.[136] Many commentators see it as the most important source of diversity, innovation, and challenge in an increasingly centralized cultural sphere.[137] But the vibrancy of this expressive source depends to a large extent on the ready availability of its raw materials, namely preexisting informational works, particularly culturally dominant materials of exactly the kind where the owner is likely to have an interest in aggressive enforcement of derivative works rights.[138] The most promising source of heterodox works, in other words, is disproportionately dependent upon access to copyrighted materials. Moreover, the creators of such works, who often serve only small niche demand or pursue non-profit-oriented strategies, are more frequently unable to recoup large development costs. As a result, these works are likely to be particularly adversely affected by the access barriers imposed by a derivative-work entitlement, in the form of either flat licensing fees or prohibitive transaction cost.

The net result is that heterodox works, and especialsly the most promising source for such works in the digital age, are disproportionately prejudiced by the access cost of derivative works protection, while benefiting relatively little from its incentive function. Compared, then, to standard efficiency analyses of the derivative-work rights, a self-determination perspective is more conclusively skeptical.

Recently, however, a new economic rationale for the entitlement has been advanced. On this view, the primary purpose of the derivative entitlement is not the dubious one of increasing incentives for creation, but rather of serving as a check on wasteful rent dissipation in secondary markets.[139] The derivative work entitlement centralizes the power to authorize and coordinate secondary innovation in the hands of the copyright

136. On the tendency of mass media to produce mainstream, bland content, see C. EDWIN BAKER, MEDIA, MARKETS, AND DEMOCRACY 115 (2002); BENKLER, THE WEALTH OF NETWORKS, *supra* note 58, at 196–211 (discussing basic critiques of mass media including the tendency toward "lowest common denominator programming"); NETANEL, *supra* note 65, at 39–40 (discussing conflicting arguments in the literature and concluding that studies praising the diversity of content generated by mass media are based on a narrow understanding of diversity detached from the expressive diversity relevant for free speech theory).

137. *See* BENKLER, THE WEALTH OF NETWORKS, *supra* note 58, at 276; NETANEL, *supra* note 65, at 44; SUNDER, *supra* note 81, at 105–14; Coombe, *supra* note 76, at 1864; Fisher, *supra* note 77, at 1460.

138. NETANEL, *supra* note 65, at 41 (arguing that oppositional speakers "are far less able to acquire copyright permissions").

139. *See* Abramowicz, *A Theory of Copyright's Derivative Right and Related Doctrines*, *supra* note 33, at 358–59. Abramowicz is building here on Edmund Kitch's earlier work in the patent context. *See* Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265, 276 (1977).

owner of the primary work, thereby avoiding the wasteful overlapping activity associated with races to capture the rents from secondary uses of popular primary works.[140] This is the reason why we get just one computer game based on a successful movie rather than ten highly duplicative versions that mainly divert demand from each other. Like the traditional economic justification, this argument is far from establishing conclusive support for the entitlement. For one, preventing secondary-level rent dissipation by creating a highly valuable right to control secondary markets may simply push the rent dissipation back to the primary level, by igniting races to capture this right that are potential equally or even more wasteful.[141] It is far from clear that copyright doctrine has effective tools for applying the entitlement only when its benefits clearly outweigh this cost.[142] Moreover, centralized control is likely to have detrimental effect on the quality of innovation on the secondary level: in the highly uncertain and unpredictable area of creative innovation, the most effective way of finding out what works best is to let many minds try.[143] Thus, like the traditional efficiency analysis, this variant is somewhat inconclusive.

Once again, self-determination provides a more conclusively negative evaluation of the derivative-work right, now understood as a tool for centralized coordination. The analysis hinges once again on the right's disproportionate effect on heterodox works. For the reasons just explained, heterodox works are likely to be disproportionately represented on the side of derivative works (whose production is restricted by the centralized control of the copyright owner) and, conversely, more sparsely represented on the side of those primary works with a high probability of lucrative secondary development. Most importantly, heterodox secondary works are particularly likely to be susceptive to the pernicious tendencies of centralized control over secondary innovation. Compared to mainstream uses, heterodox works often promise a lower overall profit, higher risk, or both. The private profit-maximizing interest of coordinating copyright holders, and the known

140. Abramowicz, *A Theory of Copyright's Derivative Right and Related Doctrines*, *supra* note 33, at 358 (explaining that the concern "is that derivative works will be redundant with one another").

141. *See* Bracha & Syed, *supra* note 34, at 34. For the original critique in the patent context, see Donald G. McFetridge & Douglas A. Smith, P*atents, Prospects, and Economic Surplus: A Comment*, 23 J.L. & ECON. 197 (1980).

142. Bracha & Syed, *supra* note 34, at 34.

143. *See* Lemley, *supra* note 24, at 1068–69; Robert P. Merges & Richard R. Nelson, *On the Complex Economics of Patent Scope*, 90 COLUM. L. REV. 839, 908 (1990) (making the argument in the patent context).

tendency of mainstream media firms to minimize risk,[144] will systemically prefer the creation of conventional secondary works over more heterodox ones. If there is to be only one computer game following a successful movie, we can safely predict that, in most cases, it will be the more conventional, mainstream version that is licensed rather than a subversive or experimental variant. Similarly, if *Pride and Prejudice* were under copyright, it would be one of the more standard motion-picture adaptations that is licensed rather than *Bride and Prejudice*.[145] Were Shakespeare's plays sheltered behind a derivative works entitlement, it is doubtful that *Lear's Daughters*, a loose adaptation, would be licensed.[146] Nor is it likely that Aime Cesaire's postcolonial play *A Tempest* would ever be authorized.[147] The private coordination power created by the entitlement is systemically biased against heterodox works. In sum, the special value given to heterodox works by self-determination theory leads again to a more clear and secure negative assessment of the derivative works, when compared to economic analysis.

3. Fan Fiction

Fan fiction has fast become the source of heated debate in copyright policy.[148] The genre is a form of secondary creation that builds on well-established—usually highly popular—cultural works, and is typically produced and consumed by fans of such works. The writings are typically located within the "universe" of the primary work, and employ and build upon much of the creative material of the original: its characters, relationships, themes, scenery and plot lines. The extent to which efforts in the genre add new creative content varies considerably, from new story lines

144. *See* NETANEL, *supra* note 65, at 136–37 (discussing strategies for reducing risk by mainstream media companies).

145. Bride and Prejudice, directed by Gurinder Chadha, is a 2004 Bollywood adaptation of the novel Pride and Prejudice. *See* BRIDE AND PREJUDICE (Pathé Pictures Int'l 2004). To be sure, the question of which derivative is subversive or experimental is debatable in general and in regard to Austen's adaptations in particular. *See* Julian North, *Conservative Austen, Radical Austen: Sense and Sensibility from Text to Screen*, *in* ADAPTATIONS: FROM TEXT TO SCREEN, SCREEN TO TEXT 38 (Deborah Cartmell & Imelda Whelehan eds., 2013).

146. The 1987 play has been described as "a 'landmark' in the feminist reinvention of Shakespeare." *See* LYNNE BRADLEY, ADAPTING KING LEAR FOR THE STAGE 218 (2010).

147. For background on the play, see GREGSON DAVIS, AIME CESAIRE 156–62 (1997).

148. *See, e.g.*, FAN FICTION AND FAN COMMUNITIES IN THE AGE OF THE INTERNET: NEW ESSAYS (Karen Hellekson & Kristina Busse eds., 2006) [hereinafter FAN FICTION AND FAN COMMUNITIES]; Anupam Chander & Madhavi Sunder, *Everyone's a Superhero: A Cultural Theory of "Mary Sue" Fan Fiction as Fair Use*, 95 CALIF. L. REV. 597 (2007); Aaron Schwabach, *The Harry Potter Lexicon and the World of Fandom: Fan Fiction, Outsider Works, and Copyright*, 70 U. PITT. L. REV. 387 (2009); Rebecca Tushnet, *Legal Fictions: Copyright, Fan Fiction, and a New Common Law*, 17 LOY. L.A. ENT. L. REV. 651 (1997).

or characters folded within well-established patterns, to novel and sometimes radical or subversive transformations of meaning through changes in the creative elements of the work.[149] While fan fiction provokes a broad spectrum of reactions from copyright holders,[150] the central question in policy debates is whether, and to what extent, a copyright owner should have legal control over such amateur uses of her work. Should the copyright owner of *Star Trek*, for example, have the exclusive right to authorize or forbid fan-generated writings that explore new exploits of Captain Kirk and Mr. Spock, perhaps in ways that shed surprising light on the characters and the story?

The standard efficiency analysis of this question is somewhat inconclusive. Effects of fan fiction on supramarginal creation or inframarginal entry do not seem substantial. Other than in exceptional cases,[151] fan fiction does not form even a remote substitute for the original work or for any commercial follow-up projects in which the copyright holder may want to engage or for which she wishes to license her work. What about licensing fan fiction itself? The highly decentralized context and the difficulty in assessing a given user's usage and valuation may impose high transaction costs and impede licensing.[152] Thus it is unclear whether placing fan fiction beyond the reach of copyright is likely to substantially reduce revenue or degrade incentive. Some other factors, however, point in the opposite direction. The ordinarily low commercial value of fan fiction means that under the efficiency criterion of willingness and ability to pay, the cost of curtailed access from copyright restrictions may not be significant, and thus outweighed by whatever small incentive benefits are created. Moreover, copyright owners may derive benefits other than licensing revenue from a legal power to control and exclude certain kinds of fan fiction. For example, a copyright owner who plans future additions or extensions to the work may

149. *See* Chander & Sunder, *supra* note 148, at 598–601; Sonia K. Katyal, *Performance, Property, and the Slashing of Gender in Fan Fiction*, 14 J. GENDER SOC. POL'Y & L. 461, 481–97 (2006).

150. *See* Fisher, *supra* note 77, at 1435–41 (describing the range of intellectual property owners' responses to user innovation including in the form of fan fiction).

151. Such exceptions are particularly likely in regard to future projects that are amenable to development by peer production, especially when the fan-project is susceptible to be converted or leveraged into a centralized commercial enterprise. The Harry Potter Lexicon may have been one of those cases. *See* Warner Bros. Entm't Inc. v. RDR Books, 575 F. Supp. 2d 513 (S.D.N.Y. 2008).

152. On the other hand, blanket licenses of a portfolio of works administered by an intermediary, either for a flat fee or on the basis of royalties from exploitation of the secondary works, may make licensing both feasible and worthwhile. *See infra* text accompanying notes 157–61 (describing Amazon's Kindle Worlds platform).

see herself as materially prejudiced by fan fiction that anticipates or preempts some of those plans. Alternatively, some authors may see themselves adversely affected by what is usually described as encroachment on their moral rights, meaning uses of their creative work that—from the point of view of the author—mutilate, distort, or subvert its content. To the extent that any such possibly prejudicial uses are anticipated at the time of the original work's creation and are likely to influence the incentive for its creation, they are relevant from the normative perspective of efficiency.[153]

Product differentiation adds a possible further dimension: unrestricted production of fan fiction is likely to create high amounts of wasteful duplicative activity. The argument here is not that fan fiction is likely to be a good substitute for the original or for any official follow up. Rather, it is that much of the fan fiction produced constitutes a close substitute for other fan fiction, from the point of view of consumers. To be sure, readers of the genre have their preferences among writers or stories, and different variants are better tailored to some subset of preferences. Nevertheless, in a corpus of thousands of stories in the *Star Wars* universe there is likely to be a high degree of overlapping demand satisfaction, and thus the real efficiency problem with fan fiction may not be a jeopardizing of incentives as much as the waste of so much effort and cost expended in creating a multitude of works that are mostly close substitutes of one another.

Would applying any of the democratic rivals of efficiency change the analysis? One natural candidate is flourishing theory. The fan fiction phenomenon maps closely onto some of the central elements of human flourishing distilled above. Most importantly, the creation of fan fiction is a prime example of meaningful human activity. Those who take part in it trade in a position as passive consumer of meanings created by others—over which they have little control—for a new role of active participation and engagement with dominant expressive materials. Such an active role requires agency, the stretching of one's mental muscles (of imagination, expression, etc.), and usually a high degree of identification.[154] Fan fiction production is also typically social in nature, often providing a fulcrum around which

153. *See* Jeanne C. Fromer, *Expressive Incentives in Intellectual Property*, 98 VA. L. REV. 1745, 1817 (2012) (observing that "a robust right of integrity to authors might serve as a strong expressive incentive" but finding it to be inadvisable due to the cost associated with it).

154. Chander & Sunder, *supra* note 148, at 609 ("Mary Sues help the writer claim agency against a popular culture that repeatedly denies it."); Katyal, *supra* note 149, at 488 ("[S]lash . . . highlights the increasingly participatory culture of cyberspace and the audience's inherent challenge to the author's control over the creation of meaning and subtext.").

communities of shared interests and enterprises are built and developed.[155] This perspective thus sheds new light on fan fiction production. Instead of a wasteful, duplicative activity that produces little output value, it begins to seem as an activity of intrinsic, and potentially invaluable, worth to those who engage in it, whether as producers, members of communities, or both.

In one sense this shift in perspective is internal to efficiency analysis, rather than an indication of a difference between efficiency and flourishing. Put differently, the brief excursion into flourishing theory was simply a detour necessary for realigning the efficiency analysis. Standard efficiency analysis tends to see fan fiction in purely instrumental terms, as an activity aimed at an output, which activity is of little social value because (so it is assumed) the consumptive market value of its output is low. However, the picture changes once we realize that from the point of view of those engaged in it, the production of fan fiction is intrinsically valuable—not merely a cost, but rather a rewarding in itself. The aggregate value to producers of the activity, a previously ignored factor, becomes an important (possibly the most important) part of the social value of fan fiction. As a result, the assumption of wasteful duplication that downgrades the social value of fan fiction loses much of its force. The concern over duplicative production cost is moot when, in many cases, the productive activity is, in itself, a net benefit rather than a cost.

This still leaves open the question of divergence between the two theories. Now that we have established that efficiency needs to properly account for the value of fan fiction to its producers, does there remain any difference between its evaluation of the relevant effects and that of flourishing theory? Simply saying that the latter places a higher value on fan fiction is not enough. One needs to explain why we should place a higher value than those who engage in it, as either consumers or producers, as revealed through their preferences. The question may have considerable practical significance in view of the prospects for private mechanisms to reduce transaction costs in this domain. As explained, commercial licensing of copyrighted works is increasingly feasible, either through technology that allows finely-tailored licensing schemes or through intermediaries administering subscription and blanket-licensing schemes.[156] With Amazon's

155. *See* FAN FICTION AND FAN COMMUNITIES, *supra* note 148, at 7 ("[T]he community of fans creates a communal (albeit contentious and contradictory) interpretation."); HENRY JENKINS, TEXTUAL POACHERS: TELEVISION FANS AND PARTICIPATORY CULTURE 45 (1992).

156. On private institutions for overcoming transaction cost problems in the context of intellectual property, see MERGES, *supra* note 44, at 230–34; Robert Merges, *Contracting into*

recent launch of its fan-fiction publishing platform Kindle Worlds, the prospect of licensing in this field has begun to materialize.[157] Kindle Worlds is a scheme for the commercial exploitation of fan fiction based on: blanket licenses issued to fan writers by rights-holders of the original works, a massive transfer of exclusive rights in the resultant fan fiction works to Amazon and a royalty-based arrangement for sharing the profits generated from consumers of these works between the three parties.[158] The likely proliferation of similar such schemes will enable copyright owners to capture the demand for fan fiction at relatively low transaction costs, while still allowing much of this demand (by both producers and consumers) to be satisfied. Under such circumstances, efficiency analysis may shift toward firmer support of copyright protection.[159]

Why might flourishing theory still insist on exempting fan fiction, and thus take the view that the opportunities to engage in the activity may be inadequate even when deemed sufficient by the lights of those who produce and consume it?[160] This difficulty facing flourishing theory—of justifying non-deference to subjective preferences—is not particular to the fan fiction context, but extends more generally, across a variety of copyright policy questions. The pervasiveness of the challenge is attributable to this theory's departure from one of liberalism's central tenets, namely neutrality in social choice toward individuals' subjective conceptions of the good.[161] And there exist a range of possible answers to that challenge, which can be arrayed along a continuum from greatest to least degree of departure from the liberal principle of neutrality.

One answer is forthright paternalism, namely the assumption that many people simply do not know what is good for them in this context.[162] To be sure, this is controversial. If, however, one is willing to affirm that commitments to human flourishing require placing a greater value on access

Liability Rules: Intellectual Property Rights and Collective Rights Organizations, 84 CALIF. L. REV. 1293 (1996).

157. *See* KINDLE WORLDS, https://kindleworlds.amazon.com (last visited Nov. 8, 2013).

158. *Kindle Worlds Publishing Agreement*, KINDLE WORLDS (June 27, 2013), https://kindleworlds.amazon.com/faqs?topicId=A1MMH2I71OJWTR.

159. Copyright protection accompanied by a finely tailored, low transaction cost licensing scheme allows most of those who sufficiently value the production of fan fiction to engage in it and allows producers of the originals to internalize much of this value thereby increasing incentive.

160. We bracket here the possibility that the producers and consumers of fan fiction are disproportionately those with lower income or who otherwise merit a purely distributive priority. For discussion, see *infra* Part IV.

161. *See* sources cited *supra* note 88.

162. *See* Fisher, *supra* note 77, at 1472; Fisher, *supra* note 24, at 1762–66.

to meaningful activities and social interaction than what many of the beneficiaries of that access may place on it, an obvious gap opens up vis-à-vis efficiency analysis. The policy implication of this affirmation is that copyright law should nourish the conditions for fan fiction, even when a market that adequately reflects current subjective preferences would not. In effect this is a form of a subsidy that, while not coercing anyone to do anything, does "nudge" people toward activities considered to be of higher substantive value. This channeling, achieved through keeping opportunities for the valued activity open and accessible, is bought by sacrificing some satisfaction of existing preferences that would occur under a solely efficiency-based legal framework.

An alternative (or perhaps supplementary) response would be to eschew paternalism while still departing from the neutrality principle. The core of this position is that in some cases, the subjective preferences of some individuals should be privileged over those of others because, according to substantive considerations external to the individuals' subjective evaluations, the privileged preferences are more valuable for realizing the good life. Applied here, this means that to some extent those persons seeking to satisfy their preferences for engaging with expressive works in more active and participatory ways should be empowered even if that comes at some cost to frustrating the more subjectively-valued preferences of others for consumptive interaction. Paternalism's possibly offensive element of telling people what is good for them is absent now. Copyright's aim here is not to nudge anyone into a choice that he does not currently realize is good for him. Rather, the purpose is to prioritize, in a situation of an inevitable clash between preferences, some peoples' ability to realize what they think is good for them over others' ability to do the same, on the non-neutral ground of substantively affirming one set of subjective desires as against another.[163]

Yet another answer to the challenge of illegitimately overriding individual preferences points to the endogenous nature of preferences. Rather than being completely exogenous to the social and legal environment, preferences about fan fiction are partly shaped by it. What people prefer and value depends on what possibilities they have been exposed to in the past and what

163. To put the point in terms of liberal theory, this departure from neutrality is not based on withholding "respect" for certain individuals' subjective conception of the good, by foisting upon them an alternative good in tension with their freedom, as does paternalism. Rather, it is based on not according "*equal* respect" toward different individuals' subjective conceptions, with the offense to liberalism residing not in governmental disregard for individual freedom but rather in governmental endorsement of some ways of life over others.

habits they have had opportunity to develop.[164] Habits of consumptive or passive interaction with creative materials, and lack of exposure to alternatives, skew preferences toward such interaction.[165] Legal background rules, including copyright, that erect or lower barriers for more active forms of interaction take part in this preference-formation process.[166] The upshot is that no legal regime is completely neutral or deferential with regard to individual preferences. The choice, then, is not between a neutral legal framework that completely defers to individual preferences and another that nudges preferences in a desirable direction. All frameworks nudge, and none are neutral. Accordingly, flourishing theory's support for exempting fan fiction does not override of individual preferences in the name of a more valuable good, but rather amounts to choosing, among legal regimes each of which unavoidably shapes preferences, the one whose preference-shaping effect is more consistent with the theory's notion of the good.

A last possible answer to the difficulty of diverging from individual preferences focuses on the character of the benefit derived by those who are engaged with fan fiction. Much of the value of some human activities depends on their non-commodified character.[167] As Yochai Benkler puts it, leaving a fifty-dollar check on the table at the end of a friendly meal at a friend's house is likely to reduce the value of the event for all its participants.[168] The category of activities whose value depends on their non-commodified character is elusive and culturally contingent, but there seems to be a fair preliminary argument that, for many, the value of fan fiction would be substantially reduced if the activities associated with it were conducted through commercial market interactions. The form of commodification may matter a great deal. Simply paying a subscription fee for a blanket license may erode the inherent motivation of some for participation in a fan fiction community, but the erosion may be mild. A more robust commodification, such as Amazon's Kindle Worlds platform, that subjects the consumption of fan fiction to use-based fees so that

164. *See* Cass Sunstein, *Legal Interference with Private Preferences*, 53 U. CHI. L. REV. 1129, 1161–64 (1986) (discussing the role of habits in shaping preferences as a justification for "paternalistic" regulation). Ultimately Sunstein insists that the endogenous nature of preferences can justify non-neutral laws on welfare and autonomy grounds alone, while flourishing theory may also rely on direct appeals to a substantive conception of the good.

165. *See* MILL, *supra* note 101, at 65 ("The despotism of custom is everywhere the standing hindrance to human advancement, being in unceasing antagonism to that disposition to aim at something better than customary . . . .").

166. *See* Sunstein, *supra* note 164, at 1146–50 (discussing how adaptive preferences may be shaped by the legal rules).

167. *Id.*

168. BENKLER, THE WEALTH OF NETWORKS, *supra* note 58, at 96–97.

copyright owners can internalize a greater share of the secondary use of their works, is more likely to significantly erode the intrinsic value of the activity. It is quite possible that, for many, much of the inherent value of producing fan fiction, as part of a non-commercial community bound by shared interests, would evaporate with a shift to sales transactions between producer-sellers and consumers. To the extent that this assumption holds, commodification of fan fiction through copyright protection would defeat rather than serve the purpose of preference satisfaction.

Admittedly, the persuasiveness of each of these four justifications for departing from strict adherence to subjective preference satisfaction will vary according to context and one's normative commitments. In combination, however, they furnish flourishing theory's rejection of the liberal neutrality principle, and that theory's concomitant recommendations for copyright policy, significant support.

4. Personal Uses

Personal uses of copyrighted works by users—a topic that received little attention in the past—is attracting growing interest in recent copyright scholarship.[169] Personal uses are defined in the literature in varying and sometimes loose ways. Our focus here is on uses of copyrighted works that involve at least a modicum of expressive interaction but are not necessarily highly transformative of the original. An additional trait of these uses is that, while not necessarily free of any income-generating aspect (a personal blog may sell advertisement space), they do not primarily revolve around creating and selling an expressive commercial product designed to extract a market price from consumers. Within this universe there is much variation. Some personal uses are simply expressive acts, which often facilitate interaction with others: a home video of one's toddler dancing to the sounds of a copyrighted song posted on Facebook, an amateur "cover" video clip of the song posted on YouTube, or a wedding video playing the song in the background posted on a personal blog. Other uses are geared toward facilitating dialogic interaction with others: quotations from a literary work or video snippets out of a film in a blog post discussing those cultural artifacts. Still others involve using works in order to assert one's identity or share with others significant aspects of one's persona: a blog post about a personal

169. *See* Julie Cohen, *The Place of the User in Copyright Law*, 74 FORDHAM L. REV. 347 (2005); Jessica Litman, *Lawful Personal Use*, 85 TEX. L. REV. 1871 (2007); Joseph P. Liu, *Copyright Law's Theory of the Consumer*, 44 B.C. L. REV. 397 (2003); Rothman, *supra* note 53; Rebecca Tushnet, *Copy This Essay: How Fair Use Doctrine Harms Free Speech and How Copying Serves It*, 114 YALE L.J. 535, 562–81 (2004).

aspect of one's life accompanied by a song that the writer strongly associates with that aspect, or publication of one's diary containing extensive quotations from a poem.

Historically, most personal uses of this nature would, as a practical matter, pass under the radar of copyright enforcement.[170] Even a zealous copyright owner would find it prohibitively expensive to detect an analog physical copy of his work exchanged interpersonally within a limited circle of acquaintances. Technological developments and associated social-cultural patterns have changed this situation. Digital technology now allows cost-effective detection of many personal uses that in the past were invisible, especially when the information is exchanged through the Internet.[171] So-called digital rights management technology, backed by legal sanctions, allows "self-enforcement," meaning technological restrictions that simply prevent users from engaging in certain uses.[172] And the digital environment is sprinkled with intermediaries whose control power, surpassing any equivalent in the analog world, can be leveraged through legal threats and commercial inducements to detect and restrict the activities of individuals that were once beyond the reach of the copyright owner.[173] Examples involving all three elements include the content filtering systems applied by various dominant intermediaries[174] and the recent proliferation of graduated response policies by Internet Service Providers.[175] These changes in the prospects of detection and enforcement forces us to confront an important policy question, one that in the past lay fallow due to technological constraints: to what extent should copyright's exclusive entitlements encompass personal uses?[176]

---

170. Rothman, *supra* note 53, at 467.

171. Julie E. Cohen, *Comment: Copyright's Public-Private Distinction*, 55 CASE W. RES. L. REV. 963, 964–65 (2005); Jessica Litman, *Reforming Information Law in Copyright's Image*, 22 U. DAYTON L. REV. 587, 606 (1997).

172. Julie E. Cohen, *A Right to Read Anonymously: A Closer Look at "Copyright Management" in Cyberspace*, 28 CONN. L. REV. 981, 983–89 (1996). The main legal support for technological protection measures is provided by the anti-circumvention sections of the Digital Millennium Copyright Act. 17 U.S.C. § 1201 (2012). *See* Glynn S. Lunney Jr., *The Death of Copyright: Digital Technology, Private Copying, and the Digital Millennium Copyright Act*, 87 VA. L. REV. 813 (2001); Pamela Samuelson, *Intellectual Property and the Digital Economy: Why the Anti-Circumvention Regulations Need to Be Revised*, 14 BERKELEY TECH. L.J. 519, 534–37 (1999).

173. Litman, *supra* note 171, at 606.

174. *See* Annemarie Bridy, *Graduated Response and the Turn to Private Ordering in Online Copyright Enforcement*, 89 OR. L. REV. 81, 103–06 (2010).

175. On graduated response, see *id.*; Mary LaFrance, *Graduated Response by Industry Compact: Piercing the Black Box*, 30 CARDOZO ARTS & ENT. L.J. 165, 171 (2012).

176. We bracket here an important related question about an affirmative user's right to access information. Unlike the issue we analyze in the text, this related question is not whether users should enjoy a privilege to engage in personal uses, in the sense of being free

From the vantage of efficiency, the answer depends mainly on a straightforward transaction-cost analysis, hinging on whether or not transaction-cost problems hinder smooth and relatively inexpensive licensing of rights for personal uses.[177] In the absence of significant transaction-cost impediments, personal uses (now subject to actual enforcement ability) provide a channel for copyright owners to internalize the social value of works, thereby increasing incentives for production of supramarginal works.[178] Under such conditions, there is also no reason to expect that the inframarginal social cost imposed by copyright on these uses will be unusually high compared to other entitlements.[179] The analysis may be complicated by other considerations such as the fact that the newfound enforcement power in regard to personal uses largely depends on technological means and actions by intermediaries, which may be over-inclusive in their sweep.[180] In light of this, it may be difficult to generate definite efficiency-based conclusions. Nevertheless, the analysis will likely be driven by the transaction-cost factor and decided by one's empirical conclusions about the nature of transacting over personal uses.

How does the analysis change under some of the democratic alternatives? Perhaps surprisingly, any divergence between self-determination and

from legal rights by another to prevent such uses or obtain remedies for them. It is, rather, whether users have any affirmative rights to compel those who own or control protected works to grant access to them. *See* Dan L. Burk & Julie E. Cohen, *Fair Use Infrastructure for Copyright Management Systems*, 15 HARV. J.L. & TECH. 41 (2001).

177. Note that transaction costs divide here into two main categories: impediments to transactions between the copyright owner and the user, and impediments to internalizing the positive effects of the use on other parties that may cause a failure of the bargaining process to reflect the true social value of the use. *See* Loren, *supra* note 100, at 1.

178. PAUL GOLDSTEIN, COPYRIGHT'S HIGHWAY: FROM GUTENBERG TO THE CELESTIAL JUKEBOX 23 (2003) (praising the technological promise of fine-grained ability "to compensate copyright owners each time their works are chosen").

179. To be sure, the more fundamental question is whether, given the current level of copyright protection, any social value derived from the increased incentive outweighs the access cost of extension of copyright to encompass private uses. To avoid path dependence, however, the rational way to conduct this inquiry is to ascertain the incentive/loss ratio of the personal uses entitlement and decide its desirability relative to the ratio of other entitlements. *See* Fisher, *supra* note 24, at 1707–17. The analysis in the text suggests that absent special transaction-cost problems, a personal use entitlement is not likely to generate a particularly low incentive/access ratio compared to other entitlements and therefore it is unlikely to be found inefficient.

180. Mark A. Lemley & R. Anthony Reese, *Reducing Digital Copyright Infringement Without Restricting Innovation*, 56 STAN. L. REV. 1345, 1386 (2004); Lunney, *supra* note at 172, 830–44; Samuelson, *supra* note 172, at 535–37; Pamela Samuelson, *Mapping the Digital Public Domain: Threats and Opportunities*, 66 LAW & CONTEMP. PROBS. 147, 160–62 (2003); Alfred C. Yen, *Internet Service Provider Liability for Subscriber Copyright Infringement, Enterprise Liability, and the First Amendment*, 88 GEO. L.J. 1833, 1886 (2000).

efficiency is not immediately apparent here. Unlike critical uses, personal uses as broadly described above do not strongly correspond with higher-order conditions for self-determination that may take precedence over satisfaction of existing preferences. The poster on YouTube of a video of a young child dancing to the sounds of a copyrighted song finds, no doubt, the act rewarding and valuable.[181] But the broad availability of materials of this kind is not a precondition for genuine self-authorship. The video and many personal uses like it hardly seem germane either for critical reflection and possible revision of one's choices, or for a meaningful variety of robustly different alternatives. The value for the user in exercising his free choice by posting the video should be counted, of course. But it does not enjoy any normative priority over other subjective preferences, such as those of other users who prefer more consumptive choices and fewer opportunities to engage in personal uses. The analysis seems to collapse into efficiency's aggregation of effects on subjective preferences.

A subset of personal uses, however, enjoys a different status. One category within this subset is strong identity-based uses. Consider Jennifer Rothman's example of the woman who accompanies a blog post about an assault she experienced with the song that played on the radio during that assault.[182] Is there anything about this use of a copyrighted song that gives it a normative priority over mere preference satisfaction? It is not the sort of use that is closely related to higher-order interests in self-determination. Yet it does have another normatively significant feature—the use involves an "urgent" interest, one going to the user's subjectively highly-valued or intense interest in publicly asserting her self-identity. Uses essential for such urgent interests form a distinct normatively privileged category.[183] They merit privileged treatment because they implicate core personal interests of users,

181. *Cf.* Lenz v. Universal Music Corp., 572 F. Supp. 2d 1150 (N.D. Cal. 2008) (holding, in a case involving a twenty-nine-second clip posted on YouTube of children dancing to Prince's "Let's Go Crazy," that copyright owners must consider fair use before issuing takedown notices pursuant to the Digital Millennium Copyright Act).

182. Rothman, *supra* note 53, at 516.

183. Within self-determination, then, we can distinguish between three levels of interests and corresponding uses of copyrighted works: (1) those uses going to ordinary preference-satisfaction, which are to be analyzed as part and parcel of standard efficiency analysis (subject to any distributive considerations—see Part IV, *infra*); (2) those uses implicating the user's higher-order interests in reflectively forming her preferences, which are to be given lexical priority over the satisfaction of existing preferences of others or even of the user herself; and (3) those uses that pertain to a user's urgent interests (typically by strongly going to self-expression or self-identification), which merit some priority against the aggregation of others users' interests in preference satisfaction.

interests so important and intense that they should not be placed on the same scale as ordinary preference satisfaction.

More specifically, some cultural materials become closely entangled with people's identity or persona.[184] Publicly asserting such aspects of one persona requires resort to these cultural materials. Placing such high-intensity uses, which go to the core of one's ability to publicly realize aspects of one's personal identity, on the same scale of comparison with mere consumptive uses seems wrong. The tendency of market-based economics to reduce everything to fungible values located along a single-scale is inadequate here. In both private and public contexts we usually do not treat such goods as fully commensurable. We do not make marginal adjustment of two units of self-identity to gain five units of entertainment. The specter of fungibility is particularly troubling when aggregating effects over multiple individuals because of the prospect of numerous miniscule individual interests sweeping away very high-intensity interests of the few. To many, except those thoroughly immersed in the market's reduction of all interests to a single aggregative scale, sacrificing one person's ability to engage in uses that go to the core of her persona for the sake of satisfying a multitude of small consumptive interests by others would seem problematic. At the end, however, a consequence-sensitive theory, even one that denies universal fungibility, must allow for some tradeoffs. One way of dealing with this tension is to insist on minimal thresholds for certain high-intensity interests independent of other interests. In our context this translates into allowing strong-intensity, identity-based uses of copyrighted works without consideration of their effects on consumptive uses by others. This is about as close as a consequence-sensitive self-determination theory gets to according "absolute" rights-like protection to particular individual interests.

It follows that self-determination theory would shield the interest of users to engage in strong identity-based uses irrespective of effects on consumptive interests, which means an exemption from copyright liability. Note that the justification is only for a targeted norm that exempts users who can plausibly establish their need to use the material for purposes of an urgent personal interest, rather than all personal uses.[185] The category of

184. Gordon, *supra* note 46, at 1569; Rothman, *supra* note 53, at 496–500. For the persona argument in the context of real and personal property, see Margaret Jane Radin, *Property and Personhood*, 34 STAN. L. REV. 957 (1982).

185. This leaves open hard second-order questions about how to implement the exemption: through an open ended standard that relegates the identification of high-intensity interests uses to a case-by-case ex-post analysis or through rules that identify such uses ex ante on a more precise level.

publicly asserting one's identity is not exhaustive and other subsets of personal uses may enjoy a similarly favorable treatment.[186] Nevertheless self-determination provides only a limited ground for allowing personal uses beyond what is justified by efficiency. It extends a normatively preferred status only to that subset of personal uses that either are plausibly understood as implicating high-intensity interests not fungible with consumptive preferences or as falling within the categories of critical uses or heterodox works that go to higher-order conditions for self-determination.

Flourishing theory offers a more inclusive support for a favorable treatment of personal uses. The key element here is that personal uses constitute one of the main sites of actual opportunity for many to engage in a meaningful activity in the cultural sphere. Personal uses, even those that do not involve a high-intensity interest such as realization of one's persona, are meaningful expressive activities in the sense of active and somewhat creative interaction with cultural materials. To be sure, the bulk of these users are not "strong poets" who radically break with conventions and challenge existing meanings and norms.[187] Nonetheless, they interact with cultural materials in ways materially different from passive consumption. Sometimes, these are ways that "require skill and concentration," involving challenge and problems that can be overcome "only through the exercise of initiative and creativity."[188] Other times, however, the active engagement may be even more modest, involving simply the exercise of one's critical judgment—or, simply, cultural tastes—among different works to match a specific occasion or mood. This exercise, when repeated and, importantly, communicated to others, contributes not only to a further cultivation of one's critical-creative faculties, but also to a deeper sense of the pleasures associated with their

186. It is even possible to imagine a situation in which a user has a valid claim that a private consumption use implicates a privileged core personal interest and therefore should be exempted. (We thank Yochai Benkler for pointing this out.) Suppose, for example, that John regards listening to a large variety of classical music not as a merely entertaining activity but as a core, constitutive aspect of his identity. Further assume that John's ability to effectively realize this aspect of his identity would be materially constrained by copyright-priced access. John may have a valid claim that his unauthorized reproduction of copyrighted classical music should be exempted in order to protect his privileged personality-based interest. Verifying and enforcing this type of claims involve, however, obvious concerns of over-inclusiveness and administrability that cut against applying the exemption in circumstances of private, consumptive uses, or at least for stipulating a very high standard of proof in such cases.

187. The term "strong poet" is taken from HAROLD BLOOM, THE ANXIETY OF INFLUENCE: A THEORY OF POETRY (1973). For an interpretation of the strong poet as a normative ideal, see RICHARD RORTY, CONTINGENCY, IRONY, AND SOLIDARITY 28–29 (1989).

188. Fisher, *supra* note 24, at 1747.

exercise and to a richer, more solidified sense of self, through recognition by others of one's public, externalized and embodied, expressive acts. To limit a concern for such meaningful activity to avant-garde uses, excluding milder variants such as the family video clip or the amateur song cover, would be to ignore it exactly in the contexts where it is most likely to be currently available for large swaths of ordinary people.

Because of the special value that flourishing theory places on making widely available opportunities for meaningful activity in the cultural sphere, it would insist on a privileged status for the broad category of personal uses even when efficiency is ambivalent or even opposed. It would support exempting from copyright liability a broad swath of personal uses, both traditional ones which in the past fell under copyright's enforceability threshold and new ones enabled by new technologies.[189] Here again flourishing theory supports creating the conditions that would nudge people toward activities and choices that are considered significant for the good life, even at the cost of falling short of maximizing existing subjective preferences.[190]

IV. DISTRIBUTIVE EQUITY IN COPYRIGHT POLICY

The distributive analysis of copyright faces a challenge similar to the one confronting democratic theories. Any consequence-sensitive approach to distributive equity must also attend, like economic analysis, to the multiple effects that variations in copyright protection will produce, effects tending to point in countervailing directions. Simply advancing distinct criteria to those of efficiency for evaluating such countervailing effects does not suffice to alter our resolution of the tradeoffs they pose. For instance, it is sometimes suggested that placing greater weight on the interests of the poor should, on that ground alone, result in a stronger emphasis (than that placed by efficiency) on the access costs of IP rights as compared to their incentive

189. There are important doctrinal implications to such a favorable stance toward personal uses. These include: a concrete statutory exemption for non-commercial personal uses; a modification of the current importance of the transformative nature of the use under fair use to cover, in appropriate cases, even mildly transformative personal uses; narrow readings of certain entitlements such as the public performance entitlement in 17 U.S.C. § 106(4) (2012) to exclude personal uses; and restricting third party liability as well as extending safe havens to internet intermediaries in ways that avoid incentives for over-zealous filtering of personal secondary works.

190. The ways such a stance may or may not be paternalistic and how it may, in any case, be justified, are discussed above. *See supra* text accompanying notes 161–71.

benefits.[191] But heightened concern for the poor does not by itself lead to this conclusion; just as it leads us to place greater value on their burdens from curbed access, so it should also lead us to place greater value on the benefits they would have reaped from any new innovations now foregone.[192] In other words, there is again a structural affinity to economic analysis that offers an important opportunity: the strong prima facie overlap between the prescriptions of efficiency and those of distribution suggests that we need to specify with greater precision those circumstances where taking into account distributive concerns may actually make a significant policy difference. Moreover, sharpening our analysis in this respect not only clarifies the practical differences between efficiency and distributive equity; it also uncovers some telling points of divergence between distributive and democratic theories.

One further challenge for distributive analysis of copyright is the relative paucity of theoretical writing in this vein, especially when compared to other, more richly developed strands of IP theory. Consequently, we begin our analysis by situating copyright-specific distributive concerns within a broader frame of distributive justice theory. Next, we briefly isolate those positive parameters of copyright protection that are most germane to pursuing the relevant normative concerns. We then take up some specific considerations of distributive equity, specifying how their prescriptions for the level and type of copyright protection may differ from those of efficiency. Finally, we briefly respond to some objections generally advanced in the literature against the pursuit of distributive concerns in fields of private law such as copyright.

---

191. *See, e.g.*, Abramowicz, *supra* note 17, at 105–06 (suggesting that a distributive concern for "low-income" individuals or communities justifies giving greater weight to access over incentive effects as compared to efficiency analysis); Margaret Chon, *Intellectual Property and the Development Divide*, 27 CARDOZO L. REV. 2821, 2891 (2006) (arguing that distributive "equality tilts the balance towards static efficiency and away from dynamic efficiency arguments, at least for resource-poor areas of the world").

192. *Cf.* Amy Kapczynski, *The Cost of Price: Why and How to Get Beyond Intellectual Property Internalism*, 59 UCLA L. REV. 970, 997 (2012). Kapczynski states:

> The problem with distributive justice arguments that favor IP exceptions is . . . the apparently self-defeating nature of these exceptions. Because IP both helps to create and constrains access to the goods that we want, we have the appearance of a paradox: If we make IP more flexible so that the poor have more access, the poor will also have less access—because the information that we want the poor to access is less likely to be created.

Id. Kapczynski goes on to state that "[t]here is in fact no true paradox here, because sometimes we can give the poor access to information goods without substantially undermining innovation." *Id.* Our focus here is on cases where a distributive justification for trimmed protection may remain strong even where incentives are reduced.

A. VARIETIES OF DISTRIBUTIVE CONCERNS

Theories of distributive justice are concerned with the fair distribution of goods across individuals or groups. The goods at issue in copyright are expressive activities, expressive works, and the surpluses associated with these. A full distributive theory of copyright would apply equitable principles of individual access to each and all of these, which is what we undertake to do here (with some limitations, identified below).

There is, however, a threshold objection to pursuing any distributive concerns in copyright that is worth attending to at the outset, since responding to it provides helpful guidance in delimiting and structuring our analysis. This is the view, as expressed by Jessica Litman, that copyright law should not attempt to address "distributional or equality issues" since these "stem from the outside, non-copyright world" and thus raise concerns that are "alien to copyright's rationale" or "purpose."[193] Further, distributive equity poses concerns that, in any case, copyright "would address . . . badly."[194] Similar arguments have been advanced in the more general law-and-economics literature against the pursuit of distributive equity within the confines of specific fields of private law, such as torts, contracts, and so forth. And, as with Litman's, such objections typically take two related forms. First, since distributive concerns stem from general, background inequities, it makes little sense to attempt to ameliorate them within the context of specific legal fields, each having their own particular, internal concerns (such as the regulation of risky activities in the law of accidents, or enforceable promises in contracts, and so forth). Second, the rules of private law typically furnish poor tools for achieving distributive aims. Regarding the first objection, our focus here is not on how copyright law might serve as a vehicle for correcting general distributive inequities at large, but rather on how we might attend to copyright's own distributive impacts. That is, if the subject of copyright is to regulate the production and dissemination of expressive works, there seems little reason in principle to bar taking into account how its rules might affect the distribution of benefits and burdens resulting from said production and dissemination. As for the objection that copyright law will tend, in practice, to be a poor tool for addressing any of the distributive concerns it implicates, this general worry can in fact be broken down to a number of distinct criticisms or obstacles, to which we attend implicitly throughout our analysis and address explicitly at the end of this Part.

193. Litman, *supra* note 171, at 617–18.
194. *Id.* at 618.

Distributive equity, in contrast to efficiency's aim of maximizing overall social benefits net of costs, is concerned with how such benefits and costs are spread across particular individuals or groups in society.[195] And a wide array of individuals or groups have been canvassed in the IP literature as possible candidates for special distributive concern, including: consumers as a group against producers;[196] poor consumers in particular;[197] resource-strapped follow-on users;[198] innovators as a group vis-à-vis society;[199] creators early in their careers;[200] and host of other discrete groups singled out by various provisions of the copyright statute, including schools, churches, veterans' groups and small businesses.[201] Less forthcoming, unfortunately, has been the supply of any normative criteria for determining which of these individuals or groups merit distributive priority, on what grounds, and by how much. Further, an additional ambiguity arises from the unclear boundary between strictly distributive concerns and values of the sort discussed above under the auspices of cultural democracy and flourishing

195. To be sure, this is a somewhat more restrictive sense of "distributive" considerations than has sometimes been deployed in the legal literature, where at times the term has been taken to refer to any value falling outside of efficiency and to cover a bewilderingly diverse array of concerns. Thus, it has been taken to refer, inter alia, to individual constitutional rights. *See generally* Laurence H. Tribe, *Constitutional Calculus: Equal Justice or Economic Efficiency?*, 98 HARV. L. REV. 592 (1985). In the private law sphere, it has been taken as an umbrella term to cover concerns of corrective justice, substantive morality, and the protection of settled expectations. *See* Duncan Kennedy, *Law-and-Economics from the Perspective of Critical Legal Studies*, *in* THE NEW PALGRAVE DICTIONARY OF ECONOMICS AND THE LAW 465, 470 (Peter Newman ed., 1998). Kennedy, it should be noted, suggests that this broader usage is implied by legal economists' own arguments regarding non-efficiency concerns. *Id.* at 470, 472. Nevertheless, the most prominent legal economists writing on this issue have tended to advance a rather more clearly delimited notion of "distributive" considerations. *See, e.g.*, KAPLOW, *supra* note 15 (distinguishing a concern with the overall distribution of benefits and costs, which may fall within the purview of welfarist analysis, from various additional "fairness" considerations that remain outside welfarism); A. MITCHELL POLINSKY, AN INTRODUCTION TO LAW AND ECONOMICS 7 (3d ed. 2003) (defining "equity," in contrast to "efficiency," as having to do with the general or overall distribution of benefits and costs, and excluding from its ambit case-specific conceptions of corrective justice or cognate notions of morality). Our reasons for adopting a more restrictive conception of distributive concerns, one closer to but still somewhat broader than that of the legal economists are given *infra*, text accompanying note 211.

196. *See, e.g.*, Daniel A. Farber & Brett H. McDonnell, *Why (and How) Fairness Matters at the IP/Antitrust Interface*, 87 MINN. L. REV. 1817 (2003); Yoo, *supra* note 33.

197. *See, e.g.*, Abramowicz, *supra* note 17, at 105; Daniel Benoliel, *Copyright Distributive Injustice*, 10 YALE J.L. & TECH. 45 (2007); Margaret Chon, *supra* note 191.

198. *See* Van Houweling, *supra* note 110, at 1547.

199. *See* Shubha Ghosh, *The Merits of Ownership*, 15 HARV. J.L. TECH. 453, 475 (2002).

200. *See* Robert P. Merges, *The End of Friction? Property Rights and Contract in the "Newtonian" World of On-Line Commerce*, 12 BERKELEY TECH. L.J. 115, 134 (1997).

201. *See* Litman, *supra* note 171, at 617.

theories (such as aspirations to decentralize meaning-making power or to broaden opportunities for more individuals to engage in a more active and meaningful way with their cultural environment).[202] Thus, before examining the difference made in evaluating the positive effects of copyright protection by switching our normative lens from efficiency to distributive equity, we need first to sharpen the latter lens by refining our sense of the relevant distributive concerns. We do so here by briefly distilling some central features and principles of theories of distributive justice in general, after which we draw out their key implications for the specific distributive context of copyright law and policy.

Theories of distributive justice address the question of the fair distribution of the burdens and benefits of social life. In their essentials, such theories take up issues along two dimensions: what counts as a fair distribution of some good, and what good or goods should we be principally concerned with fairly distributing? Somewhat more formally, the debate between contending distributive justice views may be organized in terms of the answers to two questions: (1) What good(s) should be the appropriate *space(s)* of distributive concern?; and (2) What *distributive principle(s)* should apply to the chosen space(s)?[203] Needless to say, each of these in turn raises enormously complex and vexing controversies. Fortunately, we do not need to delve too deeply into these here. For our purposes, it will suffice to distill some core features of the leading positions along three central aspects.

First, the main candidates for the appropriate space(s) of distribution are: resources (primarily, income and wealth as general all-purpose means);[204] utility or welfare (as the satisfaction of subjective preferences or hedonic end-states of experiential well-being);[205] opportunity for welfare;[206] and capabilities

202. *See, e.g.*, Fisher, *supra* note 77, at 1458–60 (grouping under the aegis of "distributive justice" normative aspirations to decentralize semiotic power and to facilitate a more participatory engagement with one's cultural environment). Our reasons for drawing a sharper distinction between these values and strictly distributive considerations are given below. *See infra* text accompanying note 211.

203. Or alternatively: (1) what is the appropriate currency or metric for distribution?; and (2) what is the appropriate distributive function?

204. *See* ERIC RAKOWSKI, EQUAL JUSTICE (1991); JOHN RAWLS, *Social Unity and Primary Goods*, *in* UTILITARIANISM AND BEYOND 159 (Amartya Sen & Bernard Williams eds., 1982); Ronald Dworkin, *What is Equality, Part 2: Equality of Resources*, 10 PHIL. & PUB. AFF. 283 (1981), *reprinted in* RONALD DWORKIN, SOVEREIGN VIRTUE: THE THEORY AND PRACTICE OF EQUALITY 65–119 (2000).

205. *See* Kenneth Arrow, *Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice*, 70 J. PHIL. 245 (1973); John C. Harsanyi, *Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility*, 63 J. POL. ECON. 309 (1955); Louis Kaplow, *Primary Goods, Capabilities, . . . or Well-Being?*, 116 PHIL. REV. 603 (2007).

(understood as an irreducibly plural set of discrete capacities to realize various doings and beings considered essential to "truly human" functioning,[207] or a life of flourishing[208]). These candidates differ significantly in their foundational premises, as well as in their practical implications for a wide variety of policy contexts. Nevertheless, all the variants in these different camps would acknowledge, as a component of each of their chosen spaces, the distributive significance of the goods disposed of by copyright law, namely the benefits and burdens associated with the production and dissemination of expressive works and their resulting surpluses.

Second, turning to the appropriate distributive principles to apply to the chosen space(s) of the good, although again debate is fiercely contentious, we can nevertheless identify three broad principles that would garner wide if not universal assent. The first is a priority principle, of giving greater weight to a certain amount of benefits or burdens when it is borne by those especially badly or worse off through no fault of their own (where badly or worse off is understood in terms of whatever is the chosen space(s) of the good).[209] The second, which may be called a desert principle, accords due weight to considerations of individual responsibility in labor-leisure and related choices,

206. *See* Richard J. Arneson, *Equality and Equal Opportunity for Welfare*, 56 PHIL. STUD. 77 (1989).

207. *See* MARTHA C. NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT (2000); AMARTYA SEN, INEQUALITY REEXAMINED (1989) [hereinafter SEN, INEQUALITY REEXAMINED]; Amartya Sen, The Tanner Lecture on Human Values: Equality of What? (May 22, 1979) [hereinafter Sen, Equality of What?].

208. G.A. Cohen, *On the Currency of Egalitarian Justice*, 99 ETHICS 906 (1989); Fisher, *supra* note 24, at 1755.

209. This principle is formulated broadly enough so as to encompass a variety of distinct views, which otherwise disagree, both on the normative basis for giving priority and on how much priority (or weight) should be given. For alternative normative premises, see, e.g., LARRY TEMKIN, INEQUALITY 245–82 (1993) (advancing the "telic equality" view that distributive inequality is in itself bad); Harry Frankfurt, *Equality as a Moral Ideal*, 98 ETHICS 21 (1987) (arguing that what matters is whether someone is "badly" off according to an absolute standard, rather than whether someone is "worse" off than others); Derek Parfit, Lindley Lecture at the University of Kansas: Equality or Priority? (Nov. 21, 1991); Talha Syed, *Equality, Priority & Justice in Needs* (2009) (unpublished manuscript) (on file with authors) (arguing that although distributive equality is not in itself valuable, comparative levels of well-being remain intrinsically significant on grounds of fairness, by contrast to considerations of either telic equality or non-comparative sufficiency). For discussion of alternative priority principles or weights, see *id.* (discussing, inter alia, absolute or lexical priority for the worst off; weighted priority for the worse off; absolute or lexical priority for the very badly off, up to a sufficiency threshold; and weighted priority for the very badly off, up to a sufficiency threshold). In addition to the preceding sources, others adopting some version of the principle in text include RAKOWSKI, *supra* note 204; RAWLS, *supra* note 88; Arneson, *supra* note 206; Cohen, *supra* note 208; Dworkin, *supra* note 204; Fisher, *supra* note 24; Sen, Equality of What?, *supra* note 207.

where these are deemed relevantly present and discernible.[210] And the third, which may be referred to as an efficiency principle, holds that everything else being equal, it is better to confer a greater added benefit on one person than a smaller one on another. We take up the implications of these principles for the copyright context momentarily.

Finally, a third central feature of distributive justice theories relates to how they delimit their domain, so as to mark a boundary between those spaces of concern to which their distributive principles apply, and other, distinct domains of political morality to be regulated by other considerations (such as "rights" or "democracy"). This of course pertains to the question raised above about how to distinguish between more or less strictly distributive values and those having to do, say, with cultural democracy. Two delimiting considerations are paramount in this respect. First, for many theorists, not all individual goods are properly conceived as susceptible to the sort of commensurable treatment involved in applying distributive principles to make tradeoffs across them within a single domain of "benefits and burdens" borne by individuals.[211] Thus, for resourcists like Rawls and Dworkin, considerations of individual "rights" (or "fair equality of opportunity") lie outside the domain of the distributive principles applicable to income, wealth, and similar goods, and should be governed by distinct internal considerations and enjoy special priority relations vis-à-vis economic interests.[212] In a related but distinct vein, capability theorists, whether or not they accord special priority to certain capacities above others, tend to demur from making straight tradeoffs between capacities, seeing them as qualitatively distinct and thus restricting the application of distributive principles to *within* rather than *across* various separate spheres (e.g., self-determination, preference-satisfaction, meaningful activity, health, and so forth).[213] A second consideration is that some interests or domains may best

210. *See, e.g.*, RAKOWSKI, *supra* note 204; SEN, INEQUALITY REEXAMINED, *supra* note 207; Arneson, *supra* note 206; Cohen, *supra* note 208; Dworkin, *supra* note 204; Fisher, *supra* note 24. *But cf.* RAWLS, *supra* note 88, §48 (rejecting any role in distributive justice for considerations of responsibility or desert).

211. Welfarists may be an exception here. To the extent that welfare is conceived in terms of being measurable along a single-scale metric, any categorical or qualitative boundaries will not be recognized and hence one set of distributive principles will be generically applicable across the entire gamut of "benefits and burdens."

212. RONALD DWORKIN, TAKING RIGHTS SERIOUSLY (1977); RAWLS, *supra* note 88.

213. Amartya K. Sen, *Plural Utility*, 81 PROC. ARISTOTELIAN SOC'Y. 193 (1981); G.A. Cohen, *Equality of What? On Welfare, Goods and Capabilities*, *in* THE QUALITY OF LIFE 9 (Martha C. Nussbaum & Amartya Sen eds., 1993); Martha Nussbaum, *Flawed Foundations: The Philosophical Critique of (a Particular Type of) Economics*, 64 U. CHI. L. REV. 1197, 1199–1203 (1997).

be conceived as irreducibly relational rather than individuated, requiring the application of distinct principles to those of discrete equitable distribution across individuals. Values such as democracy or community membership are salient examples. Democracy, involving the exercise of collective agency for the carrying out of irreducibly social ends, poses questions about the appropriate shares of individuals in participation, influence, etc., questions that are best conceived not in terms of distributive tradeoffs but rather in terms of inherently relational principles of equal standing. The upshot of these delimitations is that strictly distributive considerations will be taken to apply only when the benefits from expressive works are understood primarily in terms of satisfying the subjective preferences of individual consumers and producers. Once such benefits implicate others' interests, say in meaningful activity or self-determination or democratic participation, we need to supplement our distributive analysis with other regulative criteria—meaning, at the least, additional priority principles and even, perhaps, distinct relational norms.

Against this theoretical backdrop, we can identify the following as key concerns of distributive equity in the context of copyright. First, with respect to fair access to expressive works, distributive concerns may modify or supplement those of efficiency in the following central ways. Efficiency values access to works solely in terms of preventing deadweight loss from priced-out transactions, and it measures the magnitude of such loss in terms of foregone consumer and producer surplus. Distributive equity, however, may modify efficiency's evaluation in both respects: first, by adjusting upwards the cost represented by deadweight loss, and second, by factoring in as a distinct added cost an effect toward which efficiency is indifferent, namely the fact that the copyright-protected prices paid by consumers who nevertheless do access the work represent a transfer of surplus from them to producers. Regarding the measure of deadweight loss, in efficiency terms its magnitude is understood in terms of foregone "effective" preferences, or preferences backed by willingness and ability to pay. As is widely recognized, however, ability to pay, and hence the intensity of effective preferences, are strongly shaped by the existing distribution of income, wealth, and legal entitlements, for which there is no compelling justification.[214] Consequently,

214. *See, e.g.*, C. Edwin Baker, *The Ideology of Economic Analysis of Law*, 5 PHIL. PUB. AFF. 3 (1975); Ronald Dworkin, *Is Wealth a Value?*, 9 J. LEGAL STUD. 191 (1980); Lucian Bebchuk, *The Pursuit of a Bigger Pie: Can Everyone Expect a Bigger Slice?*, 8 HOFSTRA L. REV. 671 (1980); Duncan Kennedy, *Cost-Benefit Analysis of Entitlements: A Critique*, 33 STAN. L. REV. 387 (1981); Richard Posner, *The Problematics of Moral and Legal Theory*, 111 HARV. L. REV. 1637, 1669–70 & n.62 (1998).

from the point of view of distributive equity, application of the priority principle will tend to give greater weight to the unsatisfied preferences of individuals who are worse off through no fault of their own (such as those with disability or at lower income levels).

Next, turning to the incentive side, similar considerations apply regarding access to, and surplus from, any expressive works that are generated or foregone by changes in copyright protection. Perhaps more surprisingly, distributive values may also modify our assessment of the actual composition of the output of new or supramarginal works. The extent to which this may occur depends on the degree to which the subset of consumers with disability or lower incomes vary significantly, from the market at large, in their demand profile for expressive works. We canvass possible instances of such divergence below.

Finally, considerations of distributive equity may also, of course, influence what we regard as fair compensation for creators of expressive works. Three distinct issues may be singled out in this regard: (1) ensuring that creative workers receive fair compensation for their labors, both from employers and the consuming public; (2) ensuring an equitable division of the returns from follow-on or derivative works, between original creators and those building upon, adapting, or otherwise using the original work; and (3) ensuring that producer returns from expressive works are "reasonable" rather than excessive.[215] Unfortunately there exists, at present, a formidable barrier to directly pursuing these concerns through the vehicle of copyright: namely, the absence of any broadly accepted way of operationalizing plausible normative principles of equity-in-earnings (such as "to each according to effort and sacrifice")[216] and especially of doing so at the fine-grained level of copyright rules. We may observe, however, that to some extent efficiency's concern with ensuring the provision of necessary incentives for creative activity will tend to address (although crudely) the issues raised under the first two branches.[217] Similarly aspects of the third

215. *See, e.g.*, Ghosh, *supra* note 199, at 475–77.

216. For a review of the philosophical arguments and literature that converge on this as a rough principle of equity-in-earnings, see William W. Fisher & Talha Syed, *Global Justice in Healthcare: Developing Drugs for the Developing World*, 40 UC DAVIS L. REV 581, 670–72 (2007); see also RAKOWSKI, *supra* note 204; James Dick, *How to Justify a Distribution in Earnings*, 4 PHIL. & PUB. AFF. 248 (1975); Dworkin, *supra* note 204.

217. Efficiency calls for sufficient incentives to elicit from creative producers the effort and resources needed to generate those expressive works the value of which exceeds their costs of creation. In principle, such incentives need be just enough to ensure that the resources are deployed in this avenue rather than whatever is their owner's next-best alternative use. That is, efficient incentives are responsive to the bargaining power of market participants. And while equity may frown on some instances of paying producers only what

issue will be addressed, again only indirectly and partially, by attending to the above equitable concerns toward distributively prioritized subsets of consumers.[218] Thus, in what follows we do not give close attention to distributive aspects of innovator returns.

The foregoing previews the distributive issues raised by one half of the goods disposed of by copyright, namely access to expressive works (and their associated surpluses). What about the other half—access to expressive opportunities (or the ability of individuals directly to participate in expressive activity)? Where the barriers to such participation are on account of such individuals being worse off due to disability or lower-income levels, then distributive equity counsels prioritizing their access to such expressive opportunities in much the same way as for access to expressive works.[219] However, access to expressive activity may also merit special concern, on distinct grounds of the sort canvassed in Part III, such as cultural democracy's commitment to decentralizing of meaning-making power or flourishing's substantive emphasis on meaningful activity as valuable for its own sake. In such cases, how such opportunities should be valued in ways different from efficiency raises questions not so much of distributive priority within one space of the good, but of priority relations or tradeoffs between distinct spheres or spaces of the good.

B. KEY PARAMETERS OF POSITIVE ANALYSIS

In what circumstances might attention to the foregoing distinct concerns of distributive equity significantly alter our appraisal of the effects of copyright from those of efficiency? Recall that once we recognize that our alternate evaluative concerns must apply not only to one dimension of the analysis but, like efficiency, must assess the full range of relevant parameters, simply adopting alternative normative criteria to efficiency may often not

their bargaining power fetches on the market, the class of skilled producers that are involved at least in more traditional creative expressive work will tend not to prominently feature in that group. Of course some participants in a creative project—such as those on the "lower rung" of the creative hierarchy or those at more vulnerable stages of their career cycle—may be vulnerable to inequitably low payments; which is one reason why efficient incentives will only crudely address these concerns.

218. "Reasonable" rather than "excessive" producer returns would mean those returns deemed fair according to a general theory of the fair distribution of the social surplus, taking into account overall equity-in-earnings and equity-in-needs. Priority to individuals with disability and at lower incomes, while certainly not a comprehensive addressing of equity-in-earnings and -in-needs, is also not a negligible aspect thereof. Admittedly, the priority here is limited to access to expressive works and their associated surpluses, but so are the surpluses being disposed of.

219. We thank Terry Fisher for pushing us to clarify this point.

result in any appreciable policy difference. At the same time, however, *how* different positive effects are relevant may vary depending on the normative perspective adopted. Here, we briefly set out the distinct general parameters brought to light by a distributive as compared to efficiency analysis, before turning to specific circumstances in which these general differences may matter.

Adopting a standard incentive-access frame, assume we have at our disposal only crude levers of providing "more" or "less" copyright protection, generically applicable across all classes of expressive works. The efficiency analysis runs as follows: From a positive point of view, reducing protection will likely lower prices and hence increase access to inframarginal works, but also forego the generation of some supramarginal works. Normatively, this will only be worth it if the reduction in deadweight loss provides a greater benefit in terms of the satisfaction of effective preferences than the loss in producer and consumer surplus associated with the foregone supramarginal works. What do distributive considerations, specifically the placing of a priority weight on the preferences of worse-off consumers, add to this? A common supposition is that they may favor a reduction in protection even when efficiency does not, that is, when the efficiency loss in foregone innovations is greater than the efficiency gain from reduced deadweight loss. Why? Because we now give greater weight to the access gain, over and above its measure in terms of foregone surplus, by counting the interests of our distributively-prioritized group for more than what is reflected in efficiency's metric of effective preferences.[220] The trouble with this analysis, however, is that a similar distributive adjustment must also be made to the loss side of foregone innovations so that we place greater weight on the interests of our prioritized group's lost surplus from such ungenerated works, resulting in its amplification. Absent special reason to assume otherwise, a plausible default would seem to be that adjustments are symmetrical: as the access gain is enhanced over and above its efficiency measure, to that extent the loss from foregone works should also be adjusted upward. The result: a stalemate.

Upon closer analysis, however, crucial additional factors are disclosed, ones potentially obscured by an exclusive focus on the parameters relevant to efficiency. These factors increase the likelihood of a divergence between efficiency's prescriptions and those of distributive equity. Adopt, again, a standard incentive-access analysis with crude, generically applicable protection levers. Assume we are contemplating a move from protection

220. *See, e.g.*, Abramowicz, *supra* note 17; Chon, *supra* note 191.

level A to increased protection level B. Doing so will mean adding some protection over those works inframarginal to protection level A. This results in: (a) increased inframarginal deadweight loss, and (b) increased transfer of surplus from inframarginal works from consumers to producers. Both of these effects, while incentive-unjustified from the strict point of view of generating the inframarginal works, are incurred as an unavoidable byproduct (unavoidable if our tools are limited to crude, generic protection levers) of providing the added protection necessary to induce generation of those works supramarginal to protection level A. The benefits from such supramarginal works are: (c) the realized consumer surplus from the works, and (d) the realized producer surplus from them.

Efficiency cares about effects (a), (c), and (d), and is indifferent about (b). Distributive equity differs in two main respects. First, the distributive analysis may also care about (b), the incentive-unjustified transfer of surplus from consumers to producers. Second, distributive analysis may value the other three variables differently than efficiency. And these differences may matter *when the distributively-prioritized group is disproportionately affected by any of these effects*. The point is perhaps easiest to see with respect to factor (d)—the realized producer surplus from supramarginal works. To the extent that a distributively-prioritized population is significantly underrepresented within the group enjoying this benefit—i.e., producers—distributive analysis will downgrade its relative value compared to efficiency. The same principle applies, albeit likely less frequently, to (a) and (c). When a distributively-prioritized population is equally represented within the groups suffering from increased inframarginal deadweight loss and enjoying increased supramarginal consumer surplus, the significance of any different value placed by distributive equity on these two countervailing effects relative to efficiency will be nullified. However, the same does not hold when a distributively-prioritized group is disproportionately affected by either of these factors, as in a case where a disproportionate share of the increased deadweight loss falls on members of this group or when such individuals are unlikely to enjoy the increased supramarginal consumer surplus. In a nutshell, then, efficiency compares supramarginal benefits (c) and (d) against inframarginal cost (a) (all valued equivalently in terms of effective preferences); in contrast, distributive analysis potentially adds a further inframarginal cost (b) and may assign different values to (a), (c), and (d) without each of these adjustments offsetting each other.[221] Whether such

221. By contrast, it is unclear whether, at a general level, democratic theory would have any more grounds than efficiency for caring about (b)'s inframarginal transfer of surplus from consumers to producers. At the same, it would presumably share efficiency's concern

potential sources may plausibly become actual grounds of divergence depends on further normative and institutional considerations, including the identity of, and appropriate normative priority to be given to, different subgroups of consumers and producers, and the ability of copyright policy levers to tightly target specific subsets of such consumers or producers for distributive benefits or burdens.

C. APPLICATIONS

Which groups should merit distributive priority in copyright policy? We lay aside the possibility of giving such priority to the group of consumers as a bloc vis-à-vis producers as a bloc.[222] In the early days of the law-and-economics movement, regulatory policies based on such en masse favoring of consumers suffered sharp criticism due to their unclear normative basis, notwithstanding their significant roots in American political and legislative history.[223] In the absence of figures confirming substantial disparities between the average income/wealth profiles of the typical consumer and those to whom producer surpluses ultimately redound, the normative footing of any

for (a)'s inframarginal deadweight loss, (c)'s supramarginal consumer surplus, and (d)'s *supra*marginal producer surplus. This highlights, then, some important distinctions in the way that distributive and democratic theories relate to each other as well as to efficiency: For distributive theories, the prices of, and hence relative shares of surplus from, expressive works may matter a great deal (depending on the identity of the consumers and producers). On the other hand, all expressive works are valued equivalently, according to the subjective preferences of market actors (albeit suitably adjusted in magnitude, where priority principles demand it). For democratic theories, on the other hand, although a robust level of production and access to expressive works matters greatly for the sake of individual self-determination and social deliberation, who gets what share of the surplus from such works would not, in the first instance, be of much direct concern. At the same time, by contrast to a purely distributive focus, great emphasis may be placed on qualitative distinctions in the types of works and forms of engagement made available, such that even the satisfaction of existing preferences of distributively-prioritized groups, through the generation of further supramarginal works, may not be worth restrictions placed on critical uses of inframarginal works, even if those uses are engaged in by individuals or groups not meriting any distributive priority. One refinement of this important distinction is that democratic theories may care about the distribution of surplus from expressive works to the extent it influences effective access to particular, highly-valued, types of works.

222. *But cf.* Abramowicz, *supra* note 17, at 69 (treating consumers versus producers as a relevant group comparison for distributive analysis); Farber & McDonnell, *supra* note 196 (same); Yoo, *supra* note 33 (same).

223. For the criticism, see, e.g., RICHARD POSNER, ANTITRUST LAW: AN ECONOMIC PERSPECTIVE (1976). For discussion of the role of pro-consumer, as well as pro-small business, distributive concerns in shaping early American antitrust law and policy, see HERBERT HOVENKAMP, FEDERAL ANTITRUST POLICY: THE LAW OF COMPETITION AND ITS PRACTICE § 2.1 (4th ed. 2011); Robert H. Lande, *Wealth Transfers as the Original and Primary Concern of Antitrust: The Efficiency Interpretation Challenged*, 34 HASTINGS L.J. 65 (1982).

such distributive priority seems too unsure. Accordingly we focus on groups that more clearly merit distributive priority, namely individuals with disability or at lower-income levels. If these groups represent a non-negligible subset of the consumers for inframarginal works, there is a clear basis for divergence between efficiency and equity's appraisal of the value of lower protection and hence of increased and cheaper access (i.e., lower deadweight loss and smaller transfer of consumer surplus to producers). But the fact that a distributively-sensitive evaluation will tend to value an access gain from curbed protection more than efficiency does not mean that such gain is worth foregoing the benefits from stronger protection, when those benefits are also assessed from a distributive point of view. This is because the evaluation of the beneficial effects of stronger protection on the prioritized group must also be adjusted. This includes the supramarginal increase in consumer and producer surplus and the gain to producers from inframarginal transfer. The more evenly spread the distributively-prioritized group's share of the affected consumers and producers is on both sides of the infra-supramarginal tradeoff, the greater the chance of agreement between efficiency and distributive equity.[224]

If we wish, then, to pursue effectively distributive concerns in copyright, our focus should be on cases of significant asymmetry between our priority group's share of the inframarginal or supramarginal benefits of a measure vis-à-vis its share of the concomitant supramarginal or inframarginal costs. One implication of this is that, as a rule, sweeping or generic rules or policy levers—those ostensibly affecting the entire market in roughly similar ways, such as a shorter term or tighter enforcement—will tend not to be suitable candidates for pursuing specifically distributive considerations (put more precisely, their distributive evaluation is more likely to align with that of

224. Arguably the group of lower-income individuals is generally unevenly affected by the benefits and costs of protection. To the extent that lower-income individuals are significantly underrepresented within the group of producers, distributive analysis will downgrade the value of increased supramarginal producer surplus compared to efficiency. Similarly, the inframarginal transfer to producers will be of little countervailing value regarding the corresponding transfer from consumers. It follows that if this underrepresentation of lower-income individuals within the group of producers obtains systemically, distributive analysis will be systematically tilted in favor of weaker protection relative to efficiency. Difficulties in assessing the pervasiveness and magnitude of this effect, which is essentially a variant of the general argument for favoring consumers over producers, leads us to focus on other stronger cases of asymmetric effects on distributively-prioritized groups.

efficiency).[225] Rather, we need more targeted measures, those focused predominantly on benefiting our priority group(s), with the costs falling more evenly on between them and the rest of the market, thus creating the needed asymmetry.

1. Special-Needs Individuals

Section 121(a) of the Copyright Act provides an exception to copyright in nondramatic literary works, to allow "authorized entities" to make and distribute copies or phonorecords of such works in "specialized formats exclusively for use by blind or other persons with disabilities."[226] A similar, but somewhat distinct, provision is included in the Marrakesh Treaty recently adopted by the World Intellectual Property Organization.[227] The provision requires all member countries to provide in their national laws "a limitation or exception to the right of reproduction, the right of distribution, and the right of making available to the public . . . to facilitate the availability of works in accessible format copies for beneficiary persons."[228]

To a considerable extent, both the existing domestic and the international provisions are precisely the sorts of targeted exemptions that the preceding analysis suggests are best suited for the effective pursuit of distributive aims. To be sure, each might also pass muster on efficiency grounds as well. The impact on incentives of such exemptions is likely negligible, while the gain in consumer surplus for the serviced population is likely high (even when measured in strictly effective preference terms). But suppose that this were not the case—say because somehow specialized formats for the blind were a reasonably lucrative domestic market or the international provisions were somehow broad or leaky enough to pose a non-trivial threat to dynamic incentives. The point is that, nevertheless, these provisions would remain

225. This is subject to our discussion below of how, under certain conditions, generally applicable measures may have asymmetric effect on lower-income individuals. *See infra* paragraph following note 250.

226. 17 U.S.C. § 121 (2012).

227. The Marrakesh Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired, or Otherwise Print Disabled was adopted on June 27, 2013. Marrakesh Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired, or Otherwise Print Disabled, June 27, 2013, *available at* http://www.wipo.int/edocs/mdocs/copyright/en/vip_dc/vip_dc_8_rev.pdf [hereinafter Marrakesh Treaty]. It will go into force after it is ratified by twenty member countries. *See Historic Treaty Adopted, Boosts Access to Books for Visually Impaired Persons Worldwide*, WORLD INTELL. PROP. ORG. (June 27, 2013), http://www.wipo.int/pressroom/en/articles/2013/article_0017.html.

228. *See* Marrakesh Treaty, *supra* note 219, art. 4(1)(a). Article 3 contains a lengthy definition of a "beneficiary person" that amounts to a person who is blind or suffers from other impairments detailed in the Article resulting in a reading disability. *Id.* art. 3.

firmly supported on distributive grounds. The reason for divergence with efficiency is the asymmetry of the measure's effects. Because the rule exclusively channels the access benefit to blind and print-disabled people, a clearly prioritized group reaps the lion's share of inframarginal benefits (which are thus highly magnified by our distributive criteria over and above their efficiency measure), while suffering only a proportionate share of supramarginal costs (which are correspondingly less divergent from their efficiency measure).[229] By contrast, the efficiency analysis would compare the same access benefit with the entire supramarginal cost without magnifying the cost and benefits effects on the prioritized group. As a result, distributive equity supports the exemption even when efficiency does not, and supplies a stronger justification for it even when efficiency points in the same direction.

The targeted exemption for reproduction in formats suited for the blind supplies a template for a legal lever that can create asymmetrical effects in the service of a distributive policy. Specifically, when a copyright exemption or limitation targets its access benefits exclusively or almost exclusively to a distributively-prioritized group (and there is no reason to assume that the prioritized group is similarly overrepresented among those incurring the supramarginal cost), it will confer disproportional benefits on that group, thereby increasing the likelihood of it being justified by the distributive policy.[230] There are two main doctrinal routes for achieving this effect: precise rules and case-specific applications of fuzzy standards.[231] The rule track embodied in section 121(a) targets the access benefits to the distributively-prioritized group through a precise definition of the realm of the exemption. The alternative route involves applying a general, open-ended standard such as "fair use" in a way that relies heavily on the existence in a specific case of access benefits narrowly targeting a favored group. In *Authors Guild v. Hathitrust* the court did just that when it held the unauthorized reproduction of copyrighted literary works, as part of a service supplying

229. As for the inframarginal transfer ignored by efficiency, it is unlikely to add negative drag to the distributive analysis and it probably pushes it further in the positive direction. This is so because the prioritized group enjoys the entire benefit of the transfer to consumers while suffering only a fraction of its negative effect on producers according to its relative share of its members among the group of producers.

230. To be sure, theoretically even a measure that disproportionately benefits a prioritized group may impose a large enough cost on that group (and others) that it may not be justified from the perspective of the distributive policy. The point is merely that the more disproportionate the benefit conferred on the group, the higher the likelihood it will be distributively-justified and the greater the divergence of the analysis from efficiency.

231. On the rule/standard distinction, see Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L. J. 557 (1992); Duncan Kennedy, *Form and Substance in Private Law Adjudication*, 89 HARV. L. REV. 1685 (1976).

access to blind and print-disabled individuals, did not infringe on a ground alternative to section 121(a). The court found that because of "the unprecedented ability of print-disabled individuals to have an equal opportunity to compete with their sighted peers," such a service constituted fair use.[232]

Having clarified the mechanism that facilitates effective distributive prioritization through structuring of copyright exemptions, a few additional observations that bear generally on using copyright to promote distributive equity ends are in order. First, what is the normative justification for giving priority to persons with disabilities? The basic grounds for such distributive priority are two-fold: (a) as a result of their disability, such individuals may have considerably less earning potential than average through no fault of their own; and (b) even if we took care of undeserved disparities in earnings, such individuals will tend to have "expensive needs," meaning conditions that make it comparatively more resource-costly to reach the same level of well-being (on any plausible candidate metric) as someone without disability, such that even with the same bundle of resources they will tend to be significantly worse off.[233] Even if we bracket here the first ground, the second by itself provides ample support for the subsidized production of disability-suitable formats of expressive works.[234]

It follows that the existing exemptions are likely too narrow. Both section 121 and the international provision are limited in regard to both the beneficiaries of the exemption[235] and the categories of works within its

232. Authors Guild, Inc. v. HathiTrust, 902 F. Supp. 2d 445 (S.D.N.Y. 2012). Specifically the court found that the use was highly transformative and had no negative impact on relevant markets for the copyrighted work, two crucial factors in the fair use analysis. *Id.* at 463.

233. This way of framing the distributive justice issues raised by disability is developed in Syed, *supra* note 209.

234. This leaves open the thorny question of the extent of the justified priority accorded to those who are undeservedly badly-off, or worse-off than others, in the queue for social resources over those well- or better-off. We leave full treatment of this question, which is a matter of considerably greater controversy than the basic priority principle, for another day. One of us, however, has offered in other work the following as a benchmark ideal: we should aim to improve the access of those with disability up to the point that the next improvement per-unit resource that they will reap will be smaller as a ratio of their overall capacity/well-being than the foregone improvement of those without disability would be as a ratio of *their* overall capacity/well-being. *See* Syed, *supra* note 209.

235. While the international provision is clearly limited to the blind or otherwise visually impaired (see Article 3), section 121(a) uses the deceptively broad term "blind or other persons with disabilities." The term, however, is specifically defined in section 121(d) by reference to another statute. Examination of the referenced legislation and related regulations shows that the category is limited to various forms of reading disabilities. *See*

ambit.[236] The distributive equity rationale, however, furnishes no general reason for a different treatment on the basis of either the nature of the disability or the categories of the works involved. As long as the assumption of expensive needs applies, the justification for distributive priority holds. At the same time, whether the assumption of expensive needs applies in regard to a specific disability and a particular category of works is a function of the interaction between technological, social and economic contingencies.[237] These considerations open up fertile ground for further investigation of the proper scope of disability-based copyright exemptions domestically and internationally.

A further question is whether it is justified to give priority to disabled individuals through a form of cross subsidy internal to the copyright regime, rather than through a more general tax and transfer scheme. Recall that the basis for distributive priority here is two-fold: undeserved disparities in earnings and expensive needs. Even supposing that the question of equitable earnings is best handled through general redistributive tax and transfer, issues of expensive needs will still often be better addressed in context-specific settings. This is so for two distinct reasons. First, the relevant dimensions of well-being (or spaces of the good) to which we want to improve access for those with disabilities strike many as irreducibly plural or "sphere-specific," with improvements in, say, pain amelioration, physical sensorimotor abilities, mental capacity, health, education, cultural capabilities, etc., each involving qualitatively distinct spaces of valued "doings and beings," spaces that are implausible to treat as commensurately substitutable with one for the

JOANNE KARGER, NIMAS AND NIMAC: A DISCUSSION OF THE LEGAL ISSUES ASSOCIATED WITH THE PROVISION OF ACCESSIBLE INSTRUCTIONAL MATERIALS TO STUDENTS WITH PRINT DISABILITIES 3 (2010), *available at* http://www.madisoniep.com/AssistiveTechnology/NIMASIDEA2004.pdf.

236. Section 121(a) applies to reproduction or distribution of a "previously published, nondramatic literary work." Article 2(a) of the Marrakesh Treaty defines "works" to which the exemption applies as "literary and artistic works within the meaning of Article 2(1) of the Berne Convention for the Protection of Literary and Artistic Works, in the form of text, notation and/or related illustrations, whether published or otherwise made publicly available in any media." An official note clarifies that this definition "includes such works in audio form, such as audiobooks." Article 4.1(a) requires member states to limit "the right of reproduction, the right of distribution, and the right of making available to the public," and Article 4.1(b) makes it optional for member countries to limit the right of public performance.

237. For example, it is an open question whether under current technological and economic conditions subtitles in films can reasonably be seen as an expensive need of individuals with a hearing impairment, or whether the market already meets this need with little or no extra cost to such individuals.

other.[238] Second, on the most plausible views of the appropriate distributive principle, the extent of distributive priority to those worse off in any one domain will depend at least in part on how resource-costly specific improvements are, and that is something more easily ascertained in concrete local settings rather than through a single overall monetary transfer. These reasons combine to yield the following conclusions: there will be a variety of distinct, capacity-specific, policy domains or spheres where attention to disability-equity is warranted, and in each sphere the primary focus should be on equitably addressing the domain-specific need, through improving or ameliorating the relevant dimension or capacity, rather than attempting to "compensate" for any failures to achieve equity in some other distinct sphere/capacity.

This reasoning bears two implications for copyright policy. First, it supports addressing the equitable concern of ensuring that persons with disability are able to realize the valued capacity of engaging with their cultural environment directly through the sphere-specific mechanisms of copyright law, rather than a general tax and transfer scheme. And second, the primary focus in this respect should be on addressing their "expensive needs" in this domain—namely, the need for disability-suitable formats. Addressing any general background inequity in earnings, by subsidizing access to expressive works over and above alleviating copyright burdens on the creation of specialized formats, should be secondary. Thus the exclusive emphasis of the domestic and international provisions on exemptions limited to special formats that address specific needs of disabled individuals, rather than subsidizing general access to works by such individuals through broad exemptions, receives some support.

2. Lower-Income Individuals

Under copyright, both the production and dissemination of works follow the market logic for allocating resources, namely: to each according to willingness and ability to pay. In other words, the types of expressive works that are generated is a function of the price signals sent by effective market demand, and the dissemination of such works to particular individuals is a function both of taste (or willingness to pay for accessing the work) and budget (ability to pay). A distributive concern centered on lower-income individuals reflects, then, unease about the possible effects of asymmetrical purchasing power in shaping the generation of, and/or access to, cultural works.

238. *See supra* note 213 and accompanying text.

The source of that unease lies in something like the following mix of positive and normative premises: Even accepting that some disparities in individuals' income-wealth holdings may be plausibly traced to individual choices regarding work, consumption and related decisions for which persons are reasonably held responsible,[239] nevertheless numerous other factors are also present—ranging from the contingencies of history and birth to the highly morally arbitrary character of market returns[240]—such that real-world disparities fall considerably short of being so justified. This is especially so when the disparities are of the order of magnitude confronting us today, where ten percent of the world's residents own roughly eighty-five percent of its wealth,[241] and similarly so in the U.S., with the top ten percent accounting for roughly half of all annual income and eighty percent of total wealth.[242] Application, then, of the priority principle—i.e., that those who are undeservedly worse off should be given greater weight in the claim for social resources—seems clearly to require policies that would significantly reduce such disparities. Two such policies of general scope would be a highly redistributive tax and transfer scheme (subject, of course, to said scheme's potential for baneful incentive effects that reduce overall productivity), or so-called "predistribution" measures, aimed at earlier-stage interventions to restructure market arrangements so as to attenuate the resulting disparities generated.[243] Unfortunately, such policies are clearly not in place, either in the

239. The question of which disparities are so justified is the central concern taken up by the "luck egalitarian" strand of distributive justice theory. *See* RAKOWSKI, *supra* note 204; Arneson, *supra* note 206; Cohen, *supra* note 208; Dworkin, *supra* note 204.

240. For a review of these factors, and the associated philosophical and economic literature adumbrating and analyzing them, see Fisher & Syed, *supra* note 216, at 670–72. The phrase "morally arbitrary" comes, of course, from the work of Rawls. *See* RAWLS, *supra* note 88, at 284. The concept has since been picked up and developed by the luck-egalitarian line of distributive justice theorists. *See supra* note 239.

241. With 40.1% going to the top 1%. *See* James B. Davies et al., *The World Distribution of Household Wealth* 8 (United Nations Univ.-World Inst. for Dev. Econ. Research, Discussion Paper No. 2008/03, 2008), *available at* http://www.wider.unu.edu/publications/working-papers/discussion-papers/2008/en_GB/dp2008-03.

242. With roughly half of those figures going to the top 1%. *See* EMMANUEL SAEZ, STRIKING IT RICHER: THE EVOLUTION OF TOP INCOMES IN THE UNITED STATES (UPDATE WITH 2007 ESTIMATES) 6–7 (2009), *available at* http://elsa.berkeley.edu/~saez/saez-UStopincomes-2007.pdf; Edward N. Wolff, *The Asset Price Meltdown and the Wealth of the Middle Class* 50 (Nat'l Bureau of Econ. Research, Working Paper 18,559, 2012); *see also* Thomas Piketty & Emmanuel Saez, *Income Inequality in the United States, 1913–1998*, 118 Q.J. ECON 1 (2003) (examining the trends of income inequality through the twentieth century).

243. The term "predistribution" was coined in Jacob S. Hacker, *The Institutional Foundations of Middle-Class Democracy*, *in* PRIORITIES FOR A NEW POLITICAL ECONOMY: MEMOS TO THE LEFT 33–37 (2011), *available at* http://www.policy-network.net/publications/4002/Priorities-for-a-new-political-economy-Memos-to-the-left. The underlying concept has

United States or worldwide, making existing distributions of purchasing power dramatically unjust.

Consequently, copyright's reliance on purchasing power, in principle one seemingly sensible way of translating people's preferences into signals of social value, can be troubling. This is not to say that copyright law should thus be harnessed as a vehicle for general distributive justice, in the service of ameliorating inequities in society at large. Doing so, by treating access to expressive works as a (highly imperfect) compensating currency for other inequities, would subvert copyright's aim of ensuring the robust and fair production and dissemination of expressive works. However, there is no reason in principle to reflect or reinforce background inequities here; rather, we should, to the extent possible, act to ameliorate undeserved disparities in access to expressive works.[244]

How to do so? Ideally we would try to distribute expressive works in accordance not with "willingness and current ability to pay," but with "willingness and justified ability to pay," with "justified" denoting the purchasing power that would be available to different individuals in a society where background distributive equity in earnings prevailed. In the absence of this fantastical ideal, we are left with using the imperfect proxy of giving some greater weight to benefits and burdens where they disproportionately fall on lower-income individuals. The aim is similar to that in the disability context: to identify circumstances where it is plausible to believe that a policy would produce an asymmetry between the benefits and costs experienced by

roots in two related strands of literature. One is the late Rawlsian development of James Meade's notion of "property-owning democracy." *See* JAMES MEADE, PROPERTY-OWNING DEMOCRACY (1964); MARTIN O'NEILL & THAD WILLIAMSON, PROPERTY-OWNING DEMOCRACY: RAWLS AND BEYOND (2011); JOHN RAWLS, JUSTICE AS FAIRNESS: A RESTATEMENT (2001); Martin O'Neill & Thad Williamson, *The Promise of Pre-Distribution*, POLICY NETWORK (Sept. 28, 2012), http://www.policy-network.net/pno_detail.aspx?ID=4262&title=The+promise+of+pre-distribution. The other strand is the parallel development of legal-realist insights into how markets are always already publicly constructed through law, into proposals for "democratizing the market" through "institutional experimentation." *See* ROBERTO UNGER, DEMOCRACY REALIZED (1998); ROBERTO UNGER, WHAT SHOULD LEGAL ANALYSIS BECOME? (1996).

244. One obvious question that arises at this point is whether we should concern ourselves with domestic or global disparities, especially where attending to one may exacerbate the other. Our focus here will be primarily on the domestic front, not out of any normatively compelling reason to limit distributive concern to the nation-state, but rather the more practical one that our primary focus here is on U.S. copyright law which, to a large although far from complete extent, is territorially bound in its scope of application. One clear exception to this of course is how the rules pertaining to territorial exhaustion of the first-sale doctrine interact with global trade in expressive works—a topic we touch on briefly at the end.

priority groups in terms of, say, increased inframarginal access to existing works versus foregone creation of new supramarginal works. We discuss here three such cases.

The first and most straightforward is when, as with disability, our priority group can be directly targeted through tailored exemptions, enabling it to reap a disproportionate share of the inframarginal/access gains from the exemption, while suffering only a proportionate share of the supramarginal/incentive costs. In such cases the distributive justification for the exemption will be particularly strong and will tend to diverge significantly with efficiency.[245] Are there such realistic cases? One possibility here would be directly factoring in the defendant's income as relevant to fair-use analysis, not only for follow-on creative uses but also for purely private, consumptive copying.[246] Another option, probably more feasible, is to justify and expand existing privileges for intermediary institutions and activities that are strongly correlated with access to or uses of expressive works by lower-income individuals. The Copyright Act includes several exemptions that fit this mold,[247] with the strongest case being the limited exemption in section 108 for public libraries in regard to certain reproductions of copyrighted works.[248] The shortcoming of this strategy is obvious: to the extent that non-low-income individuals enjoy the services of the beneficiary institutions—say public libraries—the targeting of the access benefits to the distributively-prioritized group is imperfect. But this more veiled approach also has some advantages beyond securing it greater prima facie legal legitimacy. As the social-welfare literature on universal versus "means-tested" entitlements points out, broadly inclusive social benefits tend to have solidaristic and destigmatizing effects.[249] Consequently, even if the presence of some higher-income beneficiaries dilutes the fine-tailored character of an exemption targeting intermediaries, that loss in short-term distributive advantage may be worth it in terms of longer-term institutional legitimacy and sustainability.

245. *Cf.* Kapczynski, *supra* note 192, at 998 (discussing possible instances of "exceptions to IP law" that "will disproportionately benefit the poor" while having "potential dynamic effects").

246. *Cf.* Van Houweling, *supra* note 110, at 1568–69 (proposing to factor in defendant's "ability to pay" in the fair-use analysis for follow-on derivative uses, although not for purely consumptive ones).

247. *See, e.g.*, 17 U.S.C. § 110(1)–(2) (2012) (listing certain exempted performances and displays by non-profit educational institutions).

248. *See* 17 U.S.C. § 108; *see also* Kapczynski, *supra* note 192, at 998 & n.108.

249. *See* Gillian Lester, *Can Joe the Plumber Support Redistribution? Law, Social Preferences, and Sustainable Policy Design*, 64 TAX L. REV. 313 (2011).

By contrast to targeted exemptions, it would seem that more general, sweeping measures, such as an overall reduction in protection, would not be well-suited for pursuing distributive aims as distinct from those of efficiency, for the familiar reason that our priority group will tend to be similarly situated on both the inframarginal benefit and supramarginal costs sides of the equation. However two considerations present possible departures from this default presumption and give reasons why lower-income individuals may not benefit from the creation of new works in proportion to their share of benefits from increased access to existing works. Both considerations operate within a product differentiation framework. Recall that many of the supramarginal works attracted by copyright protection are imperfect substitutes for inframarginal works.[250] The higher the level of protection, the greater the share of new works which are substitutes for existing ones and the smaller are the differences between the substitutes. Both considerations pertain to the fact that lower-income individuals are likely to enjoy a disproportionately smaller share of the value of close substitute works.

The first reason why higher income individuals may derive a bigger benefit from supramarginal substitute works is that they can afford to consume a greater number of them. A lower-income individual who can afford to consume fifty films a year derives some benefit from the existence of one hundred films from which to choose. But a higher income individual who can afford to consume all one hundred films derives a bigger benefit from their existence. This effect is exacerbated by the intersection of the cultural life of works with the term of copyright protection. A contrast with the case of patents over drugs may help illustrate the point. One plausible objection to a distributively-based claim for general reduction in patent protection—i.e., a reduction motivated by improving access to inframarginal treatments for lower-income individuals—is that such individuals would also stand to gain from any new treatments incentivized by the existing, stronger levels of protection. Or, if not them, due to patent-protected prices over the new treatments, then the next generation of lower-income individuals once the drugs go off-patent. This objection is premised on the plausible assumption that the social utility of drugs exceeds the duration of the patent term—so that many will ultimately benefit from new treatments even if they can't afford them in the initial patent-protected period. However, the same may not hold to nearly the same extent in copyright. When we consider the cultural life-cycle of many expressive works (especially those belonging to "pop" rather than "high" culture) in tandem with the extremely long

250. *See supra* text accompanying notes 33–35.

duration of copyright protection, it does not seem implausible that for some subset of the population, there exist a subset of expressive works with a "cultural expiration" date such that where copyright protection makes them prohibitively priced, they are then never actually enjoyed. As a result, the creation of more such works is really of no benefit to them (or alternatively is only of the small benefit of having a greater variety out of which to choose), while greater access to existing or inframarginal works is of huge benefit. In such cases a sweeping or overall reduction in protection may find significant additional support in considerations of distributive equity over and above efficiency because lower-income individuals will systematically derive a disproportionately small share of the supramarginal benefits of stronger protection.

A final set of possible cases of asymmetry also relate to a disproportion in the benefits derived by lower- and higher-income individuals with respect to variants of cultural works. But the source now is not the lack of effective access for lower-income individuals but rather the differential tastes of the two groups. Purchasing power may not simply function as the servant of one's tastes; it may also be a driver. Having more disposable income may open one up to developing a range of cultural tastes less likely to be shared by one's poorer self. These range from the pursuit of positional goods, to an increased "connoisseur's" appreciation of subtle distinctions between works in a similar genre, to the desire for large collections of cultural works for the sake of conspicuity, connoisseurship, or simply because one can. What all these tastes have in common is that they will tend to give greater value to "variety" or "differentiated" additions to existing genres or product spaces than those without them. And such additional works will tend, as copyright protection increases, to be an increasingly larger share of new, supramarginal works enabled by such protection. The result: to the extent that cutbacks to protection mainly reduce the entry of highly differentiated variants, the supramarginal costs will be disproportionately experienced by high-income individuals, leaving our priority group to reap a disproportionately greater share of the benefits from the decrease. This will justify in turn weaker overall levels of protection on distributive grounds.

In sum, there are some reasons why distributive concerns regarding lower-income individuals may, at least in certain cases, justify copyright policy results different from those generated by efficiency. The strongest subset of cases are those where copyright rules can target the benefits of increased access to the group of lower-income individuals either directly or through intermediary beneficiary institutions such as public libraries. A somewhat less firm case may be made on distributive grounds for overall weaker copyright protection in those domains where it is plausible to assume

that lower-income individuals enjoy a disproportionately small share of the benefits of supramarginal works attracted by higher levels of protection.

D. OBJECTIONS

Is the game worth the candle? Numerous objections to the pursuit of distributive concerns in fields of private law have been voiced over the past few decades.[251] Above, we addressed one general theme common to many such objections—namely that the values of distributive equity are inapposite because they are somehow external or "alien" to the core purposes of fields like copyright. Our answer, again, is that our aim is not to graft onto copyright external agendas and to use the field as an engine for pursuing general distributive justice. Rather, it is to attend to equitable concerns where they arise organically, internal to the field itself. Nevertheless, there remain a series of more discrete variations on this general theme, each in their way suggesting that sensitivity to distributive equity is quixotic at best and distracting or even damaging at worst. The principal criticisms are four and, adapting Albert Hirschman's classic list, we can group them into two clusters: *futile/perverse* and *unwieldy/distortionary*.[252]

1. Futile/Perverse

One line of criticism, most commonly voiced in the contracts context, is that any ex post imposition of distributively-motivated terms in parties' agreement will be, ex ante, anticipated and corrected for by private ordering on the market, thus rendering such distributive efforts ineffectual or futile.[253]

---

251. Our discussion here is focused on objections that accept, in principle, the normative legitimacy of distributive concerns and, further, the institutional legitimacy of attending to them in non-legislative arenas such as courts. *Cf.* FRIEDRICH A. HAYEK, LAW, LEGISLATION AND LIBERTY: THE MIRAGE OF SOCIAL JUSTICE 67–100 (1976); ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA 150–160 (1974); Richard A. Posner, *The Ethical and Political Basis of the Efficiency Norm in Common Law Adjudication*, 8 HOFSTRA L. REV. 487 (1980); Richard A. Posner, *Utilitarianism, Economics, and Legal Theory*, 8 J. LEG. STUD. 103 (1979). In our view, criticisms on grounds of normative or institutional legitimacy have little force, a point conceded by one of their most forceful early advocates. *See* Posner, *supra* note 214. The objections we take up here query, rather, the instrumental competence or suitability of any institutional body or mechanism besides legislative tax and transfer to effectively address distributive concerns.

252. ALBERT HIRSCHMAN, THE RHETORIC OF REACTION: PERVERSITY, FUTILITY, JEOPARDY (1991).

253. *See, e.g.*, Bruce Ackerman, *Regulating Slum Housing Markets on Behalf of the Poor: Of Housing Codes, Housing Subsidies and Income Redistribution Policy*, 80 YALE L. J. 1093 (1971); Richard Craswell, *Passing on the Costs of Legal Rules: Efficiency and Distribution in Buyer-Seller Relationships*, 43 STAN. L. REV. 361 (1991); Duncan Kennedy, *The Effect of the Warranty of Habitability on Low Income Housing: "Milking" and Class Violence*, 15 FL. ST. U. L. REV. 485 (1987); Duncan Kennedy, *Distributive and Paternalist Motives in Contract and Tort Law, with Special*

And if such terms are made compulsory or nonwaiveable, the response is that "you are hurting the people you are trying to help." In other words, to impose such terms upon private parties is to force upon them deals that they would not voluntarily negotiate, resulting either in the intended beneficiary paying more for the term than they subjectively valued it or in the intended provider supplying less of the good or simply refusing to contract with those likely to be the beneficiaries of ex post judicial solicitousness. Such results being, in a word, "perverse."

Hopefully, it is by now apparent that there is a structural affinity between these critical concerns and the challenge that we have been at pains to highlight and address in our analysis: namely, that the consequence-sensitive pursuit in copyright of any one set of goals must attend to the potentially countervailing effects that will result from the interaction of legal-policy decision and the responses of private market actors. Roughly speaking, what is called "ex post" in areas like contract is analogous to what we have termed the "inframarginal" access effects; while "ex ante" goes to incentive or "supramarginal" access effects. Our entire analysis, thus, has been aimed at articulating and operationalizing the parameters relevant to avoiding the sort of one-sided focus on access/inframarginal effects that risks either futility or perversity.

2. Unwieldy/Distortionary

A distinct line of criticism, one more closely associated with the torts context, highlights the incongruous fit between the ideal set of prioritized groups and desirable effects that might be specified by a plausible normative theory and the actual groups impacted by, and effects resulting from, private law rules, which are limited in their scope and flexibility to particular subject-matter domains and administration by courts through litigation.[254] This argument may be overstated in certain contexts such as ours, where the rules of copyright tend to have quite broad effects on the shape of, and access to, culture. Nevertheless, the general point is well taken: efforts to pursue distributive goals through specially-crafted categories do raise issues of leakage, imperfect reach, etc., and thus the question of whether we would do

Reference to Compulsory Terms and Unequal Bargaining Power, 41 MD. L. REV. 563 (1982); Anthony Kronman, *Contract Law and Distributive Justice*, 89 YALE L.J. 472 (1980); Anthony Kronman, *Paternalism and Contracts*, 92 YALE L.J. 763 (1983); Richard Markovits, *The Distributive Impact, Allocative Efficiency, and Overall Desirability of Ideal Housing Codes: Some Theoretical Clarifications*, 89 HARV. L. REV. 1815 (1976).

254. POLINSKY, *supra* note 195; Louis Kaplow & Steven Shavell, *Why the Legal System Is Less Efficient Than the Income Tax in Redistributing Income*, 23 J. LEGAL. STUD. 667, 674–75 (1994).

better to pursue changes in more broadly-impacting institutional fora. As we reflect on that question, however, it is important not to commit the "Nirvana fallacy" of comparing a realistic variant of one institutional alternative with an idealized version of its competitor.[255] If a distributively equitable system of tax and transfer or more structural reforms is not presently on the cards, it hardly seems a compelling objection to the pursuit of justifiable aims in other venues that these are an institutional second-best. Moreover, the two points are not unrelated. Rather, it seems plausible that two important sources of inaction on the progressive taxation front are the political power of those whose material interests are threatened by such taxation and the discursive-framing effects associated with "taxes." On both fronts, copyright law and policy may play a significant role: a more equitable copyright regime is one that contributes less by way of rent distortions into the political arena and exceptions to exclusionary rights—taking the form of "liberty" privileges—are often more easily framed and accepted than taxes.

One distinct and influential variation of this "too costly" line of criticism is Louis Kaplow and Steven Shavell's "double-distortion" argument.[256] Kaplow and Shavell urge that the pursuit of distributive aims through private law and regulatory rules is always institutionally second-best to the use of redistributive income tax, since both tools have distortionary effects on labor-leisure incentives while the private-law/regulatory alternative also introduces an additional distortion on account of pursuing an aim besides efficiency.[257] This argument, however, does not apply to the sorts of distributive measures we contemplate above, since the only "distortion" involved in such measures—i.e., the decrease in efficiency—is precisely the same as would be incurred by a redistributive progressive tax, namely an effect on dynamic production incentives.

V. CONCLUSION

We have offered in this Article what we think is an attractive reading of two normative frameworks that recently have gained prominence in the field of copyright law and policy: democratic theories and considerations of

255. Harold Demsetz, *Information and Efficiency: Another Viewpoint*, 12 J.L. & ECON. 1, 3 (1969).

256. Kaplow & Shavell, *supra* note 254; *see also* Aanund Hylland & Richard Zeckhauser, *Distributional Objectives Should Affect Taxes but Not Program Choice or Design*, 81 SCAND. J. ECON. 264 (1979); Steven Shavell, *A Note on Efficiency vs. Distributional Equity in Legal Rulemaking: Should Distributional Equity Matter Given Optimal Income Taxation?*, 71 AMER. ECON. REV. 414 (1981).

257. Kaplow & Shavell, *supra* note 254, at 668.

distributive equity. Our construal of these theories as consequence-sensitive opens up a space between the utilitarian and natural-rights theories that traditionally have dominated the field. Like the consequentialist theory of efficiency and in contrast to natural-rights theories, consequence-sensitive theories factor into their normative analysis the full array of consequences attributable to a legal rule, including forward-looking incentive effects. The theories in our group diverge, however, from efficiency regarding both the values they seek to promote (and thus with which they identify, evaluate and prioritize the relevant consequences), and the role(s) they assign to expressive works and activities vis-à-vis these values. This combination of robust consequence-sensitivity with a nuanced, non-reductive, and (we think) attractive appraisal of the relevant consequences should serve to bolster the case for this family of normative views when compared to its long-standing rivals.

A major theme of the analysis has been the structural homology between efficiency and both democratic and distributive theories construed as consequence-sensitive. These theories, we have argued, face countervailing supramarginal and inframarginal effects of copyright that closely track familiar tradeoffs at the heart of copyright's economic analysis. Proponents of democratic and distributive theories are urged to take this structural homology seriously—it is no longer plausible simply to point at values different from efficiency and assume that this alone prescribes results materially different from those commended by economic analysis (typically on the side of greater access). This general homology does not mean, however, that democratic and distributive theories collapse into efficiency, leaving natural-rights and negative-liberty frameworks as the sole alternatives to it. Close examination reveals significant cases in which various democratic theories support copyright arrangements distinct from those mandated by efficiency, along with ways in which copyright legal levers can be tweaked to effectively pursue distributive equity goals in divergence with the scriptures of efficiency. Attention to the relevant values and the ways they interact with the actual effects produced by copyright rules in specific contexts reveals the stakes in the choice between efficiency and democratic or distributive theories. Similarly, the same analysis sharpens the differences between distributive considerations, centered on an equitable distribution of the surplus associated with creative works as subjectively evaluated by individuals, and democratic theories, for which the core values are centered around qualitative distinctions in the types of expressive works and activities differentially affected by copyright. Finally, our analysis clarifies the differences between the various outlooks we gather under the loose title of democratic theories.

Those distinctions are not merely scholastic. They make all the difference for purposes of normative justification and institutional design in regards to important contemporary copyright policy questions, such as the appropriate treatment of critical uses, personal uses, or attending to important concerns of distributive equity implicated by copyright rules. In terms of the long history of copyright theory and policy, democratic and distributive theories are still relative newcomers. Hopefully, our analysis has sharpened the analytical and normative toolkit with which the growing number of proponents of these outlooks are able to defend, elaborate, and pursue their normative commitments in copyright law and policy.

FREE TRADE IN PATENTED GOODS: INTERNATIONAL EXHAUSTION FOR PATENTS

Sarah R. Wasserman Rajec†

ABSTRACT

Modern international trade law seeks to increase global welfare by lowering barriers to trade and encouraging international competition. This "free trade" approach, while originally applied to reduce tariffs on trade, has been extended to challenge non-tariff barriers, with modern trade agreements targeting telecommunication regulations, industrial and product safety standards, and intellectual property rules. Patent law, however, remains inconsistent with free-trade principles by allowing patent holders to subdivide the world market along national borders and to forbid trade in patented goods from one nation to another. This Article demonstrates that the doctrines thwarting free trade in patented goods are protectionist remnants of long-abandoned pre-Industrial Age economic theories, and the modern arguments for restricting international trade in patented goods—most notably, the possible desirability of permitting price discrimination—provide an insufficient justification for restricting trade across national frontiers. The Article concludes that modern patent law doctrine should be modified to permit free international trade in patented goods and that, if price discrimination or other goals are thought desirable, better alternatives are available to achieve those goals.

© 2014 Sarah R. Wasserman Rajec.

† Lecturer in Law and Fellow in Law, Science & Technology, Stanford Law School. I would like to thank Sean Bottomley, Robert Brauneis, Dan Burk, John Duffy, Rebecca Eisenberg, Andrew Gilden, Paul Goldstein, Tomás Gómez-Arostegui, Timothy Holbrook, Mark Lemley, Benjamin Liu, Amelia Rinehart, Jacob Sherkow, Marketa Trimble, Molly Van Houweling, Peter Yu, and participants in the World Intellectual Property Organization's Seventh Advanced Intellectual Property Research Seminar, the 2013 Intellectual Property Scholars Conference, the Sixth Annual Junior Scholars in IP Workshop at Michigan State University College of Law, and faculty presentation at Brooklyn Law School, Drexel, Fordham, Northern Kentucky University, University of New Hampshire, The Ohio State University Moritz College of Law, The University of Pittsburgh School of Law, and William & Mary Law School. I would also like to thank Max Wallot for his excellent research assistance.

TABLE OF CONTENTS

I. INTRODUCTION 318

II. PATENT LAW AND INTERNATIONAL TRADE: CONFLICTS AND CONFLUENCES 322

A. PATENT LAW PERSPECTIVE 323
B. TRADE LAW PERSPECTIVE 328

III. RECONCILING PATENT AND TRADE THEORY: EVOLUTION IN TANDEM 331

A. MERCANTILIST TRADE AND PATENT LAWS 333
B. MODERN PATENT LAW: A FOCUS ON INNOVATION 339

IV. DOCTRINAL DEVELOPMENTS 342

A. NATIONAL PATENT EXHAUSTION: DOCTRINE AND THEORY 342
B. A DIFFERENT RULE FOR INTERNATIONAL PATENT EXHAUSTION 348
C. THE UNITED STATES POSITION IN INTERNATIONAL NEGOTIATIONS 354

V. IMPLEMENTATION AND CRITIQUES 356

A. *KIRTSAENG* AS A MODEL: COPYRIGHT'S FIRST SALE DOCTRINE GOES GLOBAL 357
B. QUESTIONING THE BENEFITS OF GEOGRAPHIC PRICE DISCRIMINATION 361
C. IMPLICATIONS FOR OTHER FORMS OF PRICE DISCRIMINATION 367
D. IMPLEMENTATION 368
E. ADDRESSING CRITIQUES: THE PHARMACEUTICAL INDUSTRY 371

VI. CONCLUSION 376

I. INTRODUCTION

Modern international trade law seeks to increase global welfare by lowering barriers to trade and encouraging competition. Multilateral treaties such as the 1947 General Agreement on Tariffs and Trade[1] have significantly lowered tariffs and led to increased trade. What is more, the theories underlying modern trade law have been applied to non-tariff barriers to

1. General Agreement on Tariffs and Trade, *opened for signature* Oct. 30, 1947, 61 Stat. A-3, 55 U.N.T.S. 187 [hereinafter GATT].

trade,[2] so that the World Trade Organization ("WTO") includes agreements addressing subjects as diverse as telecommunication, industrial and product safety standards, and intellectual property. While the overriding purpose of the WTO is to encourage free trade, this principle has not been fully applied to patent law, placing the two fields in tension. In short, patent law grants rights that are fortified by national borders while trade law aims to diminish the relevance of borders.

Patent law has shed many geographical distinctions from its pre-Industrial Age roots that conflict with the formalities of free trade. These roots were planted in patent and trade theories that emphasized gains to the state, domestic industries and workers, and domestic consumers at the expense of other countries.[3] Under a mercantilist theory of commerce, patents were granted based on the introduction of new goods to a country, without a focus on invention.[4] Continued protection was based on the patent holder's willingness to produce goods locally and employ apprentices to learn the trade.[5] Patent law was not overly concerned with rewarding innovators for the sake of innovation, but rather with spurring domestic industry and supplying the domestic market with new goods. In contrast, patents are now available without regard to the location of invention, the nationality of the inventor, or the location (or fact) of production. These changes show a commitment to encouraging innovation for its own sake and a hands-off approach to effects on domestic industry and consumer welfare. The modern approach is consistent with modern trade theory, which eschews protectionism and posits net gains in worldwide welfare through decreased barriers to competition and trade.[6] One consequence of these changes is that any attempt to retain some sort of protectionism through exhaustion rules would not just run counter to current ideas about the role of patents and trade, but it would be misguided; protection of U.S. patent holders is not protection of U.S. industry or consumers.

Despite patent law's incorporation of some free-trade principles, however, patents remain territorial rights, and in one very important respect patent law remains at odds with modern trade theory. Under current law, a U.S. patent holder may block the importation, use, or sale of patented goods

2. JOHN H. JACKSON ET AL., LEGAL PROBLEMS OF INTERNATIONAL ECONOMIC RELATIONS: CASES, MATERIALS & TEXT 410–12 (4th ed. 2002).

3. *See infra* Section III.A.

4. *See infra* notes 79–89 and accompanying text.

5. *See infra* notes 83–84 and accompanying text.

6. *See generally* Alan O. Sykes, *Comparative Advantage and the Normative Economics of International Trade Policy*, 1 J. INT'L ECON. L. 49 (1998) (offering an explanation of comparative advantage, the case for free trade, and its caveats, *inter alia*).

purchased abroad, even if purchased from a seller licensed under a foreign patent. In contrast, an unconditional purchase of a patented good within the United States exhausts the patent holder's rights with respect to that good.[7] The doctrine of exhaustion—also called the first sale doctrine—advances consumer interests by limiting restraints on alienation and fosters efficient use of goods and competition by lowering transaction costs in resale markets while limiting the patent holder to a single reward for each sale. However, because there is no doctrine of international exhaustion for patents, there is no free trade in patented goods.

This topic is particularly timely given the conflict between the Executive and Judicial branches on the question of international exhaustion of intellectual property rights. One might expect that, given the United States' traditionally strong free trade stance in negotiations, the Executive branch would push for a rule of international exhaustion for patents. However, the United States Trade Representative is currently negotiating a trade agreement—the Trans-Pacific Partnership—requiring that member countries *not* recognize international exhaustion in intellectual property rights.[8] This stance is not uniform throughout the branches of government. For example, the Supreme Court recently held in *Kirtsaeng v. John Wiley & Sons, Inc.* that the first sale of a book published with the copyright owner's permission—anywhere in the world—exhausts the owner's rights in that book.[9] Thus, there was no infringement of a U.S. copyright when a graduate student imported and sold textbooks that were lawfully purchased from a licensed publisher in Thailand. Given this conflict, the Court will likely face the question of whether an international rule of exhaustion should be applied to patent law as well, although it need not follow the same rule as it established for copyright.

While neither trade nor patent theory supports the continuation of national exhaustion,[10] the effects of a harmonized international exhaustion regime in patent law are uncertain. The standard economic argument against international exhaustion draws on the potential gains to patent holders and to consumers in low-income countries from geographic price discrimination. This argument describes the current rule as allowing patent holders to market goods worldwide, adjusting prices for countries with lower purchasing power

7. *See infra* Part IV.

8. *See infra* notes 175–76 and accompanying text.

9. Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351, 1355–56 (2013).

10. In this Article, "national exhaustion" only recognizes exhaustion of the rights associated with patents in the country of sale. In contrast, "international exhaustion" applies exhaustion to relevant patents worldwide.

while continuing to reap rewards in high-income countries.[11] An international exhaustion regime, according to this view, will push patent holders either to restrict sales to high-income markets or to offer goods at a globally uniform price, to the detriment of consumers in low-income countries.[12] However, geographical price discrimination is but one of many options for identifying and marketing to populations with differing abilities to pay; many goods, regardless of patent protection, are available in different versions at different prices worldwide. Geographic price discrimination is desirable to firms because of its effectiveness at preventing arbitrage and because enforcement costs are shared by states through customs enforcement. It may not be the most desirable form of price discrimination for consumers, however, because it is imprecise in identifying differing demand curves. This is particularly true for countries with large or growing income disparities.[13] A shift to international exhaustion would likely result in changes in how firms market goods, but would not necessarily entail the wholesale welfare losses that the standard argument suggests, because that argument compares geographic price discrimination with no price discrimination at all.

The global welfare effects of an international exhaustion rule are more complex for the drug industry, given the heavy involvement of other regulatory regimes and a number of patent law measures that currently serve to remove the industry from typical market forces. In particular, geographical price discrimination may be more desirable in the drug industry because current price differentials reflect regulatory choices rather than demand differences, other forms of price discrimination may not be ethical, and access concerns tend to be more pressing in that industry.[14] While these concerns are valid, they are better met through regulatory regimes that already control market access in the industry or through trade mechanisms, rather than through maintenance of a patent law rule that no longer makes sense for other industries.

We can also expect that in industries amenable to it, a rule of international exhaustion would lead to new forms of restrictive licensing replacing outright sales. Software is particularly susceptible to restrictive licensing because of the technological possibilities of restricting the ability to resell or reuse it. Increased restrictive licensing may shift the exhaustion rule from an absolute bar on downstream constraints to a default rule that can be licensed around and is one method to retain price discrimination that need

11. *See infra* notes 217–20 and accompanying text.
12. *See infra* notes 217–20 and accompanying text.
13. *See infra* notes 222–23 and accompanying text.
14. *See* discussion *infra* Section V.E.

not be geographic in nature. However, there is nothing inherently "international" about this problem, although its importance may increase under an international exhaustion regime. Addressing the desired bounds and strength of such licenses is beyond the scope of this Article, besides noting that adoption of international exhaustion would leave the area ripe for further inquiry.

Adopting international exhaustion for patents would make this area of patent law consistent with the free-trade theory that informs the rest of patent law. Reducing barriers to trade in order to encourage efficient production and increased consumer welfare—the primary goals of international trade law—provides a compelling argument for an international exhaustion regime in patent law.

The Article proceeds as follows. Part II discusses modern theories of patent and trade law, and shows how the doctrine of international exhaustion fits into both. Part III describes mercantilist trade theory, its influence on patent law, and the evolution of both fields toward free trade principles. This Part situates national exhaustion as anomalous in modern patent law, raising the question of whether there are reasons to maintain the current rule of national exhaustion. Part IV explains the doctrinal development of patent exhaustion in the United States—both the domestic embrace of exhaustion as a means to limit transaction costs in downstream markets and the courts' rejection of international exhaustion, despite the applicability of the values embodied in domestic exhaustion doctrine. Part V looks at potential methods of implementing an international exhaustion regime and responds to criticisms such a change faces. In particular, this Part responds to some economic critiques of international exhaustion, addresses potential strategic reactions to such a rule, and develops a potential carve-out for pharmaceutical products. Part VI concludes.

II. PATENT LAW AND INTERNATIONAL TRADE: CONFLICTS AND CONFLUENCES

Patent law is territorial and protects markets, whereas trade theory has been on a relentless march to bring down barriers to cross-border trade and foster competition in manufacturing. Thus, at first glance it appears that their treatment of the question of international exhaustion for patents would also be directly opposed. However, a closer look reveals that patent law has been moving away from protection of national markets and that trade law may sometimes deviate from pure free trade principles in its approach to intellectual property matters.

A. PATENT LAW PERSPECTIVE

Patents and copyrights are exclusive rights granted to authors and inventors in order to encourage progress in science and the useful arts.[15] The regimes are governed by separate statutes and common law. Nevertheless, the constitutional underpinnings and much of the governing theories are the same. Through a time-limited right to exclude others from exploitation of their works, authors and inventors are encouraged to create, invent, and disseminate the fruits of their labor to the benefit of society at large. Excluding competition gives a patent holder the opportunity to sell goods at a premium price, the size of which will depend on demand for the invention and the availability of noninfringing substitutes.[16] Patents are "widely considered essential . . . to provide appropriate incentives for innovation."[17] They encourage investment in innovation by allowing inventors to reap a higher reward for their inventions during the patent term.[18]

Patents are granted by the U.S. Patent and Trademark Office ("PTO") for inventions that are new, useful, and nonobvious.[19] For the duration of the patent term,[20] its holder may bring suit against infringers seeking damages and injunctive relief.[21] Unauthorized use, manufacture, sale, or importation of a patented invention constitutes infringement that can be challenged in federal district courts, resulting in damages and/or a permanent injunction.[22]

---

15. U.S. CONST. art. I, § 8, cl. 8.

16. *See* Mark A. Lemley & Carl Shapiro, *Patent Holdup and Royalty Stacking*, 85 TEX. L. REV. 1991, 1997 (2007) (discussing the relevance of noninfringing substitute goods to reasonable royalty rates); *see also* Roger D. Blair & Thomas F. Cotter, *Rethinking Patent Damages*, 10 TEX. INTELL. PROP. L.J. 1, 13–14 (2001) (noting the effect noninfringing substitutes should have on damages calculations before discussing the difficulty of assessing substitutability).

17. JACKSON ET AL., *supra* note 2, at 922; *see also* Mark A. Lemley, *Ex Ante Versus Ex Post Justifications for Intellectual Property*, 71 U. CHI. L. REV. 129, 130 (2004).

18. *See* Pfaff v. Wells Elecs., Inc., 525 U.S. 55, 63 (1998) ("[T]he patent system represents 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.").

19. 35 U.S.C. §§ 101–03 (2012) (detailing subject matter, novelty, and nonobviousness requirements). In addition, patents require disclosure of the invention sufficient to allow others to reproduce it upon expiration of the right. 35 U.S.C. § 112 (2012).

20. Currently, the patent term lasts twenty years from the date of application. 35 U.S.C. § 154(a)(2) (2012).

21. *See* 35 U.S.C. §§ 283–84 (2012) (allowing courts to "grant injunctions in accordance with the principles of equity" and compensatory and punitive damages).

22. 35 U.S.C. § 271(a) (2012) (Infringement occurs when someone "without authority makes, uses, offers to sell, or sells any patented invention, within the United States or imports into the United States any patented invention" during the patent term). The statute

In addition, a patent holder can seek to exclude infringing imports from the domestic market by bringing suit at the International Trade Commission ("ITC") and proving importation of articles that infringe U.S. patents.[23] A U.S. patent cannot be used to stop infringing conduct abroad; the patent laws do not apply extraterritorially.[24] This is one reason the Patent Act designates importation of infringing goods as an act of infringement as well.[25]

The patent grant results in the social benefits associated with spurring innovation that would not have occurred—or would have occurred later—but for the incentive[26] as well as allowing others to benefit from—and build upon—the information contained in the disclosure.[27] There are also social costs, such as diminished access to the invention during the patent term for would-be users and downstream innovators engaged in incremental innovation. The diminished access is a result of the higher prices a patent holder can command during the patent term. The utilitarian analysis of the

was amended in 1994 to include importation as a means of infringement. *See* Uruguay Round Agreements Act, sec. 533, § 271, Pub. L. No. 103-465, 108 Stat. 4809, 4988–90 (1994).

23. Complaints may be initiated at the ITC alleging sale for importation or U.S. sales following importation of patented products or products that are produced through a patented process. 19 U.S.C. § 1337(a)(1)(B) (2012). The provision also outlaws importation of copyrighted materials. *Id.* Subsequent provisions outlaw importation of articles that violate U.S. trademarks or registered semiconductor chip masks, § 1337(a)(1)(C)–(D), and, pursuant to a 1999 amendment, articles protected by design patents, § 1337(a)(1)(E).

24. Pellegrini v. Analog Devices, Inc., 375 F.3d 1113, 1117 (Fed. Cir. 2004). However, the globalization and fragmentation of manufacturing supply chains has resulted in a changing definition of what constitutes extraterritorial application of patent laws and an expanded understanding of what constitutes infringing activity subject to U.S. patent law. Timothy R. Holbrook, *Territoriality and Tangibility after* Transocean, 61 EMORY L. J. 1087 (2012) (discussing expanded extraterritoriality); Timothy R. Holbrook, *Extraterritoriality in U.S. Patent Law*, 49 WM. & MARY L. REV. 2119 (2008) (suggesting that decisions involving extraterritorial application of U.S. patent protection should look to foreign patent law); *see also* Transocean Offshore Deepwater Drilling, Inc. v. Maersk Contractors USA, Inc., 617 F.3d 1296, 1309 (Fed. Cir. 2010) (holding that an offer to sell that was made abroad, but contemplated a sale "within the U.S., for delivery and use within the U.S. constitutes an offer to sell within the U.S. under § 271(a)"); NTP, Inc. v. Research in Motion Ltd., 418 F.3d 1282, 1317–18 (Fed. Cir. 2005) (affirming a finding of infringement of patented system despite location of Relay in Canada but rejecting infringement of method claims where the Relay step was performed in Canada).

25. 35 U.S.C. § 271(a). Even without the importation provision, any of the other prohibited actions (such as use or sale) of imported, infringing goods would have been actionable. Designation of importation and offers to sell as infringing activity was a result of the TRIPS Agreement. *See* Timothy R. Holbrook, *Territoriality Waning? Patent Infringement for Offering in the United States to Sell an Invention Abroad*, 37 U.C. DAVIS L. REV. 701, 722 (2004).

26. Michael Abramowicz & John F. Duffy, *The Inducement Standard of Patentability*, 120 YALE L.J. 1590 (2011).

27. Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539 (2009); *see also* Sean B. Seymore, *The Teaching Function of Patents*, 85 NOTRE DAME L. REV. 621 (2010).

costs and benefits of a patent system is the dominant one among patent law scholars,[28] attractive because it suggests that there is an ideal balance of interests and that—where that balance is not promoted—the law can be tailored to do so.[29] These tailoring measures are usually adaptations of the limits that already exist in patent law. Thus, from the basic structure and trade-off of patent law, scholars, courts, and the government have studied, proposed, and engaged in tailoring measures to address specific instances in which the law has led to holdup or other problematic outcomes that fail to meet the theoretical ideal balance of interests.[30]

One limitation on patent holders' exclusive rights is the doctrine of exhaustion, which frees from infringement downstream sales and uses of inventions initially sold with the particular patent holder's authorization.[31] As a result, the patent holder need not "authorize" each and every sale for subsequent sales and uses to be non-infringing, and would-be purchasers need not research and understand myriad restrictions attached to all the goods they purchase.[32] The exhaustion doctrine thus reduces transaction costs of disposing of purchased goods while vindicating ideas of consumer rights in the goods they own. It also limits a patent holder's control over sales and uses that may compete with her own sales, thereby fostering competition in resale markets. Another explanation for exhaustion puts it in

28. ROBERT P. MERGES, JUSTIFYING INTELLECTUAL PROPERTY 2–3 (2011).

29. Sarah R. Wasserman Rajec, *Evaluating Flexibility in International Patent Law*, 65 HASTINGS L. J. 153, 160 (2013).

30. *See, e.g.*, DAN L. BURK & MARK A. LEMLEY, THE PATENT CRISIS AND HOW THE COURTS CAN SOLVE IT (2009); Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 89 VA. L. REV. 1575 (2003); Michael W. Carroll, *One Size Does Not Fit All: A Framework for Tailoring Intellectual Property Rights*, 70 OHIO ST. L.J. 1362 (2009); Anna B. Laakmann, *An Explicit Policy Lever for Patent Scope*, 19 MICH. TELECOMM. & TECH. L. REV. 43 (2012) (suggesting patentable subject matter doctrine as a potential policy lever for calibrating patent scope); Gideon Parchomovsky & Michael Mattioli, *Partial Patents*, 111 COLUM. L. REV. 207 (2011) (suggesting two new patent forms that would mitigate the social costs of traditional patents and increase access by subsequent inventors); Sarah R. Wasserman Rajec, *Tailoring Remedies to Spur Innovation*, 61 AM. U. L. REV. 733, 742–48 (2012) (suggesting that courts analyze the public interest in encouraging innovation and promoting access in determining the appropriateness of a permanent injunction when a patent holder lacks market share).

31. *See* Quanta Computer, Inc. v. LG Elecs., Inc., 533 U.S. 617, 625 (2008); *see also* Amelia Smith Rinehart, *Contracting Patents: A Modern Patent Exhaustion Doctrine*, 23 HARV. J.L. & TECH. 483, 484 (2010) ("The right of a purchaser to control the downstream sale and use of patented goods without obtaining consent from the patent owner conflicts with the right of a patent owner to exclude others from practicing his invention when selling or using those goods."); *infra* Part III.

32. Molly Shaffer Van Houweling, *The New Servitudes*, 96 GEO. L.J. 885, 914 (2008) (discussing information-processing obstacles and costs associated with servitudes on chattel).

terms of what the inventor deserves—namely, a single reward for each product she sells—and no more.[33] Exhaustion has developed as a common law doctrine in the United States,[34] and is consistent with property law's aversion to restrictive servitudes[35] and restraints on alienation,[36] in addition to economic ideas about reducing transaction costs.[37] Typically called simply "exhaustion," the doctrine has generally been applied only to goods that have been sold within the United States,[38] and this Article refers to it as "national exhaustion" to distinguish it from international exhaustion.

International patent exhaustion would extend the current rule to patented goods first sold abroad, so that the first unconditional, authorized sale anywhere in the world would exhaust a patent holder's rights in the U.S. patent.[39] Currently, foreign sales do not exhaust domestic patent rights and unauthorized importation constitutes infringement. This position appears consistent with a traditional understanding of the scope of rights associated with a patent. Each nation's patent law has historically provided nothing more or less than territorial exclusion rights, whereas a rule of international

33. *See infra* Part IV.

34. *See infra* Part IV. A similar doctrine is that of "implied license," according to which the sale of a good comes with an implied license to use, repair, and sell it according to the common and reasonable expectations of the buyer. Although the system in the United States adheres more closely to the exhaustion formulation, the language of applied license is still used to draw the lines between "making" a new patented product and "using" a previously purchased product. *See, e.g.*, Bowman v. Monsanto, 133 S. Ct. 1761 (2013). Exhaustion for copyright also developed as a common law doctrine, but was later codified. *See* discussion *infra* Section V.A.

35. *See* Michael J. Madison, *Law as Design: Objects, Concepts, and Digital Things*, 56 CASE W. RES. L. REV. 381, 430–34 (2005) (discussing exhaustion cases and questioning the theoretical basis for limiting servitudes by recasting use restrictions as a type of product design).

36. *See, e.g.*, Herbert Hovenkamp, *Post-Sale Restraints and Competitive Harm: The First Sale Doctrine in Perspective*, 66 N.Y.U. ANN. SURV. AM. L. 487, 493 (2011) (noting that national exhaustion reflects "the common law's strong policy against restraints on alienation . . . ."); *see also infra* Part IV (discussing the doctrinal development of national exhaustion).

37. *See* Van Houweling, *supra* note 32, at 915.

38. *See infra* Part IV.

39. International exhaustion would not apply to goods made without authorization—for such goods importation constitutes infringement regardless. Nor would it affect the treatment of goods that a U.S. patent holder has authorized someone to import—this already renders the importation (and subsequent sales and use) non-infringing. Rather, an international exhaustion regime would mean that an authorized sale abroad exhausts domestic patent rights so that importation does not constitute infringement, even if the authorization was limited to sales in a particular foreign market—it is subsequent uses and sales that would not be subject to claims of infringement.

exhaustion frustrates territorial exercise of those rights.[40] A strictly national rule of exhaustion allows a patent holder to control the first sale of patented goods in the domestic market without competition from unauthorized imports, thus securing the market and allowing her to reap whatever reward the market will bear. In contrast, international exhaustion permits those who have bought patented goods abroad to use and sell them domestically without authorization from the patent holder, diminishing the patent holder's control of the domestic market. Just as a rule of exhaustion limits rewards to an inventor, so would a move from a rule of national exhaustion to one of international exhaustion.

However, the factors motivating a national exhaustion rule can also apply to goods purchased abroad. The nature of manufacturing, sales, and personal life are ever more globalized, making the information and transaction cost arguments from national exhaustion applicable to international transactions as well.[41] Margreth Barrett espouses the single reward argument in the international context and suggests that inventors can choose whether to place their goods into the stream of commerce abroad, but once they do, the opportunity for further reward should be extinguished.[42] And arguments about the proper scope of exhaustion and in what circumstances it can be contracted around also translate to the international setting.[43] Thus, although at first glance patent laws are territorial and would appear not to support a rule of international exhaustion, the theories that support it in a domestic

40. *See* ADELMAN ET AL., GLOBAL ISSUES IN PATENT LAW 3 (2011) ("[T]here is no such thing as a 'global patent.' ").

41. *See infra* Part IV (discussing the territorial limitations—within the United States—at issue in early exhaustion cases).

42. Margreth Barrett, *The United States' Doctrine of Exhaustion: Parallel Imports of Patented Goods*, 27 N. KY. L. REV. 911, 965 (2000) ("The key is that the patentee chose to make the initial sale. If a sale is not sufficiently beneficial, or undercuts his financial position in other sales, the patentee may refrain from making it."). Barrett's article is from 2000, when the U.S. position on international doctrine was somewhat less clear. She argued that the United States had a modified international exhaustion regime on the basis of doctrinal development to that point. *See id.*

43. There is nothing specific to domestic transactions that changes the relevance of arguments that, on the one hand, allowing for contractual circumvention of exhaustion when there is notice will result in more efficient transactions, or, on the other hand, the associated information and transaction costs attached to goods circulated in high volume are so high that consumers are not able to meaningfully distinguish them, and will therefore accept restrictions they are unaware of and that are ultimately inefficient. *Compare* Vincent Chiappetta, *Patent Exhaustion: What's It Good For?*, 51 SANTA CLARA L. REV. 1087 (2011) (arguing that exhaustion ought to be a default rule that can be contracted around), *with* Van Houweling, *supra* note 32, at 914–15 (discussing the notice and information costs associated with allowing servitudes to run with chattel).

setting apply equally internationally, and a historical understanding of patent law does little to justify the distinction.[44]

B. TRADE LAW PERSPECTIVE

Patent law interests only partially explain the arguments in favor of exhaustion. The choice between a national and international exhaustion regime sits squarely at the intersection of patent and trade law and theory. And from a distance, the patent law and trade law views on international exhaustion appear to be in conflict.[45] However, there are trade arguments on both sides of the issue as well. In the context of international trade law, national exhaustion may be considered a barrier to efficient trade on the one hand, or a means of price discrimination that accords greater access to countries with lower incomes, on the other.

Modern international trade law seeks to increase global welfare by lowering barriers to trade and encouraging competition. Increased trade leads countries to specialize in industries in which they have a comparative advantage over others.[46] This specialization, in turn, leads to scaling of industries, which achieves further efficiencies.[47] Comparative advantage may arise from differences in the abundance of factors of production, such as labor (skilled or unskilled), arable land, or even innovation—either in general

44. *See infra* Part III. Just as there is nothing specific to patent law that requires only national exhaustion, in the copyright context, the Court in *Kirtsaeng* explained that, although publishers might like to charge different prices in different markets, they could "find no basic principle of copyright law that suggests that publishers are especially entitled to such rights." Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351, 1370 (2013).

45. Abdulqawi A. Yusuf & Andrés Moncayo von Hase, *Intellectual Property Protection and International Trade: Exhaustion of Rights Revisited*, 16 WORLD COMPETITION 115, 116 (1992) (noting "the tension between free trade, which is a basic objective of the international trading system, and the exercise by private entities of the exclusive rights conferred under intellectual property legislation," and its manifestation in the application of exhaustion, in particular).

46. KRUGMAN ET AL., INTERNATIONAL ECONOMICS: THEORY & POLICY 24 (2012). The book gives an example in which the United States produces 10 million roses for Valentine's Day but, with the same resources, could have produced 100,000 computers instead. In contrast, sunny Columbia can produce 10 million roses easily, and devoting those resources to making computers would only yield 30,000 computers. The difference in price ratios means that roses are relatively more expensive to produce in the United States in winter than in Columbia, and vice versa with respect to computers. Trade allows the United States to stop growing winter roses and Columbia to shift its resources out of computer manufacturing. Both countries are better off than before. This is the (very) basic explanation of the gains from trade based on comparative advantage, generally attributed to the economist David Ricardo. *See* Sykes, *supra* note 6, at 55.

47. KRUGMAN, *supra* note 46, at 24.

or for specific industries.[48] Of course, if transportation costs are high or countries impose large tariffs on imports, it is harder for foreign goods to compete, even if they are made more efficiently than domestic goods. If the costs associated with trade are low, however, we can expect greater efficiency in manufacturing, leading to higher social welfare.[49] Tariffs are barriers to efficient trade because they artificially inflate the prices of goods from abroad and result in suboptimal levels of specialization.

The traditional explanation for the benefits from international trade and the barriers presented by tariffs underlies the Generalized Agreement on Tariffs and Trade ("GATT"), a multilateral agreement signed in 1947, the purpose of which was the "substantial reduction of tariffs and other trade barriers and. . . the elimination of preferences, on a reciprocal and mutually advantageous basis."[50] The agreement and subsequent rounds of negotiation resulted in thousands of tariff reductions affecting tens of billions of dollars of trade.[51] Further efforts led to establishment of the World Trade

48. Sykes, *supra* note 6, at 55–56 (noting that "nations with lots of innovators and skilled workers may tend to have comparative advantage in producing relatively new products, but over time comparative advantage may shift toward 'imitator' nations with less-skilled but cheaper work forces"); *see also* Anupam Chander, *How Law Made Silicon Valley*, 63 EMORY L.J. 639 (2014) (describing how local laws have given rise to the United States' comparative advantage in internet innovation).

49. Lowering tariffs is Kaldor–Hicks efficient—that is, it is a move that does not make everyone universally better off, but provides a net gain to the world when taking into account the losses. For example, workers in an industry that does not have a competitive advantage will be worse off if their plants close due to competition from imports. One solution to this is to pay the "losers" from trade out of the surplus. Trade Adjustment Assistance in the United States is such a scheme, providing some form of financial and educational benefits to workers in industries affected by trade. *See* Sykes, *supra* note 6, at 61 ("the removal of impediments to trade is likely to be Kaldor-Hicks efficient, but not in general Pareto efficient"); JACKSON ET AL., *supra* note 2, at 669–71 (describing Trade Adjustment Assistance).

50. GATT, *supra* note 1, pmbl.

51. *See* Douglas A. Irwin, *International Trade Agreements*, *in* THE CONCISE ENCYCLOPEDIA OF ECONOMICS 298 (David R. Henderson ed., 2008), *available at* http://www.econlib.org/library/Enc/InternationalTradeAgreements.html (explaining that the "annual gain from removal of tariff and nontariff barriers to trade as a result of the Uruguay Round Agreement . . . has been put at about $96 billion, or 0.4 percent of world GDP"). The negotiations that led to the World Trade Organization ("WTO") that subsumed and expanded on GATT recognized that as tariffs were reduced, other non-tariff barriers to trade would become more relevant obstructions to trade. One such potential barrier was intellectual property. The TRIPS Agreement addresses the barrier of different levels of IP protection and requires all member countries to implement a baseline level of intellectual property rights, but leaves it to members to determine the appropriate stance on exhaustion. Frederick M. Abbott, Second Report (Final) to the Committee on International Trade Law of the International Law Association on the Subject of the Exhaustion of Intellectual Property Rights and Parallel Importation (Sept. 6, 2000), *in* 69TH CONFERENCE OF THE

Organization and associated agreements to reduce tariffs and non-tariff barriers to trade as a means of enhancing welfare worldwide.[52] The TRIPS Agreement that forms part of the WTO, for example, characterizes variations in the protection of intellectual property rights as a non-tariff barrier to trade, and seeks to minimize such variations through minimum requirements for protection of intellectual property rights.[53] The removal of these various trading restrictions has fostered the growth of multinational companies and transnational supply chains.[54]

From a trade viewpoint, a national exhaustion rule may be characterized as a trading cost that hinders efficient downstream sales and uses of products because of the requirement to seek authorization for each contemplated resale market. This viewpoint appears to pit patent law against trade. For example, Frederick Abbott describes the national exhaustion regime as a claim that "the value of protecting intellectual property at the national and regional level exceeds the value to the world economic system of open trade among nations and regions."[55] A rule of international exhaustion, then, would limit the patent right after the patent holder authorizes sale in one country, paving the way for subsequent importation to—and resale in—all WTO member countries. This would result in more efficient manufacture and distribution through increased competition.[56] However, because even international exhaustion allows monopoly control over initial market placement, the trade literature demonstrates concern over a possible loss in global welfare if international exhaustion leads lower-income markets to be unserved or underserved.[57] In other words, unlike the typical comparative advantage story, patent holders can choose not to compete with imports by

INT'L LAW ASSOC., July 2000, at 13 (explaining that "[t]he TRIPS Agreement was designed to assure an adequate level of legal protection for the technology and expression components of goods and services in world trade.").

52. *See* Marrakesh Agreement Establishing the World Trade Organization pmbl., Apr. 15, 1994, 1867 U.N.T.S. 154 [hereinafter Marrakesh Agreement] (stating desire of members to substantially reduce tariffs and other barriers to trade in order to raise standards of living and employment levels, expand the production of and trade in goods and services, and allow for optimal use of world resources).

53. Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299 [hereinafter TRIPS].

54. *See* Irwin, *supra* note 51.

55. Abbott, *supra* note 51, at 13.

56. *See, e.g.*, Frederick M. Abbott, *Parallel Importation: Economic and Social Welfare Dimensions*, INT'L INSTITUTE FOR SUSTAINABLE DEVELOPMENT (2007), http://www.iisd.org/pdf/2007/parallel_importation.pdf.

57. *See* discussion *infra* Section V.B.

choosing not to sell abroad, thus undermining, rather than enhancing, the benefits of trade.[58]

III. RECONCILING PATENT AND TRADE THEORY: EVOLUTION IN TANDEM

Although at first glance patent law supports rights that are confined to national borders, and trade law aims to diminish the relevance of borders, the full story is more complicated. Each discipline offers counterarguments to its respective "first glance" position. However, an understanding of seventeenth century patent and trade law gives context to the emphasis patent law has placed on domestic enforcement of rights. This understanding shows how national exhaustion is built on protectionist ideas that have consistently been excised from patent law and leads us to question whether it is still justified.

The roots of patent law are entwined with those of international trade, and though their guiding theories have evolved, it has been an evolution in tandem. From mercantilist mechanisms of domestic market and industry control to free market mores that do not discriminate based on the origin of goods and encourage a global marketplace, the natures of patent laws and trade regulation have changed to fit evolving ends. And like patent law, trade law represents a balancing of interests.[59] The content of those interests and the level and type of regulation used to promote them are radically different now than during the seventeenth century. In trade, strong government involvement and restrictive border measures have given way to free markets, nearly nonexistent tariffs, and even reductions in non-tariff barriers to trade. In patent law, invention has become an essential requirement where before a willingness to introduce a new product was more important.[60] The origin of an inventor and the place of invention are no longer relevant to patentability

58. *See infra* note 213 and accompanying text. The economic arguments for and against international exhaustion will be discussed in Part V; here, it suffices to note that there is a counterargument to the unequivocal lowering of barriers to trade that has characterized much of modern trade law and theory.

59. "Since the time of the ancient Greek philosophers," writes Douglas Irwin, "there has been a dual view of trade: a recognition of the benefits of international exchange combined with a concern that certain domestic industries (or laborers, or culture) would be harmed by foreign competition." Douglas A. Irwin, *A Brief History of International Trade Policy*, LIBRARY OF ECONOMICS AND LIBERTY (Nov. 26, 2001), http://www.econlib.org/library/Columns/Irwintrade.html.

60. The Case of Monopolies, (1603) 77 Eng. Rep. 1260 (K.B.); 11 Co. Rep. 84 b. The Statute of Monopolies limited the ability of the sovereign to grant monopolies but included an exception for invention patents, thus carving out inventions as deserving of privileges that were otherwise becoming frowned upon. Statute of Monopolies, 1624, 21 Jac., c. 3 (Eng.).

standards.[61] And provisions meant to ensure local manufacture and domestic availability of goods have given way to laws that allow patent holders to enforce patents they do not practice.[62] These shifts away from protectionist methods of controlling markets are evident in modern trade and patent law. From this viewpoint, a rule of limited, national exhaustion appears as one more vestige of a national policy that regulated the market in order to entice the production and sale of innovative products in the patent-granting country. The elimination of such measures from trade law suggests their elimination from patent law as well. In addition, other changes in patent law consistent with free trade ideas have rendered a national exhaustion rule ineffective at achieving protectionist purposes. Situating the protection of national markets for sales in patented goods in its historical context makes a strong theoretical case for an international exhaustion rule.

---

61. William Hubbard explains:

> [U]nder the Patent Act of 1793, U.S. patents could only issue to "a citizen or citizens of the United States." In 1800, Congress amended the Patent Act to extend patent eligibility to foreign inventors, but only if those inventors resided in America for two years and took an oath of their intention to become United States citizens. For more than forty years, the United States offered no patent protection to the discoveries of nonresident foreign inventors, so that "foreign inventions could be introduced to America without the additional cost of the inventor's monopoly rights." In 1836, Congress amended the Patent Act to allow nonresident foreign inventors to obtain U.S. patents, but simultaneously introduced a protectionist scheme of patent application fees. The fee schedule required that a U.S. citizen pay $30 to file a patent application, a "subject of the King of Great Britain" pay $500, and any other nonresident noncitizen pay $300 to file an application for a U.S. patent. Protectionist application fees remained in place until 1870, when Congress established uniform fees for all patent applicants regardless of nationality or residency.

William Hubbard, *Competitive Patent Law*, 65 FLA. L. REV. 341, 356–57 (2013) (citations and quotations omitted). In addition, until 1994, U.S. inventors had advantages over foreign patent applicants in establishing relevant invention dates. *Id.* at 357. However, the TRIPS Agreement requires that patents be available "without discrimination as to the place of invention . . . ." TRIPS, *supra* note 53, art. 27. As a result of the TRIPS Agreement, the United States changed its law to allow use of foreign activity to prove invention dates. More recently, the United States has eliminated requirements of proving dates of invention in passing the America Invents Act, which requires that a patent issue to the first inventor to file as opposed to the first to invent. Leahy–Smith America Invents Act, Pub. L. No. 112-29, 125 Stat. 284 (2011).

62. For example, 35 U.S.C. § 271(d)(4) was enacted in 1988 allowing recovery for infringement even when the patent holder did not practice the patent and refused to license it. Patent Misuse Reform Act of 1988, Pub. L. No. 100-703, 102 Stat. 4674.

A. MERCANTILIST TRADE AND PATENT LAWS

The availability of goods from other lands has been recognized as desirable for thousands of years.[63] Patent law, too, has a long history; its origins have been traced to fifteenth century Venice.[64] This Section focuses on theories that emerged as dominant in the seventeenth and eighteenth centuries and which coincide with the rise of the industrial era.[65] During this time, the scale of trade and technological innovation grew dramatically, concurrent with "the rise of nation-states as political entities."[66] Although there were parallel themes in other countries, this Section focuses on England, both because it is representative and because English patent law provided the model for United States law.

The trade literature in England before and during the Industrial Revolution included varied ideas on the benefits of trade, but certain interconnected themes run through it. These pro-trade and pro-regulation views are broadly referred to as mercantilism. Mercantilist literature placed

63. DOUGLAS A. IRWIN, AGAINST THE TIDE 11 (1996) (quoting 12 PLUTARCH, *On Whether Water or Fire is More Useful*, *in* PLUTARCH'S MORALIA 299 (Loeb Classical Library 1927) ("the sea brought the Greeks the vine from India, from Greece transmitted the use of grain across the sea, from Phoenicia imported letters as a memorial against forgetfulness, thus preventing the greater part of mankind from being wineless, grainless, and unlettered")). And for just as long, it has been recognized that limits on that trade might be desirable as well. *Id.* at 14–15 (discussing Aristotle's suggestion that "import trade should stop at the provision of certain essential items, such as food for consumption and timber for shipbuilding").

64. CHRISTINE MACLEOD, INVENTING THE INDUSTRIAL REVOLUTION, THE ENGLISH PATENT SYSTEM, 1660–1880, at 10 (1988). Earlier exclusive grants for creative behavior have also been chronicled. *See, e.g.*, Ramon A. Klitzke, *Historical Background of the English Patent Law*, 41 J. PAT. OFF. SOC'Y 615, 617 (1959) (describing a decree by Roman Emperor Constantine in A.D. 337 exempting "artisans of certain trades" from all civil duties for working on their craft "and instructing their sons" in it). Klitzke also cites an account of an exclusive right to produce for inventors from 500 B.C.: "in Sybaris, a Greek colony famous for luxurious living and self-indulgence, if any confectioner or cook invented a peculiar and exclusive dish, no one else was allowed to make it for a year." *Id.* at 617 (citing Athenaeus, *The Deipnosophists*, *in* 3 BOHN'S CLASSICAL LIBRARY 835 (1854)).

65. *See* John F. Duffy, *Inventing Invention: A Case Study of Legal Innovation*, 86 TEX. L. REV. 1, 20–23 (2007) (describing fourteenth and fifteenth century grants of exclusivity as "proto-patents" and "quasi-patents," before suggesting that mid-fifteenth century Venetian law provided the model for other European countries to institute invention-based patent grants (quoting Hansjoerg Pohlmann, *The Inventor's Right in Early German Law: Materials of the Time from 1531 to 1700*, 43 J. PAT. OFF. SOC'Y 121, 122 (1961); F.D. Prager, *The Early Growth and Influence of Intellectual Property*, 34 J. PAT. OFF. SOC'Y, 106, 123 (1952))).

66. IRWIN, *supra* note 63, at 28 (generally providing excellent incarnations of theories that gained momentum and began to have worldwide impact when the industrial revolution led to a remarkable expansion in both trade and material amenable to patenting); MACLEOD, *supra* note 64 (same); Klitzke, *supra* note 64 (same).

particular value on regulating trade in order (1) to reach a favorable balance of trade—promoting national wealth and economic growth through the accumulation of gold and silver, (2) to increase employment, and (3) to protect and foster domestic industry.[67] Although mercantilism emphasized the potential benefits of trade to domestic interests, government regulation was considered important to counter situations where "merchants might pursue profitable commercial activities that could prove detrimental to the nation as a whole."[68] At the same time, mercantilist thought was by no means entirely based on consumer interests. The result, explains Thomas Nachbar, was that "not only profit but also free competition was discouraged, for while competition might maximize supply, it would result in prices too low for craftsmen to live on."[69] According to the mercantilists, national governments could provide incentives that would shape the actions of traders to enrich the nation as a whole by encouraging exports while discouraging over-consumption of luxury imports.[70] This in turn would maintain a favorable balance of trade. It also encouraged a composition of trade that would promote economic development and employment in manufacturing.[71]

The value of a favorable balance of trade was that a net exporting country became richer relative to other countries because of its increase in gold and silver stores.[72] The pursuit of a favorable balance of trade demonstrates how mercantilists were focused on the benefits of trade to a particular nation and the protection of national interests.[73] In this sense, mercantilists saw international trade as a zero-sum game. Regulation of the composition of traded goods was one of the means mercantilists suggested for controlling the balance of trade. For example, importing raw goods that were relatively cheap, promoting domestic manufacturing, and exporting the resulting goods would increase England's store of wealth. Regulation of the composition of trade served other purposes as well, promoting employment and the advance of domestic industry. The prevailing view was that the export of raw materials that would be used by manufacturers abroad was not

67. IRWIN, *supra* note 63, at 26 (quoting A. W. COATS, *Mercantilism: Economic Ideas, History, Policy*, *in* ON THE HISTORY OF ECONOMIC THOUGHT 40, 46 (1992)).

68. *Id.* at 31–32.

69. Thomas B. Nachbar, *Monopoly, Mercantilism, and the Politics of Regulation*, 91 VA. L. REV. 1313, 1318–19 (2005).

70. IRWIN, *supra* note 63, at 32–33.

71. *Id.* at 33–34 (discussing low or zero-interest government loans to merchants who were engaged in export).

72. *Id.* at 34–35.

73. *Id.*

in the national interest, whereas exporting manufactured goods was a benefit.[74] As Irwin explains, mercantilists thought that "[b]ecause processing activities generated more value and employment than other sectors, the economy should be oriented toward importing raw materials and exporting finished goods."[75]

It is also interesting to note the secondary role played by the interests of consumers. While international trade certainly made a greater variety of goods available to a greater swath of society, this was not seen as a driving interest for trade law. Trade regulation was effected through the grants of "exclusive trade privileges" of different sorts, including grants that resemble modern patent grants to inventors or grants for inventions and industries that the recipient had introduced to England, regardless of inventorship.[76] All of these exclusive grants were termed "letters patent," although this association was not necessarily beneficial—exclusive grants came under heavy criticism as they were frequently bestowed by the Crown upon favorite merchants.[77] The passage of the Statute of Monopolies in 1623 aimed to curb abusive grants of letters patent by the Monarchy, but did not necessarily end the use of monopoly grants for trade regulation by the state.[78]

Because there was no formal distinction between patents for invention and these other grants, it is no stretch to say that our patent system has its roots in trade law. Mercantilist values informed the development of the patent system and are evident in many of its early characteristics. Holger

74. *Id.* at 38 (suggesting that this is a proposition with which "[v]irtually all mercantilists would agree").

75. *Id.*

76. Nachbar, *supra* note 69, at 1323–25 (describing four categories of such privileges, all of which were called "letters patent": (1) patent-type privileges, (2) the exemption from other regulations, such as requirements to use English shipping or export prohibitions, (3) rights to supervise a particular trade, and (4) common trade monopolies unrelated to inventions).

77. *See* Duffy, *supra* note 65, at 25 (detailing how monopolies were granted for commonly available commodities such as "vinegar, salt, horns, iron, bags, [and] bottles . . . ."); Klitzke, *supra* note 64, at 632–33.

78. Scholars also point to the earlier *Case of Monopolies*, invalidating a monopoly on the manufacture and import of playing cards for its role in curbing abusive monopoly grants. *See, e.g.*, Duffy, *supra* note 65, at 26; *see also* Nachbar, *supra* note 69, at 1327–34 (providing a detailed recounting of the case and pointing to routine grants of exclusive trade privileges that occurred for years following the *Case of Monopolies* to argue that the case was "an assault on the monarchy, not on exclusive trade privileges"). For a discussion of one later, state-granted monopoly, see William A. Pettigrew, *Free to Enslave: Politics and the Escalation of Britain's Transatlantic Slave Trade, 1688–1714*, 64 WM. & MARY Q. 3 (2007) (discussing the Royal African Company's monopoly on slave trade, the political fight to end it, and the resulting expansion of slave trade).

Hestermeyer suggests that patent law developed primarily as "a means to promote the industrial advancement of the nation."[79] Similarly, Stephen Van Dulken writes that "[t]he patent system in England gradually evolved out of the royal prerogative used to encourage new trades, especially from abroad."[80] The English progenitors of patents, "letters of protection," were meant to encourage the development of new crafts and industry in the country. Letters of protection were granted "by the English crown to named foreign craftsmen, mainly weavers, saltmakers and glassmakers, with the intention of encouraging them to settle in England and transmit their skills to native apprentices."[81] The characteristics of these grants of privilege, and later monopolies, to those who brought industrial know-how to England reveal them as encouraging employment for citizens, self-sufficiency through technology transfer, and curbing the flow of money out of England.[82]

Local manufacture and the promise to take apprentices were seen as the important gains from patent grants.[83] The requirement that a patent be worked locally is consistent with mercantilist ideas about the balance of trade by ensuring that manufacture was done in England. Local working encouraged the importation of raw materials and the export of worked materials. It also supported domestic industry. In particular, by requiring that patent holders take apprentices, patent grants ensured that the industry could continue to operate even following the expiration of the patent term.[84]

---

79. HOLGER HESTERMEYER, HUMAN RIGHTS AND THE WTO, THE CASE OF PATENTS AND ACCESS TO MEDICINES 21 (2007).

80. STEPHEN VAN DULKEN, BRITISH PATENTS OF INVENTION 1617–1977, A GUIDE FOR RESEARCHERS 2 (1999). The first patent monopoly in England may have been granted in 1449, to John of Utynam, who returned to England from Flanders to "instruct divers lieges of the king in many arts never used in the realm besides glass making," and who was thereby allowed the right to grant or withhold consent for others to practice such arts for twenty years. Klitzke, *supra* note 64, at 627 (quoting ARTHUR ALLAN GOMME, PATENTS OF INVENTION: ORIGIN AND GROWTH OF THE PATENT SYSTEM IN BRITAIN 5–6 (1946)).

81. MACLEOD, *supra* note 64, at 10. MacLeod explains how the fifteenth-century Venetian patent system spurred the development of other, modern patent systems: "Emigrant Italian craftsmen, seeking protection against local competition and guild restrictions as a condition of imparting their skills, disseminated knowledge of their patent systems around Europe." *Id.* at 11.

82. Klitzke, *supra* note 64, at 628 (citing 1 WALKER ON PATENTS 3 (Anthony William Deller ed., 1937) (citing a letter from Chancellor Moreton to Parliament during the reign of Henry VII)).

83. Klitzke, *supra* note 64, at 624 (explaining how patent grants were conditioned on taking apprentices).

84. A number of scholars have suggested that the purpose of the fourteen-year patent term was to allow for teaching two generations of apprentices. *See, e.g.*, MACLEOD, *supra* note 64, at 18; P.J. Federico, *Origin and Early History of Patents*, 11 J. PAT. OFF. SOC'Y, 292, 304 (1929) ("The term of apprenticeship was seven years. Fourteen years may have been chosen

The characteristics of early invention patents align with mercantilist trade purposes rather than focusing solely on encouraging innovation.[85] For example, invention was not the sine qua non of the patent that it is today. As Christine MacLeod explains, "[t]he connection between inventing and patenting is historically tentative; it only started to be firmly established in the second half of the eighteenth century."[86] Instead, it was important that a merchant be introducing a previously unknown technology or good to England—this act, more than showing responsibility for invention, was of importance.[87] John Duffy further recounts how nonobviousness, a central characteristic of patentable inventions today, was of no import at all in the seventeenth century in England and did not become important there until the nineteenth century.[88] From the absence of these requirements, the

as the proper duration of a patent to allow for the teaching of several generations of apprentices."). *But see* EDWARD COKE, THE THIRD PART OF THE INSTITUTES OF THE LAWS OF ENGLAND: CONCERNING HIGH TREASON, AND OTHER PLEAS OF THE CROWN, AND CRIMINAL CAUSES 184 (W. Clarke & Sons 1817) (1644) (suggesting that seven years was an appropriate duration for patents for inventions and noting that this was the length of a single apprenticeship); Edward C. Walterscheid, *Defining the Patent and Copyright Term: Term Limits and the Intellectual Property Clause*, 7 J. INTELL. PROP. L. 315, 326–27 (2000) (suggesting that the term limit was instead a compromise between those who favored the traditional monopoly term of twenty-one years and those, like Lord Coke, who thought seven years was a sufficient term).

85. MacLeod explains that the romanticized notion of the individual inventor, deserving of a patent grant, was not associated with seventeenth and early eighteenth century patenting:

> There was no glory attached to being a patentee. The purchase of a patent was a commercial transaction. Patents were expensive to obtain, and nobody sought them without an economic end in view. This aim might have been to protect and exploit an invention; or it might have been to impress potential customers or investors; to escape the control of a guild, or to replace a guild's protective cloak, when that began to grow threadbare and competition to increase.

MACLEOD, *supra* note 64, at 7.

86. *Id.*; *see also* Klitzke, *supra* note 64, at 624. *But see* Sean Bottomley, Patent Cases in the Court of Chancery, 1714–1758, at 14 (unpublished manuscript) (on file with author) (pushing the dating back to the first half of the eighteenth century by showing that some inventors in that period were obtaining patents and enforcing them in court).

87. The purpose remained the encouragement of introduction of inventions "into the realm," rather than the encouragement of invention, qua invention. *See* Klitzke, *supra* note 64, at 627.

88. *See* Duffy, *supra* note 65, at 33. Duffy recounts the development of the doctrine of nonobviousness in the United States, tracing it to language imported from the French Patent Law of May 25, 1791 and enacted as part of the U.S. patent law in 1793. *Id.* at 35–36.

mercantilist values underlying early patent law are apparent: the availability of inventions was important; rewarding the inventor was far less so.[89]

This discussion of pre-industrial age English patent law has not specifically addressed the prohibitions on importation that require a choice between a national and international rule of exhaustion. The reason for that should be clear; these early grants were aimed at encouraging domestic manufacture of goods. Where patents were granted for importation, however, they carried the exclusive right to it.[90] The strong control of imports and exports by the government meant that the idea that someone else might be able to bring in goods and compete with a patent holder during its term was not contemplated and would have run counter to the purposes of the patent. Patents secured the national market to the patent holder, but came with the requirement of domestic manufacture—a requirement that would have been vitiated by the import of manufactured goods.[91] For

89. I do not mean to suggest that there was no interest in encouraging innovation, only that it was not the primary driving force behind regulations. *See, e.g.,* Jeremy Phillips, *The English Patent as a Reward for Invention: The Importation of an Idea*, 3 J. LEGAL HIST. 71, 74 (1982) (detailing a patent grant to George Cobham that states it is made "to encourage others to discover like good engines and devices." (citation omitted)).

90. A sample of cases ruling on the circumstances in which patents for imported goods could be extended show the disfavor in which importation was held. *See* VALE NICHOLAS, THE LAW AND PRACTICE RELATING TO LETTERS PATENT FOR INVENTIONS 138–39 (2d ed. 1904). Nicholas quotes various cases that held, somewhat enigmatically, that:

> The merit of an importer is less than that of an inventor. The fact that the invention was imported does not take away the merit, but it makes it much smaller . . . . If the imported invention is of considerable commercial value, and the importer has embarked a large capital in endeavouring to introduce it, the patent may be extended . . . .

Id.

91. Thus, the only cases from that time period of which I am aware address the validity and extension of a British patent for importation of goods manufactured abroad, under a foreign patent. One such case holds that a patent on such imported goods shall not be granted an extension in England. *Id.* at 139; *see also* ROGER WILLIAM WALLACE & JOHN BRUCE WILLIAMSON, THE LAW AND PRACTICE RELATING TO LETTERS PATENT FOR INVENTIONS 372 (1900). Wallace and Williamson describe the case of *Johnson's Patent* from 1871, in which a patent:

> had been taken out in England for an invention also patented in America, France, and Belgium. An extension of the patent had been granted in America, and the value of the invention was great; but it was shewn that the articles protected were manufactured exclusively in America and thence imported into England. The Committee refused prolongation.

Id. The court explained that with patents in both countries, the importer had brought goods to England that would have been imported even if there had been no English patent, with the only difference that "the patentee would not have had the monopoly of his patent, and would not, therefore, have had any opportunity of securing the monopoly of prices." *Id. But*

mercantilist England, the entire domestic market was the reward that was offered, but the price was local working, in keeping with notions of maintaining a favorable balance—and composition—of trade.

It was only after the Statute of Monopolies that invention became more important[92] and the requirement for a written description eliminated the requirements of local working—or practicing a patent at all.[93] These developments in patent law mirrored those in trade, and show that import restrictions are a relic of mercantilist theories that are at odds with current law.

B. MODERN PATENT LAW: A FOCUS ON INNOVATION

This Article has discussed the question of exhaustion from modern patent and trade perspectives, suggesting that they appear to point in different directions.[94] Approaching the possibility of an international exhaustion regime with the understanding of the evolution of patent law in concert with changes in trade theory casts the issue in a new light. In particular, the changes in patent law that increased the importance of innovating activity by a focus on rewarding the rightful inventor and limiting grants to inventions that are nonobvious are in line with the movement of trade toward a global welfare model. The intentions of the inventor—whether to introduce technology to the U.S. market or not—are no longer of such importance. The move from a working requirement to a disclosure requirement is also in line with the idea that invention and disclosure are themselves the benefits that patent law seeks to promote. Under this understanding, no specific protection of the national market is necessary, because there is no quid pro quo of local manufacturing or even local

see Sean Bottomley, The British Patent System During the Industrial Revolution, 1700–1852, at 78 (July 2012) (unpublished Ph.D. dissertation, University of Cambridge) (on file with author) (discussing *Berry's Patent*, an 1850 case in which the patent holder on an imported invention was granted an extension, with Lord Brougham stating that importation provides "a benefit to the public incontestably, and, therefore, they render themselves entitled to be put upon somewhat, if not entirely, the same footing as inventors." (citing PETER HAYWARD, 6 HAYWARD'S PATENT CASES 1600–1883, at 30 (1987)).

92. *See, e.g.*, Pennock v. Dialogue, 27 U.S. (2 Pet.) 1, 20 (1829) (discussing the importance of invention and novelty in the English Statue of Monopolies: "That act, after prohibiting monopolies generally, contains, in the sixth section, an exception in favour of 'letters patent and grants of privileges for fourteen years or under, of the sole working or making of any manner of new manufactures within this realm, *to the true and first inventor* and inventors of such manufactures, *which others, at the time of making such letters patent and grants, shall not use*.' " (emphasis added) (citing Statute of Monopolies, 21 Jac. ch. 3 (1624))).

93. John M. Golden, *"Patent Trolls" and Patent Remedies*, 85 TEX. L. REV. 2111, 2155 n.164 (2007).

94. *See supra* Part II.

availability. The aspects of mercantilist patent law that made a rule against international exhaustion a foregone conclusion are therefore discredited.

Modern patent law has shed many of the mercantilist trappings of its early incarnation. The United States has no local working requirement—and no requirement to serve the domestic market at all. John Golden explains how United States patent law has mostly avoided a working requirement and has "instead viewed disclosure of a patentable invention as essentially full compensation for a right to exclude . . . ."[95] The notion that the patent is issued "[i]n consideration of its disclosure and the consequent benefit to the community"[96] is a far cry from the rights discussed above that came with requirements to manufacture locally and train apprentices.[97] The focus on invention over introduction to the domestic market similarly supports the notion that spurring innovation, wherever it may happen and whatever its result, underlies modern patent law. These rules admittedly may result in less access to goods for domestic consumers and the potential for manufacturing industries to be relocated, but these possibilities are contemplated and accepted by free trade theorists.[98] By positioning patent holders to control market entry and manufacturing decisions, the current system embraces the free market notion that worthwhile inventions will become available and those that are not worthwhile will not receive undue encouragement.[99] The focus on invention (in contrast to earlier conditions that simply related to the availability of goods) and an acceptance that some patents might not be practiced both show a reluctance to engage in regulation of markets.

In addition, patent law is not protectionist in its formalities, and does not distinguish based upon the native country of applicants and companies or the

95. *See* Golden, *supra* note 93, at 2155 n.164; *see also* 35 U.S.C. § 271(d)(4) (2012) (stating that relief for infringement may not be denied for "refus[al] to license or use any rights to the patent").

96. Bonito Boats, Inc. v. Thunder Craft Boats, Inc., 489 U.S. 141, 151 (1989) (quoting United States v. Dubilier Condenser Corp., 289 U.S. 178, 186 (1933)).

97. In *Continental Paper Bag Co. v. Eastern Paper Bag Co.*, the Supreme Court held that a competitor might choose to practice one among a number of patents and that such a choice was a reasonable business decision that should not deprive the patent holder from excluding others from selling patented goods. 210 U.S. 405 (1908). While *eBay v. MercExchange*, 547 U.S. 388 (2006), limited the availability of injunctions for non-competitor firms, injunctions remain available for companies that do not practice their patents if they are able to show that the infringement will result in a loss of market share. *See* Rajec, *supra* note 30, at 751–58.

98. *See* Sykes, *supra* note 6, at 61.

99. Of course, this notion is subject to criticism: sometimes we may find that access is more important than uniform levels of encouragement to innovate and sometimes competitive firms may choose to repress new technologies so as not to cannibalize their own markets. This claim is therefore descriptive.

location of invention.[100] This was not always the case; however, the requirement of TRIPS that patents be available "without discrimination as to the place of invention, the field of technology and whether products are imported or locally produced," has resulted in the elimination of disparate treatment for foreign inventive activity or publications for the purposes of determining patentability.[101] The mercantilist notion of enticing innovators to one's shores to manufacture goods for one's population is anathema to the current system. There is nothing constitutive of patent law that requires protecting the national market for the patent holder's benefit when the patent holder does nothing *for* the domestic market in exchange. Moreover, this vestige of market protection will not, on its own, further the purposes of aiding domestic industry or encouraging the introduction of goods that would otherwise not be available.[102] From the domestic perspective, the benefits of the national exhaustion regime are available for all U.S. patent holders, regardless of whether they manufacture or market goods domestically,[103] and there is a consensus that the current rule results in higher prices for American consumers.[104] In addition, with greater harmonization among patent regimes worldwide, there should be less need to ensure patent holders of their rewards in the domestic market. The strong baseline level of intellectual property rights in WTO member countries allows patent holders to reap rewards in foreign markets that they could not have expected in the past. These gains were ostensibly for the sake of fewer barriers to trade. However, the control patent holders hold over trade in patented goods remains a barrier to trade.

---

100. Although under the Patent Act of 1793, U.S. patents were only available to citizens of the United States, Patent Act of 1793, ch. 11, § 1, 1 Stat. 318, 318–19, patents have been available to foreign inventors living elsewhere since 1836, Patent Act of 1836, ch. 357, sec. 8, 5 Stat. 117, 120–21.

101. TRIPS, *supra* note 53, art. 27(1); *see* U.S. Patent Statistics Chart Calendar Years 1963–2012, http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm (showing the radical increase in percentage of patents resulting from foreign filings from 18% in 1963 to 40% in 1982, and again to 52% in 2012); *see also* Hubbard, *supra* note 61, at 356–58.

102. *See* Hubbard, *supra* note 61, at 374 (suggesting that changes to U.S. patent laws are not useful tools for making American firms more competitive globally, in part because the laws are non-discriminatory to foreign inventors and companies filing U.S patent applications, as required by TRIPS).

103. The availability of an exclusion order at the ITC is dependent on a patent holder's ability to show injury to a domestic industry. However, that requirement is not as strong as it sounds, and a patent holder who could not meet that low bar would still have a case for infringement through importation in federal district court. *See* Colleen V. Chien, *Protecting Domestic Industries at the ITC*, 28 SANTA CLARA COMPUTER & HIGH TECH. L.J. 169 (2011).

104. *See, e.g.*, Abbott, *supra* note 51, at 18; Jeffery Atik & Hans Henrik Lidgard, *Embracing Price Discrimination: TRIPS and the Suppression of Parallel Trade in Pharmaceuticals*, 27 U. PA. J. INT'L ECON. L. 1043, 1044 (2006).

IV. DOCTRINAL DEVELOPMENTS

The arguments about undesirable downstream control of goods and the benefits of competition, on the one hand, and the ability and potential gains of dividing markets, on the other, are not new. They inform the common law development of the principle of national exhaustion in patent law and have reemerged in considerations of an international patent exhaustion regime.

A. NATIONAL PATENT EXHAUSTION: DOCTRINE AND THEORY

The evolution of the national exhaustion doctrine in U.S. law shows concerns for many of the same principles discussed so far. Early cases wrestled with the extent to which patent holders should be able to subdivide the national market for goods, weighed against the interests of consumers in using and reselling their goods as they wished. The case law also treats the issue in terms of the ability of licenses to constrain downstream behavior. The question of restrictive licensing is one that is becoming ever more important when many software-based goods are licensed rather than sold, and as technological blocks to resale are becoming more prevalent. In fact, it may be that the ability to distinguish licenses from sales is becoming ever more difficult and will only become more important if exhaustion is extended internationally.[105] However, even in the late nineteenth century courts were balancing freedom to contract with competitive marketplaces.

The Supreme Court case establishing patent exhaustion involved geographic limitations, albeit domestic limitations imposed by license. In *Adams v. Burke*, the Court reviewed a patent infringement claim arising from the use of patented coffin lids by an undertaker who bought the lids from a licensed manufacturer.[106] The manufacturer had been assigned all rights in the patent by the patentee, *within ten miles of Boston*.[107] The manufacturer sold a patented coffin lid to an undertaker within the prescribed area, but the lid was subsequently used farther away.[108] The Court held that the right to sell may indeed have been restricted by geographical area, but that the purchaser "acquired the right to [the] use of it freed from any claim of the patentee . . . ."[109] The Court based its holding on the single reward theory—

105. *See infra* Part V.

106. Adams v. Burke, 84 U.S. (17 Wall.) 453 (1873).

107. *Id.* at 456.

108. The use was in Natick, Massachusetts, a full seventeen miles from Boston. *Id.* at 454.

109. *Id.* at 457; *see id.* at 457–58 (Bradley, J., dissenting). Justice Bradley explained that the Patent Act of 1836 expressly authorized:

> not only an assignment of the whole patent, or any undivided part thereof, but a "grant and conveyance of the exclusive right under any patent, to make and use, and to grant to others to make and use the thing

the idea that once the patent holder has received her royalties for the use of the invention, "it is open to the use of the purchaser without further restriction on account of the monopoly of the patentee."[110] The opinion recognized both the interests of patent holders in dividing territory for the purposes of licensing and the interests of purchasers in using the goods they buy without further restrictions, considerations that inform the debate on international exhaustion of intellectual property now. In addition, the Court noted the single-use nature of the invention at issue.[111]

In these early cases, the exhaustion doctrine can be seen as balancing an interest in allowing patent holders contractual freedom to license their patent as they see fit with an interest in reducing complex limitations on downstream use.[112] Such restrictions were considered a restraint on competition, and the early exhaustion decisions "used the Patent Act to create limitations on both vertical and horizontal territorial restraints that the Supreme Court would apply under the antitrust laws as well."[113] Thus, early exhaustion cases—in similar fashion to their modern counterparts—frequently involved questions of the scope and ability of licenses to restrict downstream uses.[114] Occasionally, restrictions on use that are included in a

> patented within and throughout any specified part or portion of the United States."

Id. (emphasis omitted).

110. *Id.* at 456; *see also* United States v. Masonite Corp., 316 U.S. 265, 278 (1942) (patent exhaustion depends on "whether or not there has been such a disposition of the article that it may fairly be said that the patentee has received his reward for the use of the article"); Bloomer v. McQuewan, 55 U.S. (14 How.) 539, 553–54 (1852) (explaining that the congressional power to grant an extension of a patent does not imply "that Congress may, from time to time . . . reinvest in [an inventor] right of property which he had before conveyed for a valuable and fair consideration").

111. The Court limited its holding to "the class of machines or implements we have described" referring to "an instrument or product of patented manufacture which perishes in the first use of it, or which, by that first use, becomes incapable of further use, and of no further value." *Adams v. Burke*, 84 U.S. at 456–57.

112. *See* Christina Mulligan, *A Numerus Clausus Principle for Intellectual Property*, 80 TENN. L. REV. 235, 286–89 (2013) (suggesting that a *numerus clausus* principle in intellectual property would result in a first sale doctrine in digital works, and looking to the real property analog of the historic prohibition on servitudes in chattel for theoretical support); *see also* Andrew T. Dufresne, Note, *The Exhaustion Doctrine Revived? Assessing the Scope and Possible Effects of the Supreme Court's* Quanta *Decision*, 24 BERKELEY TECH. L.J. 11, 14–15 (2009) (explaining how the exhaustion doctrine exhibits "an aversion to personal property servitudes").

113. Herbert Hovenkamp, *Innovation and the Domain of Competition Policy*, 60 ALA. L. REV. 103, 110–11 (2008).

114. In *Bloomer v. McQuewan*, the Court addressed the issue of whether an assignment of rights, title, and interest in a patent made during the first term of the patent allowed the licensee to continue using a machine made pursuant to the license during the second term. The Court held that continued use of the patented machines did not constitute infringement.

first sale have been honored such that violations are treated as cases of patent infringement.[115] However, there is some uncertainty as to what characteristics define allowable licenses and what makes such a license unenforceable, either through contract or patent law.

In 2008, the Supreme Court decided *LG Electronics v. Quanta* and expanded the reach of the exhaustion doctrine again, restricting a patent holder's ability to limit downstream uses of patented technology through patent infringement suits.[116] The *Quanta* Court discussed a series of cases in which the Court had limited restrictive licenses that constrained purchasers' use of patented products, situating the exhaustion doctrine as a counterweight to attempted expansions of patent rights beyond their

55 U.S. (14 How.) 539 (1852). In *Mitchell v. Hawley*, the Court explained that the right to "mak[e] or vend[] the patented machine" was dependent "upon the nature of the conveyance," whereas someone who had gained complete title to a patented machine has full right to use the machine, which "ceases to be within the limits of the monopoly." 83 U.S. (16 Wall.) 544 (1872) (affirming a judgment of infringement because the patent holder had issued a license that restricted the licensee's right to make, use, and license others to use the machines to the original patent term, specifying that the licensee was not to license rights beyond that original term). The Court therefore held that when two downstream licensees continued to use their machines after the original patent term had ended—but the extended term had not—they were engaging in an unauthorized use and therefore infringing the patent.

115. In one notable case, the Federal Circuit ruled on whether a patent holder's "single use only" restriction on a patented medical device could limit use of the devices, such that a company infringed the patent when it reconditioned and sterilized the parts to make them suitable for reuse. Mallinckrodt, Inc. v. Medipart, Inc., 976 F.2d 700, 709 (Fed. Cir. 1992). The court relied on the idea that sales may be conditioned on particular uses if the license does not go beyond the patent grant "and into behavior having an anticompetitive effect not justifiable under the rule of reason." *Id.* at 708. Some suggest *Mallinckrodt* was likely overruled, sub silentio, by the Supreme Court's decision in *Quanta Computer, Inc. v. LG Elecs.*, 553 U.S. 617 (2008). *See* Hovenkamp, *supra* note 113, at 111 n.35 (stating that *Quanta* overruled *Mallinckrodt*); *see also* Static Control Components, Inc. v. Lexmark Int'l, Inc., 615 F. Supp. 2d 575, 585 (E.D. Ky. 2009) ("this Court is persuaded that *Quanta* overruled *Mallinckrodt sub silentio*"). However, it is possible that *Mallinckrodt* is distinguishable because it involved an initial sale of the patented article that was conditional on single use, whereas *Quanta* involved restrictions on a downstream sale. *See, e.g.*, Chiappetta, *supra* note 43, at 1113–15 (2011) (discussing the absence of *Mallinckrodt* from the Court's analysis and the potentially narrow scope of the Supreme Court's decision).

116. In *Quanta*, patent holder LG Electronics granted Intel a license to method patents practiced in Intel's chipsets, permitting Intel to manufacture, use, sell, and import its own products that practice the patents; but, under a separate agreement, Intel agreed to notify customers that while none of the products they purchased infringed patents, the license did not extend to combining Intel products with non-Intel products. Quanta purchased chipsets and combined them with other products in a computer, and LG Electronics sued. *Quanta*, 553 U.S. 617 (2008).

intended scope.[117] The Court then found that the patent license between the parties did not restrict what Intel could sell, and that therefore Intel had sold the chipsets to Quanta with no restrictions and thus exhausted LGE's patent rights.[118]

These cases leave the impression that exhaustion was developed to target a type of patent misuse through limiting restrictions on downstream prices and uses of goods. What emerges is a general rule granting deference to the terms of a license between a patent holder and manufacturer or retailer, limited by an abiding suspicion toward restrictions that accompany patented products beyond the privity of the initial contracting parties and thereby attempt to cabin the behavior of downstream consumers or resellers. The application of the doctrine of exhaustion to restrictive licenses thus targets anti-competitive behaviors.[119]

Courts have been more lenient toward restrictions that travel with goods when those goods are long lasting and easy to replicate perfectly. One area where exhaustion is consistently revisited is in regards to self-replicating technology.[120] The Court in *Adams v. Burke* relied in part on the single-use nature of the patented coffin lids,[121] a characteristic that ensures the single reward limitation does not deprive the patent holder of rewards from multiple future uses of a good. Although someone who purchased a machine could use it as he pleased without being subject to claims of infringement,

117. *Id.* at 625–26 (citing Bauer & Cie v. O'Donnell, 229 U.S. 1, 14–17 (1913) (holding that the right to sell exclusively secured in the patent statute does not include the right to downstream price-fixing that "prevents competition by notices restricting the price at which the article may be resold"); Motion Picture Patents Co. v. Universal Film Mfg. Co., 243 U.S. 502, 518 (1917) (disallowing a claim of patent infringement against downstream users who disobeyed a notice affixed to the patented projecting-kinetoscopes they had purchased and used film produced by another manufacturer in the machines); United States v. Univis Lens Co., 316 U.S. 241 (barring, under a doctrine of exhaustion, price-fixing by a patent holder who sold unfinished lenses for glasses to licensed wholesalers and retailers who ground the lenses into patented products—with the requirement that they be sold to consumers at a fixed rate).

118. *Id.* at 636–37; *see also* F. Scott Kieff, Quanta v. LG Elecs.*: Frustrating Patent Deals by Taking Contracting Options Off the Table?*, 2008 CATO SUP. CT. REV. 315, 316 (2008) (criticizing *Quanta* for its likely restrictions on contracting freedom, arguing that the case "may greatly frustrate the ability of commercial parties to strike deals over patents").

119. Depending on the observer, patent misuse is either coextensive with—or consumed by—antitrust doctrine. *See, e.g.*, USM Corp. v. SPS Tech., Inc. 694 F.2d 505, 512 (7th Cir. 1982) (Posner, J.) (suggesting that antitrust "consumes" the misuse doctrine); HERBERT HOVENKAMP ET AL., IP AND ANTITRUST, 3–12 (supp. 2012) ("Generally speaking, patent misuse doctrine is largely coextensive with antitrust doctrine.").

120. *See, e.g.*, Bowman v. Monsanto, 133 S. Ct. 1761 (2013); Capitol Records, LLC v. ReDigi Inc., 934 F. Supp. 2d 640 (S.D.N.Y. 2013); discussion *supra* note 115.

121. Adams v. Burke, 84 U.S. (17 Wall.) 453, 456–57 (1873).

the purchase did not entitle him to manufacture, use, or sell replicas of that machine. And yet, some technologies replicate themselves by force of nature or design. Most recently, the Supreme Court was faced with a patent infringement claim brought by Monsanto, which holds patents on soybean seeds, genetically modified to withstand being sprayed with the herbicide Roundup.[122] The Court held that the farmer who bought patented seeds from a grain elevator and, without permission, planted the seeds and harvested the newly-grown seed crop had infringed Monsanto's patents, based in part on Monsanto's restrictive licensing agreement.[123] When Monsanto sued Bowman, Bowman claimed that the authorized sale of seeds to the grain elevator exhausted Monsanto's patent rights, such that the company could not claim infringement based on how a subsequent purchaser used the seeds.[124] The Supreme Court held that while Bowman could have consumed the seed, fed it to animals, or sold it to others, his use of the patented seeds to "make" new, patented seeds was infringement.[125]

In *Bowman v. Monsanto*, the Court recognized that while the exhaustion doctrine is meant to limit the patent holder to a single reward on each patented object, it should not be used to limit the patent term to "only one transaction" by allowing a purchaser to make unlimited copies of the patented good.[126] In so doing, the Court had to deal with the strict license that accompanied sales of the patented seed. It may be that courts have a higher tolerance for restrictive licenses over technologies amenable to multiple uses and easy replication.

---

122. *Bowman*, 133 S. Ct. 1761.

123. *Id.* Monsanto protects its patented seeds through a licensing agreement that allows a grower to plant the seeds in a single season. The resulting crop may be harvested and sold, and is often sold to a grain elevator, which consolidates crops from multiple farmers and sells the seeds for human or animal consumption. Monsanto's licensing agreement forbids the grower from saving harvested soybeans to replant or resell for planting. Bowman did not save harvested seeds to plant, but rather bought seeds from the grain elevator that were meant for human or animal consumption (and which therefore did not require a license from purchasers), planted them, sprayed them with Roundup, and then harvested and replanted the resulting seeds in subsequent years. *Id.* at 1764–65.

124. *Id.* at 1765.

125. *Id.* at 1766–67. The Court explained that planting the seeds and growing a crop from them was likely always infringement, but that the license granted by Monsanto (allowing farmers to plant the seeds in one season only) rendered that use noninfringing. *Id.* at 1767 n.3. Thus, although the purpose of the patented inventions was to create seeds that grow well, the initial sale of seeds did not confer the right to plant them without a license. The Court did note that, in the instance of the sale by Monsanto or its authorized seed sellers, the right to plant the seeds once might also be inferred if there had not been a license agreement, presumably because that was what the seeds were created for and because the characteristic that warranted the patent could only be realized through replication. *Id.*

126. *Id.* at 1768.

Behind the *Bowman* opinion was also recognition of another area of potentially self-replicating technology that may try the exhaustion doctrine—domestically and internationally—namely, software. Software provides an interesting example because it can be protected by both copyright and patent, and because strict definitions of copying or making under the respective acts do not neatly apply to software.[127] Thus, the Copyright Act contains a provision limiting the exclusive reproduction right by allowing the owner of a copy of a computer program to make a copy of the software in order to use it and allowing the transfer of such copies (when coupled with the transfer of related rights).[128] A provision that would expand this limitation and create a first sale doctrine for digital material was proposed but has not been adopted.[129] Recently, in *Capitol Records, LLC v. ReDigi Inc.*, a court found there was no such right under the current statute.[130] Software is also of interest because it has gravitated toward a model of complex licensing as opposed to sale. This can be seen as a direct result of its place on the spectrum from single-use to self-replicating technology. Single-use technologies may be easier to rule with an exhaustion regime because an item may be sold and resold, but once it is used it will not be used again. As a result, when the rights holder exhausts her rights in the patent by selling an article embodying it, she need not track its whereabouts or worry about infringement. In contrast, self-replicating technologies and digital technologies that allow for the creation of exact copies are particularly susceptible to infringement. Thus, for single-use articles, limiting a rights holder to a single reward does not put in jeopardy her ability to reap subsequent rewards from other sales of a good, whereas other technologies lend themselves to infringement, thus replacing some of the demand for the product with infringing copies and threatening

127. The eligibility of software for multiple types of intellectual property protection is not unique to the technology. It does, however, point to the desirability of theoretical and doctrinal consistency between the areas in regards to international exhaustion, inter alia.

128. 17 U.S.C. § 117 (2012). This provision allows those who buy software to install it on their computer hard drives and make backup copies—both of which constitute copying but are central to the industry. For example, copyright owners fully expect—and desire—their customers to install software on their computers, much as Monsanto fully expects farmers to plant their seeds, though such planting results in making a copy of protected seeds.

129. Proponents of this exception suggested that making a copy of material should not be considered an infringement if that material was simultaneously deleted from the machine on which it was lawfully placed. NAT'L TELECOMM. & INFO. ADMIN., U.S. DEP'T OF COMMERCE, REPORT TO CONGRESS: STUDY EXAMINING 17 U.S.C. SECTIONS 109 AND 117 PURSUANT TO SECTION 104 OF THE DIGITAL MILLENNIUM COPYRIGHT ACT (2001), *available at* http://www.ntia.doc.gov/report/2001/report-congress-study-examining-17-usc-sections-109-and-117-pursuant-section-104-digital.

130. Capitol Records, LLC v. ReDigi Inc., 934 F. Supp. 2d 640 (S.D.N.Y. 2013).

the rights holder's ability to reap rewards from other sales. All of these factors result in greater interest in licensing for easily replicable and self-replicating technologies.

Alas, complex licensing schemes may also be used by rights holders to complicate and restrict future sales of a product, even when such sales are not attempts to engage in replication or infringement. The exhaustion doctrine in domestic law must balance allowing patent holders to protect their patented goods through reasonable licensing practices with limiting restrictions on downstream use that the patent holder should have no reasonable expectation of curtailing and that may be beneficial to society at large, even if undesirable to the patent holder. This is not an easy balance. However, it is one undertaken and constantly refined by the courts in their application of the exhaustion doctrine in the domestic context. National exhaustion laws are applied in ways that allow greater licensing restrictions for technologies susceptible to self-replication or easy and exact replication by others. Nonetheless, overaggressive use of such licensing may backfire and result in findings that what the parties took to be a license was actually a sale.[131] These same types of consideration apply equally in the international context and will merit more scrutiny if international exhaustion is adopted.

B. A DIFFERENT RULE FOR INTERNATIONAL PATENT EXHAUSTION

The Supreme Court has not recently weighed in on the question of international exhaustion of patents, declining opportunities in 2002 and 2013 to hear cases that presented the issue.[132] Older case law indicates that there

131. Shubha Ghosh explains:

> Sales trigger exhaustion while licenses or leases do not. Through these mechanisms, the parties can structure the transaction to suit their particular needs, and the rights holder can price accordingly. The difficulty is deriving workable criteria to distinguish between a sale and a lease. Such criteria can be difficult given the complexities of actual transactions that will contain myriad terms dealing with the allocation of various risks under different contingencies. Nonetheless, in principle, the exhaustion doctrine can deal with consumer heterogeneity through threshold rules of applicability based on the transaction.

SHUBHA GHOSH, INT'L CTR. FOR TRADE & SUSTAINABLE DEV. 8 (2013), *available at* http://ictsd.org/i/publications/181092; *see also* Brian W. Carver, *Why License Agreements Do Not Control Copy Ownership: First Sales and Essential Copies*, 25 BERKELEY TECH. L. J. 1887 (2010) (in the copyright context); Timothy D. Greene, *"All Substantial Rights": Toward Sensible Patent Licensee Standing*, 22 FED. CIR. B.J. 1 (2012) (in the context of standing).

132. Jazz Photo Corp. v. Int'l Trade Comm'n, 536 U.S. 950 (2002), *denying cert. to* 264 F.3d 1094 (Fed. Cir. 2001); Ninestar Tech. Co. v. Int'l Trade Comm'n, 133 S. Ct. 1656 (2013), *denying cert. to* 667 F.3d 1373 (Fed. Cir. 2012). This does not necessarily mean that the Supreme Court will not take a case on the issue in the future; these particular cases may have

was a limited rule of exhaustion when the owner of patents in multiple countries sold a patented product without restriction. In 1890, the Supreme Court ruled in *Boesch v. Graff* that U.S. patent rights were not exhausted by lawful manufacture and sale in Germany.[133] Although the sale was indeed lawful, it was not authorized by the patent holders, who held patents in both the United States and Germany.[134] Rather, the manufacturer was entitled to manufacture and sell the products under a type of prior user right in the German patent law, which allowed those who were preparing to produce a patented article at the time the patent was filed to do so without authorization.[135]

Following *Boesch*, the potential for patent exhaustion through a first, authorized sale abroad appeared to depend on whether the entity that authorized the foreign sale was the entity trying to enforce the U.S. patent. In *Dickerson v. Matheson*, the Second Circuit Court of Appeals explained that if a foreign purchase was from the owner (or licensee) of *both* the foreign and U.S. patents, then U.S. patent rights were exhausted, but held that a purchase from a licensee of *only* the foreign patent did not exhaust U.S. patent rights.[136]

---

been poor vehicles for the issue. However, it appears likely that the Court was considering whether the exhaustion question must be answered for both the patent and copyright law simultaneously as it declined review of *Ninestar* the same day that it granted, vacated, and remanded two copyright cases in light of *Kirtsaeng*.

133. Boesch v. Graff, 133 U.S. 697, 703 (1890). The *Boesch* Court determined:

> The right which Hecht had to make and sell the burners in Germany was allowed him under the laws of that country, and purchasers from him could not be thereby authorized to sell the articles in the United States in defiance of the rights or patentees under a United States patent. . . . The sale of articles in the United States under a United States patent cannot be controlled by foreign laws.

Id.

134. *Id.*

135. *Id.* at 701. It is an open question whether the Court would have come to the same conclusion had the manufacturer operated with authorization from the holder of the patents (both U.S. and German).

136. Dickerson v. Matheson, 57 F. 524, 527 (2d Cir. 1893). The court explained:

> A purchaser in a foreign country, of an article patented in that country and also in the United States, from the owner of each patent, or from a licensee under each patent, who purchases without any restrictions upon the extent of his use or power of sale, acquires an unrestricted ownership in the article, and can use or sell it in this country.

Id. This is consistent with the pre-*Boesch* case of *Holiday v. Mattheson*, in which a U.S. patent holder sold patented goods in England with no restrictions or conditions. Holiday v. Mattheson, 24 F. 185, 185 (C.C.S.D.N.Y. 1885). The court refused to enjoin a downstream purchaser from reselling the goods in the United States. *Id.* ("[w]hen the owner sells an article without any reservation respecting its use, or the title which is to pass, the purchaser acquires the whole right of the vendor in the thing sold"). *Cf.* Featherstone v. Ormonde

The 1909 case *Daimler Manufacturing. Co. v. Conklin* also granted an injunction to an exclusive licensee of a U.S. patent against someone who purchased a car—with patented components—while in Germany, from the company authorized to sell the patented goods there, and later imported it to the United States for personal use.[137] Soon after, the court in *Curtiss Aeroplane & Motor Corp. v. United Aircraft Engineering Corp.* found international exhaustion and distinguished *Daimler*, basing its decision on the patent holder's ownership of all relevant rights.[138] The court found that the holder of patents in both the United States and Canada exhausted its rights under both patents through its wartime sale of airplanes and manufacturing licenses granted to the British government, although those licenses only referred to the Canadian patents, finding the patent holder had not placed any restrictions on the sale or licenses.[139] Thus, there was no infringement by the company that purchased the airplanes from the British government following the war and made offers to sell them in the United States.[140] Ultimately, the impression given by *Boesch* and the cases decided soon after it is that the owner of patents in multiple countries exhausts rights in them all through his first, unrestricted sale.[141] The cases rely on the agency of the patent owner in

Cycle Co., 53 F. 110, 111–12 (C.C.S.D.N.Y. 1892) (granting to an assignee of U.S. patent rights an injunction against licensees of British patent who had rights to use the patented tires in Britain but subsequently tried to sell them in the U.S., with the court explaining that "the purchaser does not acquire any right greater than those possessed by the owner of the patent").

137. Daimler Mfg. Co. v. Conklin, 170 F. 70, 72 (2d Cir. 1909) (explaining that "[t]he sale by a German patentee of a patented article may take it out of the monopoly of the German patent," then asking "but how can it take it out of the monopoly of the American patentee who has not sold?"); *see also* Daimler Mfg. Co. v. Conklin, 160 F. 679, 681 (C.C.S.D.N.Y. 1908) (detailing terms of exclusive license for import and sale of cars with patented components).

138. Curtiss Aeroplane & Motor Corp. v. United Aircraft Eng'g Corp., 266 F. 71 (2d Cir. 1920).

139. *Id.* at 79 ("It is admitted that, if the aeroplanes which are alleged to infringe had been built in Canada under a limited license, or under a Canadian patent, and then brought into the United States, infringement would have been made out. But that is not this case.").

140. *Id.*

141. *See id.* at 79. The court approvingly cites a British judge's recitation of the rule, the cited case found no exhaustion:

> When an article is sold without any restriction on the buyer, whether it is manufactured under either one or the other patent, that, in my opinion, as against the vendor gives the purchaser an absolute right to deal with that which he so buys in any way he thinks fit, and of course that includes selling in any country where there is a patent in the possession of and owned by the vendor.

Id.

deciding to sell the patented product and on the reward he is due to collect for it—but only once.[142]

In *Sanofi, S.A. v. Med-Tech Veterinarian Prods, Inc.*, a district court reached different conclusions on exhaustion for patent holder Sanofi and its licensee American Home Products, which had exclusive U.S. rights to sale.[143] Because Sanofi's subsidiary had sold veterinary drugs in France—without written restrictions on future United States sales—the court found that Sanofi had exhausted its rights in the patent and could not obtain an injunction against later importation and sale.[144] However, the court looked to the terms of American Home Products' exclusive license to determine what rights Sanofi had maintained. The court concluded that Sanofi could not sell (in France) what it did not have (the right to sell in the United States), and therefore allowed American Home Products to obtain an injunction.[145]

The Court of Appeals for the Federal Circuit ("Federal Circuit")[146] did not wrestle with these interpretations of *Boesch* in its 2001 decision, *Jazz Photo Corp. v. International Trade Commission* (*Jazz Photo I*). Instead, it held more broadly that there was no rule of international exhaustion of patents,[147] and in later cases strengthened the rule in contradiction to the early interpretations offered by the Second Circuit. In *Jazz Photo I*, the Federal Circuit reviewed a proceeding at the International Trade Commission that

142. Ultimately, limiting international exhaustion to cases of multiple patent ownership has the potential to cause much mischief. Ignoring the incentive for companies to set up complicated subsidiary and affiliated entity-structures (and the difficulties for courts of unraveling those that already exist), an inventor who sells all her rights to the patent in a foreign country presumably is being "rewarded" for all future sales under the patent. The value she puts on this is embodied in the sale price and will depend on whether there is international exhaustion or not. For that reason, an international exhaustion regime should be based on authorization from the patent holder, but authorization to sell in a particular market should be sufficient; there is no reason to require that the patent holder place the particular good on the market herself.

143. Sanofi, S.A. v. Med-Tech Veterinarian Products, Inc., 565 F. Supp. 931 (D.N.J. 1983).

144. *Id.* at 938.

145. *Id.* at 939–40 (holding that patent rights are only exhausted "where the sale is one which the seller had the authority to make in this country").

146. The Court of Appeals for the Federal Circuit has jurisdiction over appeals in cases involving patents, including cases brought at the International Trade Commission under 19 U.S.C. § 337. The Federal Circuit also has jurisdiction over other trade-related cases appealed from the United States Court of International Trade, however the § 337 cases have provided the vehicle for rulings on international patent exhaustion. *See* 28 U.S.C. § 1295(a)(5)–(6) (2012).

147. Jazz Photo Corp. v. Int'l Trade Comm'n (*Jazz Photo I*), 264 F.3d 1094, 1105 (Fed. Cir. 2001) ("United States patent rights are not exhausted by products of foreign provenance. To invoke the protection of the first sale doctrine, the authorized first sale must have occurred under the United States patent." (citing Boesch v. Graff, 133 U.S. 697, 701–03 (1890))).

resulted in an exclusion order on disposable cameras covered by patents held by Fuji Photo Film Co.[148] A number of firms, primarily in China, bought used disposable camera cartridges from film developers in many countries and reloaded them. Fuji Photo Film brought the case against importers who purchased and imported the reloaded cameras for sale in the United States.[149] The opinion from the ITC focused on whether reloading the cameras constituted noninfringing "repair" of a patented product or infringing "reconstruction."[150] Repair of a patented article is considered "use," which cannot be controlled following an authorized sale, whereas reconstruction, which "requires a more extensive rebuilding of the patented entity," is considered "making" a new, patented article.[151] The Federal Circuit reversed the ITC's holding that reloading the cameras was reconstruction, and found instead that it constituted repair.[152] Because the initial camera sales had been authorized, that might have been the end of it. However, the court then went on to hold that because many of the initial, authorized sales had been in foreign countries, they did not exhaust rights in the U.S. patents; as a result, the repaired cameras were excludable and only cameras that had been the subject of an authorized first sale in the United States were noninfringing.[153] The *Jazz Photo I* case has been criticized for imposing an inefficient "default" rule on sales of patented products,[154] for its potential to seriously restrict traditional forms of commerce and behavior,[155] and for deciding the issue without briefing, *sua sponte*.[156] There was soon an opportunity for fully

148. *Id.* at 1099.

149. It is interesting that this case, like *Adams* and *Mallinckrodt*, involved sales of items intended for single use. The court distinguished *Mallinckrodt* with the finding that the statements on each camera cautioning consumers not to remove film and to return the camera to the photoprocessor did not constitute a limiting license on their use. *Id.* at 1107.

150. *Id.* at 1101.

151. *Id.* at 1104.

152. *Id.* at 1107.

153. *Id.* at 1105.

154. *See* Chiappetta, *supra* note 43, at 1122.

155. Petition for a Writ of Certiorari at 11–12, Jazz Photo v. Int'l Trade Comm'n, 536 U.S. 950 (2002) (No. 01-1158), 2002 WL 32134396 (suggesting that, among other problems, "the owner of a patented property item purchased abroad (such as a camera, watch, or car) can no longer use it in the United States without infringing the patent. No firm may repair that item to prolong the owner's use without risk of liability for contributory infringement.").

156. *Id.*; *see also* Adam Mossoff, *Commercializing Property Rights in Inventions: Lessons for Modern Patent Theory from Classic Patent Doctrine*, *in* COMPETITION POLICY AND PATENT LAW UNDER UNCERTAINTY: REGULATING INNOVATION 348–49 (Geoffrey A. Manne & Joshua D. Wright eds., 2011).

briefing international exhaustion, when Jazz Photo appealed a district court judgment against it.[157]

In *Fuji Photo Film Co., v. Jazz Photo Corp.* (*Jazz Photo II*), the Federal Circuit again refused to apply exhaustion to authorized foreign sales, holding that "[t]he patentee's authorization of an international first sale does not affect exhaustion of that patentee's right in the United States."[158] The court's reasoning was that the foreign sales did not occur "under" a U.S. patent and that a contrary ruling would contravene the rule against the extraterritorial application of U.S. law.[159] After the Supreme Court's decision in *Quanta*, one district court held that *Jazz Photo I & II* were no longer good law, as its theory was undercut by the Supreme Court's ruling that blocked a patent holder from authorizing a sale but suing a downstream purchaser for infringement.[160] Of course, the *Quanta* Court had merely voiced the same concerns that are present in all domestic exhaustion cases; restating it did not answer the question of its applicability to foreign transactions. Another district court case that attempted to cabin *Jazz Photo I & II* held that where the parties had negotiated a worldwide license, exhaustion might occur even if the first sale was made in a foreign country.[161] However, the strong rule against exhaustion from *Jazz Photo I & II* has mostly held, although in complex technology areas with multiple assembly steps the question of whether a sale is "under" a U.S. patent is not entirely straightforward.[162] This

157. Fuji Photo Film Co., Ltd. v. Jazz Photo Corp. (*Jazz Photo II*), 394 F.3d 1368, 1370 (Fed. Cir. 2005) (beginning description of the litigation history by noting that "Fuji and Jazz are no strangers to this court").

158. *Id.* at 1376.

159. *Id.*; *see also* John A. Rothchild, *Exhausting Extraterritoriality*, 51 SANTA CLARA L. REV. 1187, 1211–12 (2011) (arguing that an international exhaustion rule would not constitute extraterritorial application of U.S. law). This contrasts with the requirement discussed in *Kirtsaeng* that a work be made "under this title" to qualify for exhaustion. Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351, 1355 (2013). Under current patent law, an authorized U.S. sale of a patented good made abroad would exhaust further rights.

160. LG Electronics, Inc. v. Hitachi, Ltd., 655 F. Supp. 2d 1036, 1046–47 (N.D. Cal. 2009).

161. STMicroelectronics, Inc. v. Sandisk Corp., No. 4:05CV45, 2007 WL 951655, at *3 (E.D. Tex. Mar. 26, 2007). The court stated:

> STM gave Toshiba a license in all types of patents with respect to the licensed products in all countries of the world. All the countries of the world includes the United States of America. Therefore, Toshiba (or its subsidiaries) had the right to sale any of the licensed products under the United States Patents '626 and '184 in the United States or anywhere in the world.

Id. at *3.

162. *See, e.g.*, Cornell Research Found., Inc. v. Hewlett–Packard Co., No. 5:01-CV-1974 (NAM/DEP), 2007 WL 4349135, *50–52 (N.D.N.Y. Jan. 31, 2007) (denying summary

complexity highlights the potential savings for companies engaged in complex technologies that would come from a default rule of international exhaustion. In *Fujifilm Corp. v. Benun*, the Federal Circuit held that *Quanta*, which did not involve foreign sales, did not impact the rule against international exhaustion.[163] And although the Supreme Court had the opportunity to address international patent exhaustion in an appeal from *Ninestar* recently, it declined.[164]

C. THE UNITED STATES POSITION IN INTERNATIONAL NEGOTIATIONS

Despite arguments that the Federal Circuit rule is overly restrictive, its approach appears to be consistent with the United States' policy in its trade discussions. These policies have been consistently expressed in negotiations for patent law harmonization across various platforms. However, international exhaustion is a rare example of an area of intellectual property law specifically and consistently left out of the extensive multilateral treaties that treat minimum requirements of patent protection.[165] The TRIPS Agreement represents an enormous step toward harmonized patent law.[166] It built upon previous agreements that primarily implemented procedural harmonization, adding substantive harmonization through requirements of minimum levels of protection.[167] In addition to substantive measures, the TRIPS Agreement has enforcement mechanisms that previous treaties did not. Disputes among member countries over TRIPS violations are resolved

judgment on the issue of where the sales occurred for purposes of an exhaustion determination, following an exhaustive discussion of the evidence); *see also* Laserdynamics, Inc. v. Quanta Storage America, Inc., No. 2:06-CV-348-TJW-CE, 2009 WL 3763444, at *1 (E.D. Tex. June 29, 2009); Minebea Co., Ltd. v. Papst, 444 F. Supp. 2d 68, 140–41 (D.D.C. 2006); Minebea Co., Ltd. v. Papst, 374 F. Supp. 2d 202, 215–16 (D.D.C. 2005).

163. Fujifilm Corp. v. Benun, 605 F.3d 1366, 1371 (Fed. Cir. 2010).

164. Ninestar Tech. Co. v. Int'l Trade Comm'n, 133 S. Ct. 1656 (2013), *denying cert. to* 667 F.3d 1373 (Fed. Cir. 2012). The case was the result of an enforcement action brought against the wholly owned subsidiaries of a Chinese producer of ink printer cartridges that had continued to import and sell cartridges subject to exclusion and cease and desist orders in an earlier proceeding. *Ninestar*, 667 F.3d 1373. The Federal Circuit did not take the opportunity to revisit its international exhaustion jurisprudence, likely because the case was an appeal of penalties assessed for "deliberately and in bad faith" violating an earlier exclusion order. *Id.* at 1378–79. For procedural and prudential reasons, then, this may not have presented an attractive case for the Supreme Court to use in addressing the question of international patent exhaustion.

165. *See* TRIPS, *supra* note 53, art. 6.

166. Margaret Chon, *Intellectual Property and the Development Divide*, 27 CARDOZO L. REV. 2816 (1999) (describing the TRIPS Agreement as effecting a "tectonic shift in the landscape of intellectual property law").

167. Requirements address patent-eligible subject matter, standards of patentability, and the duration and scope of rights. Rajec, *supra* note 29, at 16–17.

by the Dispute Settlement Body of the WTO.[168] Where violations are found, the offending member country may find itself facing serious trade sanctions.[169]

Although TRIPS has resulted in significant movement toward harmonized global patent law, the agreement has some flexibility for countries to craft their own laws. And, because patent law is territorial, there are also de facto variations in how it is applied. One subject countries could not come to an agreement about was exhaustion of intellectual property rights. Despite reaching agreements on many hotly debated topics, countries were unwilling to compromise.[170] As a result, TRIPS explicitly states that "nothing in this Agreement shall be used to address the issue of the exhaustion of intellectual property rights."[171]

As further rounds of multilateral trade agreements have faltered, progress in trade agreements has been made through bilateral or regional trade agreements. The United States has pushed for and often included "TRIPS-plus" measures in such agreements—patent protection measures that go beyond the minimum requirements laid out in the TRIPS Agreement. Some of these measures address exhaustion. Thus, for example, the U.S.-Morocco free trade agreement requires that the exclusive rights to prevent importation "shall not be limited by the sale or distribution of that product outside its territory,"[172] thus requiring members not to implement international

168. Understanding on Rules and Procedures Governing the Settlement of Disputes, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 2, 1869 U.N.T.S. 401.

169. The area in which the sanction is imposed need not be related to the violation. *Id.* art. 22.3.

170. The explicit dodge of the international exhaustion issue reflects the disagreement among countries on this issue. Australia, for example, was in favor of an international exhaustion regime—at least for copyrights—arguing that domestic consumers would benefit from it. DANIEL GERVAIS, THE TRIPS AGREEMENT: DRAFTING HISTORY AND ANALYSIS 62–63 (1998) (citing Memorandum of the WIPO Bureau for the Committee of Experts on a Possible Protocol to the Berne Convention, WIPO Doc. BCP/CE/IV/2 (Mar. 15, 1994)).

171. TRIPS, *supra* note 53, art. 6. This exclusion is only subject to the articles requiring national treatment and most-favored-nation treatment. *Id.* arts. 3–4. National treatment is the requirement of treating nationals of other member countries as well as one's own members, while most-favored nation treatment requires treating nationals of any other member country as well as nationals of other member countries. Thus, while member states may maintain their rules on exhaustion, those rules must not be applied in a manner that is discriminatory to nationals of other member states.

172. United States-Morocco Free Trade Agreement, U.S.-Morocco, art 15.9(4), June 15, 2004, http://www.ustr.gov/trade-agreements/free-trade-agreements/morocco-fta/final-text. The U.S.-Australia Free Trade Agreement similarly precludes a regime of international exhaustion, "at least where the patentee has placed restrictions on importation by contract or other means." United States-Australia Free Trade Agreement, U.S.-Austl., art. 17.9(4), Jan. 1,

exhaustion. Thus, it appears that the understanding of the executive branch has been that there is currently no international exhaustion in patent or copyright law.[173] In addition, the United States has been in negotiations for an Asia-Pacific trade agreement, known as the Trans-Pacific Partnership Agreement ("TPP").[174] Although negotiations have primarily been conducted in private, leaked draft versions indicate that the United States has opposed a rule of international exhaustion.[175]

V. IMPLEMENTATION AND CRITIQUES

An international exhaustion rule for patents might have seemed highly unlikely in the years since *Jazz Photo I & II* were decided—and particularly in the context of the United States' position in trade negotiations. However, the Supreme Court's recent decision in *Kirtsaeng* signals a willingness to reconsider exhaustion in the copyright context, supported by policy considerations that apply *mutatis mutandis* to patent law.[176] Indeed, most of the reasoning is consistent with the justification for the rule of national patent exhaustion[177] and, given the global nature of commerce and consumption, their application to a global market provides a natural next step. However, practical considerations about implementation remain. In addition, international patent exhaustion may present particular concerns for the pharmaceutical industry and make more salient concerns already raised by national rules of exhaustion, such as the appropriate treatment of restrictive licensing and the need—or potential—for differential treatment for single-use versus self-replicating technologies.

2005, *available at* http://www.ustr.gov/sites/default/files/uploads/agreements/fta/australia/asset_upload_file148_5168.pdf.

173. Justice Ginsburg mentions this in her *Kirtsaeng* dissent. Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351, 1373 (2013) (Ginsburg, J., dissenting) ("The Court's bold departure from Congress' design is all the more stunning, for it places the United States at the vanguard of the movement for 'international exhaustion' of copyrights—a movement the United States has steadfastly resisted on the world stage."); *see also infra* Sections V.A–V.B.

174. Sean M. Flynn et al., *The U.S. Proposal for an Intellectual Property Chapter in the Trans-Pacific Partnership Agreement*, 28 AM. U. INT'L L. REV. 105 (2012).

175. *See* Trans-Pacific Partnership art. QQ.A.12 (draft Aug. 30, 2013), *available at* http://wikileaks.org/tpp (noting that the United States, Australia, Japan, and Mexico oppose language suggesting that "[t]he Parties are encouraged to establish international exhaustion of rights."); *see also* Flynn et al., *supra* note 174, at 130–31 (2012) (discussing the proposed provision).

176. Although the Court's statutory construction is critiqued below, it is not applicable to the patent law question. *See infra* note 189 and accompanying text.

177. *See supra* Section IV.A.

A. *KIRTSAENG* AS A MODEL: COPYRIGHT'S FIRST SALE DOCTRINE GOES GLOBAL

The U.S. position against international exhaustion for patents is called into question by the Supreme Court's recent holding in *Kirtsaeng v. John Wiley & Sons, Inc.* that the first sale provision of the Copyright Act has no geographical limitations—in other words, U.S. copyright is governed by international exhaustion.[178] Although the separate statutory and common law development of patent and copyright law would allow for different rules, the theory underlying arguments for and against international exhaustion is the same for both fields.[179] In addition, many goods are covered by multiple types of intellectual property rights, so that absent adoption of international exhaustion for patents, the purposes of the copyright rule would be thwarted when applied to products also covered by patents. For these reasons, if international exhaustion in copyright law is defensible, it should be considered in patent law, too. The *Kirtsaeng* decision is a good place to start.

In *Kirtsaeng*, the Court ruled that importation and sale of books purchased from a Thai subsidiary of a U.S. publisher did not constitute copyright infringement because the original purchase exhausted the U.S. copyright holder's rights to exclude future sales.[180] Section 109 of the Copyright Act sets forth the rule on exhaustion, also known as the "first sale" doctrine, as an exception to the exclusive distribution rights of copyright owners, and provides that "the owner of a particular copy . . . *lawfully made under this title* . . . is entitled, without the authority of the copyright owner, to sell or otherwise dispose of . . . that copy."[181] Section 109 thereby codifies the first

178. *See Kirtsaeng*, 133 S. Ct. at 1371.

179. The story is different in trademark law, which has different theoretical and statutory underpinnings. Trademark law is generally understood to follow a rule of international exhaustion, with a widely applied exception for when "genuine, but unauthorized, imports differ materially from authentic goods authorized for sale . . . because a difference in products bearing the same name confuses consumers and impinges on the local trademark holder's goodwill." Societe des Produits Nestle v. Casa Helvetia, 982 F.2d 633, 635 (1st Cir. 1992). Courts have generally emphasized the differences between imported trademarked goods and authorized goods, thus allowing for de facto geographic price discrimination through trademark law. Thus, without changes to that area of law, consistent with *Kirtsaeng* and potential changes in patent law doctrine, a regime of international exhaustion in intellectual property could be undermined. *See generally* Charles E. Colman, *Post*-Kirtsaeng, *'Material Differences' Between Copyright and Trademark Law's Treatment of Gray Goods Persist* (New York University School of Law Public Law Research, Paper No. 13-40, 2013), *available at* http://ssrn.com/abstract_id=2281562.

180. *Kirtsaeng*, 133 S. Ct. at 1355–56.

181. 17 U.S.C. § 109(a) (2012) (emphasis added).

sale doctrine that had previously emerged in common law.[182] A separate provision, § 602, explains that importation of copies acquired abroad "without the authority of the owner of a copyright under this title . . . is an infringement of the exclusive right to distribute copies . . . under [§] 106."[183] Because the first sale provision is an exception to the exclusive right referenced in the importation provision, the Court has held that the first sale doctrine also limits findings of infringement through importation.[184] In *Quality King Distributors v. L'anza Research Int'l*, the Court held that importation of an authorized copy made in the United States, sent abroad, and sold before reimportation did not violate the importation provision.[185] That ruling did not address the issue of copies made abroad.[186] In particular, the first sale doctrine's requirement that a copy be "lawfully made under this title" was clearly satisfied by the authorized copy made in the United States. The *Kirtsaeng* Court had to decide whether authorized copies made abroad were "lawfully made under this title," and thus subject to the first sale doctrine.[187] The Court framed the question as whether there was a geographical component to the "lawfully made" requirement, and determined there was not.[188]

The decision can be understood in part as choosing between two statutory interpretations, neither of which was likely contemplated by the drafters.[189] Either the first sale doctrine would apply to copies made abroad,

182. *See* Bobbs-Merrill Co. v. Straus, 210 U.S. 339, 350 (1908) (applying the first sale doctrine to copyright law, for the first time, in the case of a publisher who attached a notice setting a floor on resale prices for the book, and holding that copyright does not allow impositions on future sellers "to future purchasers, with whom there is no privity of contract").

183. 17 U.S.C. § 602(a)(1) (2012); *see also* § 602(a)(2) (deeming it infringement to import copies "the making of which either constituted an infringement of copyright, or which would have constituted an infringement of copyright if this title had been applicable"). The separate treatment of unauthorized copies in § 602(a)(2) might suggest that § 602(a)(1) was aimed at copies that were authorized, but the Court did not address this distinction in *Kirtsaeng*.

184. Quality King Distribs., Inc. v. L'anza Research Int'l, Inc., 523 U.S. 135, 145 (1998).

185. *Id.*

186. Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351, 1355 (2013).

187. *Id.*

188. *Id.* at 1357–60.

189. The Court discusses the legislative history of the importation provision and the Copyright Office's explanation of the current version, which states that importation without permission "would violate the exclusive rights of the U.S. copyright owner . . . where the copyright owner had authorized the making of copies in a foreign country for distribution only in that country." *Id.* at 1369 (citing COPYRIGHT LAW REVISION, PT. 6, PRELIMINARY DRAFT FOR REVISED U.S. COPYRIGHT LAW AND DISCUSSIONS AND COMMENTS, 88TH CONG., 2D SESS., 150 (Comm. Print 1964)). Although that quote contemplates that copyright

thus diminishing the reach of the importation ban significantly,[190] or it would only apply to copies made in the United States, with the result that rights in copies made abroad would never be exhausted, even if the copies were imported and sold by the copyright owner.[191] This second interpretation would allow publishers to control all downstream sales of books initially published abroad, creating incentives for publishers to move operations abroad and allowing for restraints on future sales that are at odds with a functioning market.[192] Although the *Kirtsaeng* opinion avoids an undesirable

owners would be able to authorize publication for sale in particular markets and then block importation of works made for that market—precisely the type of market segmentation that was disallowed domestically by the first sale provision—the Court concluded that the importation sections were not intended to address the issue of first sale. *Kirtsaeng*, 133 S. Ct. at 1369. *But see id.* at 1373 (Kagan, J., concurring) ("I think John Wiley may have a point about what 602(a)(1) was designed to do" after summarizing his argument that the rightful purpose of that section was "enabling copyright holders to segment international markets."). It is also surprising that the Court did not contrast the relevant importation provision for "work[s] that have been acquired outside the United States . . . ," 17 U.S.C. § 602(a)(1) (2012), with the subsequent clause that addresses importation of copies, "the making of which either constituted an infringement . . . or which would have constituted an infringement of copyright if this title had been applicable," a section that seems to describe works made without authorization in contrast to the previous section that only discusses where the works were acquired. Instead, when discussing this section, the Court suggests that it proves the American Copyright Act is "applicable to all pirated copies, including those printed overseas." *Kirtsaeng*, 133 S. Ct. at 1359 (emphasis omitted). Thus, the fact that the drafters distinguished between works made abroad and works made abroad that would have been infringing if the title had been applicable is used by the Court to prove that the Act is applicable everywhere. This reading is strained at best.

190. *See Kirtsaeng*, 133 S. Ct. at 1373 (Kagan, J., concurring) (agreeing that the importation provision is greatly limited by this decision, but suggesting that *Quality King* was the problem). As explained by Justice Kagan:

> allowing the copyright owner to restrict imports irrespective of the first-sale doctrine—i.e., reversing *Quality King*—would yield a far more sensible scheme of market segmentation than would adopting . . . Wiley's argument here. That is because only the former approach turns on the *intended market* for copies; the latter rests instead on their *place of manufacture*.

Id. at 1373 n.2.

191. *Id.* at 1362 (explaining that "a geographical interpretation of the 'first sale' clause . . . would grant the holder of an American copyright . . . permanent control over the American distribution chain (sales, resales, gifts, and other distribution) in respect to copies printed abroad but not in respect to copies printed in America"); *id.* at 1365 (explaining the effects of a geographical interpretation on technology companies whose products contain copyrightable software programs, as "[m]any of these items are made abroad with the American copyright holder's permission and then sold and imported (with that permission) to the United States.") (citing Brief for Retail Litigation Center, Inc. et al. as Amici Curiae, *Kirtsaeng*, 133 S. Ct. 1351 (No. 11-697), at *4).

192. *Id.*

interpretation, the new doctrine of international exhaustion of copyright was not likely contemplated by the law's drafters.

The statutory interpretation that underlies the *Kirtsaeng* decision, while interesting, does not constrain the possibilities for patent law. The patent law doctrine is not bound by statutory language, and the common law gives room for doctrinal evolution over time. Nonetheless, the policy arguments in *Kirtsaeng* brought out the same themes discussed in Part II, *supra*, recognizing the interest rights holders have in segmenting markets, but ultimately finding those outweighed by the interests in competition and a robust resale market. In particular, the Court looked at the historical movement away from protectionism and toward competition in addition to considering the effects of the rule on particular institutions and industries. Thus, the Court put its decision in context by noting legislative movement away from protectionism in the copyright law's elimination of a limitation on works manufactured outside the United States.[193] As previously discussed, a historical view of patent law provisions and their movement away from protectionism shows that the maintenance of national markets can be seen as one of the last relics of protectionism, and one that, alone, does not work.

The Court also explained the purposes of exhaustion in the domestic context through the policies driving its earlier, common law evolution.[194] In particular, the Court referenced the common law's "refusal to permit restraints on the alienation of chattels," which are "against Trade and Traffi[c], and bargaining and contracting."[195] In addition to historical ideas about restraints on alienation, the Court noted that exhaustion is supported by the antitrust law's purpose of "maximiz[ing] consumer welfare by encouraging firms to behave competitively."[196] These same concerns about downstream markets apply in patent law as well.[197] Of interest, the Court

193. The Court pointed to the elimination in 1976 of the "manufacturing clause" that limited importation of copies made outside the United States (or Canada), explaining that the purpose of that change was to "equalize treatment of copies manufactured in America and copies manufactured abroad." *Id.* at 1361 (citing H.R. REP. NO. 94-1476, at 165–66 (1976)). The Court has noted this change before. *Id.* (citing Golan v. Holder, 132 S. Ct. 873 (2012)) ("Congress has moved from a copyright regime that, prior to 1891, entirely excluded foreign works from U.S. copyright protection to a regime that now 'ensure[s] that most works, whether foreign or domestic, would be governed by the same legal regime.' ").

194. *Kirtsaeng*, 133 S. Ct. at 1363.

195. *Id.* (citing EDWARD COKE, THE FIRST PART OF THE INSTITUTES OF THE LAWS OF ENGLAND 223 (1628)).

196. *Id.* at 1363 (quoting 1 PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW: AN ANALYSIS OF ANTITRUST PRINCIPLES AND THEIR APPLICATION ¶ 100 (3d ed. 2006)).

197. *See supra* Part III.

summarily dismissed the common argument about the benefits of price discrimination, explaining that "no basic principle of copyright law" suggests entitlement to this right.[198] Finally, the Court also discussed the practical and negative effects of a geographical restriction on particular industries, such as the used book industry, libraries, art museums, technology companies whose products contain copyrightable software, and other retailers who sell goods with copyrighted packaging, logos, etc.[199] These businesses rely on their ability to import works made abroad, works which once resold are indistinguishable from works made within the United States, such that a rule allowing continued control of the goods would constitute barriers to commerce. The consequences of such a rule to these industries, the Court suggests, would be "intolerable."[200] Many of the goods to which the Court referred are likely to be covered by patent protection as well as copyright. Copyrighted packaging may well contain patented materials, and products including software are also generally protected by patents as well. To the extent the Court was concerned about the control of importation or resale of these products, *Kirtsaeng* only takes us part of the way.[201] An analogous rule in patents would free the goods.

B. QUESTIONING THE BENEFITS OF GEOGRAPHIC PRICE DISCRIMINATION

Previous Sections have argued that maintaining different rules for national and international exhaustion is at odds with the current purposes of patent and trade law. However, the economic argument against international exhaustion posits that the geographical price discrimination that is possible under national exhaustion carries benefits that would be lost in a move to international exhaustion. According to this view, the social welfare effects of international exhaustion would result in less innovation and less access for consumers in low-income countries, so that the otherwise outdated distinction between foreign and domestic sales is justified because it is best suited to meet the objectives of the patent system. There are insufficient data for an empirical conclusion either way. However, the traditional argument

198. *Kirtsaeng*, 133 S. Ct. at 1371.

199. *Id.* at 1364–65.

200. *Id.* at 1366.

201. It is true that the parade of horribles envisioned in *Kirtsaeng* would not fully apply in patent law. Remember that the Court was concerned that rights in products made abroad could never be exhausted by an authorized sale, as those goods were not "lawfully made under" the U.S. Copyright Act. For patents, even without a rule of international exhaustion, an authorized U.S. sale or importation of goods made abroad still exhausts rights. *See supra* note 159.

against international exhaustion fails to recognize the costs of geographical price discrimination—costs that are exacerbated by growing income disparity. These include access costs for poorer consumers in otherwise high-income countries, although even with geographical price discrimination poor consumers in low-income countries also have limited access.[202] In addition, the models fail to account for substitute price discrimination measures firms would likely undertake to limit the contemplated ill effects. These substitute measures could result in more access as firms find other ways to target broad and diverse markets.

Price discrimination occurs when a seller—particularly a monopolist—charges different prices to different buyers, based on some measure of their willingness to pay.[203] Perfect (or "first-degree") price discrimination describes the situation of a monopolist who sells to each consumer at the highest price she is willing to pay.[204] It is, however, a "never-attained theoretical limit."[205] If such a thing were possible, it would result in the greatest gain to the seller who could make more sales, extracting the highest possible price for each one. It would also theoretically result in greater consumer access to goods, because everyone willing to pay more than the marginal price of production of a good would be able to obtain it.[206] However, there would be no consumer surplus—that is, no one would be able to buy the good for less than the highest price they were willing to pay. Instead, the consumer surplus that would have existed without discriminatory pricing would all accrue to the monopolist.[207]

In reality, companies find many ways to engage in price discrimination, sometimes through volume discounts or "versioning" methods that differentiate consumers according to their willingness to pay—both types of second-degree price discrimination.[208] One example of versioning is the

202. *See infra* note 222 and accompanying text.

203. *See* ROBERT H. FRANK, MICROECONOMICS AND BEHAVIOR 389–95 (8th ed. 2010).

204. *Id.* at 393.

205. *Id.* at 394 (explaining that imperfect knowledge of consumer preferences make first-degree price discrimination impossible).

206. *Id.* at 394.

207. *Id.*

208. *Id.* at 395–96. Volume discounts are one type of second-degree price discrimination, in which the seller induces the buyer to reveal her preferences. One example of a volume discount is the pricing structure utilities companies use, charging less per kilowatt-hour after a certain limit has been reached. Companies can also try to differentiate consumers who are willing to pay more from those willing to pay less by setting "hurdles," such as mail-in rebates, that only some customers will undertake the nuisance of completing. "Versioning" refers to selling slightly different products, possibly introduced at slightly different times, such as hardcover, paperback, and electronic books. *See id.*; William W.

different amounts of memory or processing power that come with tablet computers and the regular introduction of newer models.[209] Third-degree price discrimination occurs when sellers "separate [buyers] into groups that correspond roughly to their wealth or eagerness.[210] The classic example is using demographic generalizations to grant discounts to certain groups, such as student or senior discounts in movie theaters.[211] All of these methods are imperfect, but allow sellers to reach consumers they might otherwise not reach while maintaining higher prices for a large portion of the market. The success of such methods from the seller's viewpoint depends in part on how well the division differentiates markets.[212] For example, versioning is successful if high-income buyers do not see a cheaper version of a product as a sufficient substitute for the more expensive one. It also requires that arbitrage be limited, so that low-cost buyers are unable to resell goods to high-cost buyers.[213]

The description and requirements of third-degree price discrimination apply to patents and a rule of national exhaustion.[214] A rule of national exhaustion allows patent holders to engage in geographical price discrimination, offering goods at lower prices in lower-income markets while preserving their ability to sell at higher prices in higher-income markets. This theoretically results in higher returns to the patent holder,[215] higher costs to consumers in high-income markets, and lower costs (and therefore greater access to goods) in lower-income markets. It follows that elimination of geographical price discrimination would result in lower returns to patent

Fisher III, *When Should We Permit Differential Pricing of Information?*, 55 UCLA L. REV. 1, 3–4 (2007).

209. This differentiates consumers willing to pay more for greater computing capacity and those willing to pay more for the newest version of electronics.

210. Fisher, *supra* note 208, at 4.

211. *See id.*

212. *Id.* at 4 ("the firm must be able to differentiate among its customers on the basis of the values they place on the firm's product").

213. David A. Malueg & Marius Schwartz, *Parallel Imports, Demand Dispersion, and International Price Discrimination*, 37 J. INT'L ECON. 167, 170–71 (1994).

214. The welfare effects of third-degree price discrimination are generally considered to be ambiguous. Hal R. Varian, *Price Discrimination and Social Welfare*, 75 AM. ECON. REV. 870 (1985) (generalizing results that price discrimination only increases welfare when it results in increased output).

215. Alan O. Sykes, *TRIPS, Pharmaceuticals, Developing Countries, and the Doha "Solution,"* 3 CHI. J. INT'L L. 47, 64 (2002) ("Parallel importation invariably reduces the rents that are earned by pharmaceutical patent holders. To the degree that those rents are important to inducing worthwhile R&D investments, as suggested above, this effect is unfortunate."). *But see* Peter Yu, *The International Enclosure Movement*, 82 IND. L.J. 827, 844–45 (2007) (suggesting that Sykes "overstate[s] the practical impact of [parallel] importation.").

holders, lower prices in high-income markets, and less access for those in lower-income countries. This is because the threat of competition from goods bought abroad and imported for sale—parallel imports—would cause patent holders either to raise prices abroad or to decline to sell in low-income markets.[216] David Malueg and Marius Schwartz suggest, in this vein, that when there are great income disparities between countries, international exhaustion (and the uniform pricing that resulted) would lead to lower welfare than a system of national exhaustion.[217] Mattias Ganslandt and Keith Maskus also critique international exhaustion, disagreeing that it necessarily results in lower prices in high-income locations because of changes in distributional structure that would result.[218] Kamal Saggi discusses the interests and motivations of different countries and companies, suggesting that the combination of intellectual property rights protection now mandated by TRIPS with an international exhaustion policy by high-income countries would result in welfare gains to companies and consumers in those countries at the expense of consumers in low-income countries.[219] This argument conflicts with the previous two with respect to patent holders, but the conclusion that national exhaustion allows for greater welfare is the same.

The economic arguments in favor of geographic price discrimination are not without caveats or critics. The caveats are that these theoretical models, while useful, make assumptions that are not necessarily borne out in reality

216. Malueg & Schwartz, *supra* note 213, at 171; *see also* Pinelopi Koujianou Goldberg, *Intellectual Property Rights Protection in Developing Countries: The Case of Pharmaceuticals*, 8 J. EUR. ECON. ASS'N. 326, 329–30 (2010) (arguing that pharmaceutical companies might not serve low-income countries or may raise prices there if there is parallel importing).

217. *Id.* If patent holders continue to place goods on separate markets at vastly different prices, then it is true that consumers in high-income markets benefit from parallel trade. However, because the patent holder controls market entry of the patented goods worldwide, we can expect that her behavior *ex ante* will change. One way it might change is that she stops selling goods at different prices and introduces a worldwide price (with some variation to account for transportation and other distribution cost differences among regions). Another option is for the patent holder to choose not to make goods available in some low-income markets in order to preserve high returns in the high-income market.

218. Mattias Ganslandt & Keith E. Maskus, *Vertical Distribution, Parallel Trade, and Price Divergence in Integrated Markets*, 51 EUROPEAN ECON. REV. 943, 944 (2007) (suggesting that the conclusion that "permitting [parallel imports] unambiguously brings down retail prices in expensive locations is misleading"). Ganslandt and Maskus argue that in response to an international exhaustion regime, intellectual property rights holders may consolidate distributors and change wholesale pricing in ways that would obviate perceived benefits to consumers. *Id.* at 945.

219. Kamal Saggi, *Market Power in the Global Economy: The Exhaustion and Protection of Intellectual Property*, 123 ECON. J. 131, 135 (2013). However, these results would change if strong intellectual property rights were necessary to induce importation of technology, for industries where there is a quality gap between innovator companies and imitators. *Id.*

or fail to account for reactions other than those modeled.[220] Gene Grossman and Edwin Lai suggest that international exhaustion may provide more support for innovation than national exhaustion by encouraging countries that use price controls to raise their prices, thus providing greater remuneration to patent holders.[221] Peter Yu raises a question about the extent to which firms currently engage in price discrimination that permits access to low-income markets.[222] In particular, Yu suggests that concerns about parallel imports in the pharmaceutical industry are overblown, both because many pharmaceutical companies have chosen not to enter lower-income markets even with a rule of national exhaustion and because the vast wealth disparities *within* some countries leads companies to target only the high-income market instead of selling at lower prices.[223] If the opportunity for geographical price discrimination is not resulting in greater access, the argument for keeping it gets weaker. Although Yu discusses the pharmaceutical market, this critique also questions the usefulness of geographic markets as demand indicators in general. As discussed above, price discrimination works best when it successfully differentiates markets and arbitrage is limited. A rule of national exhaustion satisfies the second condition, but its application to the first may be questioned. For example, even developed countries have increasingly large levels of wealth disparity,[224]

220. *See, e.g.*, Abbott, *supra* note 51, at 11 (noting Maskus's reminder that current static game theory models were incomplete and pointing to later work that shows how the reduction of trade barriers will result in increasing benefits from parallel trade).

221. Gene M. Grossman & Edwin L.-C. Lai, *Parallel Imports and Price Controls*, 39 RAND J. ECON. 378, 380 (2008) (arguing that worldwide international exhaustion leads to more innovation than national exhaustion for industries with price controls because countries face the possibility that innovator companies will choose not to sell to them). Grossman and Lai base their argument on the idea that governments with price controls will tend to raise their price caps under a scheme of international exhaustion to ensure that producers will continue to serve their market. However, in the drug industry, many countries with price caps also require companies to fulfill demand and are able to threaten compulsory licensing when the demand is not met.

222. Yu, *supra* note 215, at 844–45.

223. *Id.*; *see also* Keith E. Maskus, *Ensuring Access to Essential Medicines: Some Economic Considerations*, 20 WIS. INT'L L.J. 563, 566 (2002) (explaining that sometimes "pharmaceutical firms and their distributors in poor countries may find it more profitable to sell drugs in low volumes and high prices to wealthier patients with price-inelastic demand rather than in high volumes at low prices to poorer patients").

224. *See, e.g.*, ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *An Overview of Growing Income Inequalities in OECD Countries: Main Findings*, *in* DIVIDED WE STAND: WHY INEQUALITY KEEPS RISING 21, 22 (2011), *available at* http://www.oecd.org/els/soc/49499779.pdf. In America, the GINI coefficient, a measure of income inequality that is 0 under conditions of perfect equality and 1 under complete inequality, has steadily risen from 0.399 in 1967 to 0.466 in 2001, according to information from the U.S. Census

and therefore increasingly diverse levels of demand. The social welfare argument for geographical price discrimination becomes weaker when it does not result in greater access for low-income countries and simultaneously results in high prices for poorer members of developed countries. With regards to limitation of arbitrage: one reason that geographical price discrimination is so attractive to patent holders is that enforcement costs are relatively low. However, this is a benefit that only accrues to the patent holders. Borders are relatively easy to patrol for infringing goods,[225] and a hefty portion of the cost is borne by the government through its deployment of customs officials rather than the patent holder. This makes a national exhaustion rule more attractive to rights holders, even when national markets are not an ideal basis for distinguishing among variable levels of demand.

Another critique is that geographical price discrimination should be compared with a regime in which sellers engage in other forms of price discrimination, rather than a world with uniform pricing. Patent holders in an international exhaustion regime will be unable to sell to large portions of foreign markets if they engage in uniform pricing (or choose not to sell abroad at all).[226] If their only choices are whether to sell and how to set the price, the arguments suggesting there will be a worldwide, uniform price or abandonment of foreign markets may be correct. However, patent holders may well choose to offer more versions of patented products and engage in second-degree price discrimination mechanisms to capture more of the market. This reaction would mitigate access concerns for consumers in low-income countries. In fact, if companies faced additional pressure to develop multiple versions of a good, we could expect increased access in both low- and high-income countries (where income disparity may already omit a number of people from the market).

In sum, geographic price discrimination likely results in higher returns to patent holders. However, some of these higher returns are due to enforcement burdens taken on by customs officials. In addition, while geographic price discrimination allows for greater access by residents of low-income countries, in practice these populations are often underserved even in

Bureau. *See Historical Income Tables-Income Equality*, http://web.archive.org/web/20070208142023/http://www.census.gov/hhes/www/income/histinc/ie6.html.

225. Based on laws that require importers to declare goods and have them inspected, as opposed to domestic market transactions that may be conducted entirely privately and may therefore be difficult for patent holders to detect. 19 U.S.C. § 1484 (2012) (requiring that importers use reasonable care in making entries and classifying the imported merchandise).

226. *But see* Ganslandt & Maskus, *supra* note 218, at 4 ("It is conceivable that wholesale prices may be set in a way that offsets or even counteracts the anticipated impacts of an open [parallel importation] regime.").

our current regime. Last, proponents of geographic price discrimination fail to account for the likely move of firms to other forms of price discrimination that might better segment and identify groups with differing demand curves. A rule of international exhaustion would likely encourage *better* price discrimination. The appropriate scope of that price discrimination is taken up in the next Section.

C. IMPLICATIONS FOR OTHER FORMS OF PRICE DISCRIMINATION

This Article has suggested that patent holders will engage in other forms of price discrimination in response to an international exhaustion regime. This and other strategic moves patent holders make in response to international exhaustion merit more analysis. For example, price discrimination through versioning may be overall beneficial because it will result in greater access and the customers who pay more will in fact receive something more for the higher price. However, other potential consequences of a move to international patent exhaustion are increased use of licenses as a means for patent holders to maintain control over their goods and increased use of technological measures to maintain such control.

Restrictive licenses circumvent limitations on restraints on alienation and could redo by contract what I have suggested undoing by international treaty. In the world of electronics and software, goods are increasingly licensed rather than sold. These licenses—and technological measures that restrict their violation—may frustrate resale, but they also allow for a form of price discrimination that may be more tailored than geographic methods. Thus, one's position on licensing need not be dictated by one's support for international exhaustion. Nor should the potential for more restrictive licenses in and of itself conclusively argue against international exhaustion. Many of the same questions about the benefits of price discrimination versus those of a robust resale market and reduced transaction costs apply to the licensing debate. Indeed, the doctrinal development of the exhaustion doctrine in the United States has been a story about balancing the ability of parties to contract for limitations on patent licenses with the transaction costs associated with goods having complex limitations on all future transactions.[227] Geographic limitations are but one possible type.

Another related consideration is the extent to which common ownership of the relevant intellectual property rights should matter to a determination that an "authorized" first sale has taken place and exhausted a patent holder's

227. *See* Mossoff, *supra* note 156, at 12 (describing the ability of nineteenth century patent holders to impose a "litany of restrictions on the use of the property interest they conveyed to a licensee").

rights. This question has been presented in patent,[228] copyright,[229] and trademark[230] cases in the national and international contexts. The early patent cases ruled that exhaustion only occurs for unconditional sales made by owners in multiple jurisdictions, whereas granting exclusive, geographic licenses to different entities allows each to use the patent laws to enforce their entitlements, thereby presenting a type of default interpretation on sales that a patent holder could overcome through explicit licensing and restrictions.[231] These decisions, while perhaps right as applied among the various licensees, may need to be recalibrated to the extent they allow restrictions to travel with the goods, affecting unwitting downstream consumers. Just as the scope of permissible licensing presents questions, the potential for licensing to separate entities to avoid exhaustion will require further scrutiny and doctrinal development.

The proper scope of the exhaustion doctrine as applied to non-geographic restrictions and to particular forms of technology is a topic that merits further inquiry—particularly if there is widespread adoption of an international exhaustion regime. This move would bring into focus the issue of the allowable scope of private restrictions on the movement of goods. It does not, however, argue against bringing international patent law into conformity with domestic law, or patent law into conformity with copyright law.

D. IMPLEMENTATION

There are many ways that a rule of international patent exhaustion could be implemented. The simplest is that it could be unilaterally adopted by the Supreme Court, which could distinguish *Boesch* on the basis that the sale in *Boesch* did not involve authorization from the patent holder. In doing so, the Court would bring together patent and copyright doctrine by finding exhaustion following an authorized first sale, worldwide. However, this would diminish the possibility of benefitting from the greater gains that

228. *See supra* Section IV.B.

229. For a discussion of how licensing can be used to avoid exhaustion for copyrighted works, see Clark D. Asay, Kirtsaeng *and the First-Sale Doctrine's Digital Problem*, 66 STAN. L. REV. ONLINE 17 (2013); Carver, *supra* note 131.

230. In *Établissement Consten*, Grundig had rights to a trademark and licensed to Consten the right to register the trademark in France in Consten's name, thereby planning to enjoin imports of the trademarked goods in Germany, where Grundig held the trademark. The European Court of Justice found this violated the competition law of the European Economic Community ("EEC") Treaty by attempting to artificially maintain separate markets. Joined cases 56 & 58/64, Établissement Consten S.A.R.L. v. Comm'n, 1966 E.C.R. 299.

231. *See supra* Section IV.B.

would come from a harmonized world rule of international exhaustion. In particular, transaction costs would only be slightly lower with one more country participating in a regime of international exhaustion, whereas the ability to engage in cross-border manufacture without concerns about crisscrossing licenses would eliminate an entire class of transaction costs. Short of a worldwide exhaustion regime, it might be possible to implement various multilateral exhaustion regions (such as the European Union already has) or to engage in a reciprocal recognition of exhaustion. A worldwide exhaustion regime aligns better with the purposes of TRIPS, and a patent holder's decision to license or sell rights in another territory ought not to remove a patent from the potential for exhaustion.

From a patent law perspective, the harmonization TRIPS brought to procedural and substantive laws governing patents provides a number of benefits. One benefit is certainty to inventors and investors that their inventions will receive protection worldwide, thus inspiring advancement in new technology areas and technology transfer to new markets.[232] Harmonization also lowers costs—both direct costs (e.g., by allowing central filing for multiple countries) and the indirect costs of understanding and analyzing various sets of laws and their implications.[233] From a trade perspective, inadequate (or differential) intellectual property protection can constitute a non-tariff barrier to trade.[234] Thus, in a country with strong patent rights, patent holders are likely to commercialize and market their goods; however, those same patent holders will be wary of exporting and making the goods available in a country without patent protection because others may copy the innovation and compete with the patent holder.[235] In

232. *See* Rajec, *supra* note 29, at 41–42 (discussing the certainty justification for harmonization of global patent law); *see also* Frederick M. Abbott, *The Enduring Enigma of TRIPS: A Challenge for the World Trading System*, 1 J. INT'L ECON. L. 497, 499 (1998) (noting that TRIPS provides "the enhancement of [industrialized country-based enterprises'] legal security in a wider portion of the world market").

233. *See* Rajec, *supra* note 29, at 46.

234. 1 THE GATT URUGUAY ROUND: A NEGOTIATING HISTORY (1986–1992), at 707–08 (Terence P. Stewart ed., Kluwer Law 1993); *see also* TRIPS, *supra* note 53, pmbl. (stating that the agreement was intended "to reduce distortions and impediments to international trade, and tak[e] into account the need to promote effective and adequate protection of intellectual property rights . . . .").

235. The potential for such imitations to be exported back to the strong-patent-rights country may also lessen incentives to invest. This problem is addressed through bans on infringing imports and by increased harmonization on laws governing counterfeits. In fact, concerns about trade in counterfeit goods provided some of the strongest motivation to include intellectual property rights in the Uruguay Round negotiations. TRIPS, pmbl. ("[r]ecognizing the need for a multilateral framework of principles, rules and disciplines dealing with international trade in counterfeit goods").

this way, the minimum standards of protection set by the TRIPS Agreement were seen as reducing a barrier to trade.

A harmonized rule on exhaustion would have further reduced such barriers.[236] Instead, there is a patchwork of rules. The European Union, for example, practices regional exhaustion for goods that are placed on the market within the European Union.[237] Other countries practice international exhaustion.[238] The United States has continued to adhere only to national exhaustion. It would be possible and permitted under the WTO to unilaterally adopt an international exhaustion regime for patents (as the United States has now done for copyright), but if international exhaustion in fact makes sense (as I argue it does), it does so in a world with harmonized international exhaustion. The gains to manufacturing efficiency and the certainty of freedom to use and resell goods is much greater if manufacturers, downstream retailers, and users do not have to keep in mind the multiple different treatments of goods that occur depending on which countries' versions of a patent apply to particular goods or parts of goods. In addition, using the WTO framework to bind all member countries eliminates the potential for strategic decisions by countries to tailor their exhaustion stance in deference to current constituencies that seek short term gains.[239] For that reason, any movement toward international exhaustion from the United States should be a move toward a harmonized regime of international exhaustion.

---

236. During negotiation of the TRIPS Agreement, A.A. Yusuf and A. Moncayo von Hase suggested that the decision of whether to implement a regime of international exhaustion "has profound implications for the free movement of goods and services across national boundaries, and thus for international trade in general." Yusuf & Moncayo von Hase, *supra* note 45, at 116 (advocating for an international exhaustion regime).

237. This rule was first introduced in decisions of the European Court of Justice, Case 78/70, Deutsche Grammophon, GmbH v. Metro-SB-Großmärket, GmbH, 1971 E.C.R. 487; Case 15/74 Centrafarm BV v. Sterling Drug Inc., 1974 E.C.R. 1147, and later adopted under the Treaty of Rome. *See* Consolidated Version of the Treaty on the Functioning of the European Union arts. 28, 30, May 9, 2008, 2008 O.J. (C 115) 59–60. Switzerland has also adopted a rule of exhaustion for goods from the European Union since July 2009. Press Release, Secrétariat Général, Département Fédéral de Justice et Police, Epuisement Régional en Droit des Brevets [Regional Exhaustion in Patent Law] (May 29, 2009), *available at* http://www.ejpd.admin.ch/content/ejpd/fr/home/dokumentation/mi/2009/2009-05-29.html.

238. Some countries that have adopted international exhaustion by statute include Egypt, South Africa, Argentina, Costa Rica, India, Malaysia, and China.

239. *See* Vincent Chiappetta, *The Desirability of Agreeing to Disagree: The WTO, TRIPS, International IPR Exhaustion and a Few Other Things*, 21 MICH. J. INT'L L. 333, 348 (2000) (explaining the various exhaustion regimes by suggesting that "[n]ations (and regions) are not coming out differently on the same global calculation; they are performing independent, self-interested valuations").

A worldwide rule of exhaustion would represent a significant change in current policy. The United States was stalwart in its refusal to implement international exhaustion during TRIPS negotiations.[240] A shift in position might therefore have some worth in future multilateral negotiations relating to intellectual property. In other words, moving to an international exhaustion regime is both desirable in theory and of potential negotiating use as a practical matter.

Other possibilities are a regional or development-based multilateral exhaustion regime. Under such a regime, markets with roughly similar characteristics might create zones of exhaustion. The benefits would be that patent holders could still engage in some degree of price discrimination, assured that goods placed on the market in other "zones" would not be imported and compete with goods in high-income locations. At the same time, manufacturers could avoid transaction costs associated with cross-border manufacture by producing goods within a given zone. And consumers could still expect some level of competition between the various intrazone markets. Nonetheless, this idea is deeply problematic because of its potential to entrench countries within a particular zone, as the costs of breaking into a new one would become barriers to development. If a country were part of a lower-income zone and also a manufacturing zone, for example, the costs of trying to move into a higher-income zone would include consumer loss as prices went up and the loss of manufacturing jobs. Such a situation would be inconsistent with the purposes of the WTO agreement, if not with its substantive rules.[241]

E. ADDRESSING CRITIQUES: THE PHARMACEUTICAL INDUSTRY

The pharmaceutical industry may warrant special treatment when it comes to price discrimination because of the non-market nature of the heavily regulated industry and because of the importance access to healthcare plays in global notions of human rights. Concerns about the availability of patented drugs in low-income countries are not merely academic.[242] The

240. *See also* GERVAIS, *supra* note 170, at 60–63.

241. A regional exhaustion regime might violate most-favored nation requirements by treating imports from different countries differently, although not necessarily resting on nationality of patent holders. *See* TRIPS, *supra* note 53, art. 4 (requiring that "any advantage, favour, privilege or immunity granted by a Member to the nationals of any other country shall be accorded . . . to the nationals of all other Members").

242. *See* Ellen t'Hoen, *TRIPS, Pharmaceutical Patents, and Access to Essential Medicines: A Long Way from Seattle to Doha*, 3 CHI. J. INT'L L. 27, 27 (2002) (describing that ninety percent of those killed by infectious diseases each year are in the developing world and suggesting that "[t]he reasons for the lack of access to essential medicines are manifold, but in many cases the high prices of drugs are a barrier to needed treatments").

importance of access to medicines has spurred concerns with the requirements of minimum levels of patent protection throughout the history of the TRIPS Agreement.[243] One of the main points of contention between the global north and the global south during negotiations was the treatment of patents for pharmaceutical products. Many developing and least developed countries did not allow them, and there was great concern that implementing such regimes would lead to a crisis in access to medicine for poor countries.[244] Whereas nongeographic forms of price discrimination may be effective for many consumer goods,[245] these methods do not lend themselves easily to use in the pharmaceutical industry. For example, providing less effective versions of drugs for lower prices is conceptually and ethically problematic. As a result, one of the strongest counterarguments to a regime of international exhaustion focuses on the effects such a regime would have on patient populations in poor countries. The argument is that major pharmaceutical companies will refuse to sell in such low-income markets because the potential for parallel imports to erode the very large profits they glean in high-income markets would make it unprofitable.[246] The other side of the same argument against international exhaustion centers on the high cost of research and development that goes into drug development and the relative ease with which drugs can be reverse engineered.[247] Under this argument, the losses to pharmaceutical companies under an international exhaustion regime would lower incentives to innovate significantly. From either perspective, drugs are special. Access is of greater importance in the area of health, and profits derived from patents are valued more by innovators in this area.

The strong control of market entry exercised by the Food and Drug Administration ("FDA") (and equivalent agencies in other developed countries) could be used to exempt drugs sold to least developed countries from international exhaustion.[248] Drugs must go through an approval process

243. *See, e.g.*, Amy Kapczynski, *Harmonization and Its Discontents: A Case Study of TRIPS Implementation in India's Pharmaceutical Sector*, 97 CALIF. L. REV. 1571, 1571–72 (2009).

244. *See* CYNTHIA M. HO, ACCESS TO MEDICINE IN THE GLOBAL ECONOMY: INT'L AGREEMENTS ON PATENTS AND RELATED RIGHTS 91 (2011).

245. *See supra* Part II (discussing price discrimination).

246. Atik & Lidgard, *supra* note 104, at 1045–46 (arguing that price discrimination is important in getting pharmaceutical products to least developed countries but that there has been a strange ambivalence toward the use of price discrimination).

247. *See, e.g.*, Henry Grabowski, *Patents and New Product Development in the Pharmaceutical and Biotechnology Industries*, *in* THE PROCESS OF NEW DRUG DISCOVERY AND DEVELOPMENT 533, 535 (Charles G. Smith & James T. O'Donnell eds., 2d ed. 2002).

248. A rule against unauthorized reimportation is already part of the FDA's governing statute. *See* Federal Food, Drug, and Cosmetic Act, 21 U.S.C. § 381(d)(1) (2012) (providing

before they may be marketed. In the United States, the FDA has a complex process for companies wishing to sell drugs—patented or not—to patients that aims to ensure the drugs are both safe and effective.[249] The registration of drugs and approval of processes mean that the FDA serves as a gatekeeper for all who wish to sell drugs in the U.S. market, and an expansion in this role to exclude drugs first sold in least developed countries would not exceed the scope of the agency's current expertise.

While a patent holder could respond to the introduction of international exhaustion by not selling drugs in low-income markets at all, this would be an undesirable outcome because it might result in a lack of access to lifesaving medicines. Alternatively, the decision could raise the possibility of countries engaging in compulsory licensing to produce generic versions of lifesaving drugs.[250] This result would still insulate the pioneer drug makers from parallel imports because drugs produced pursuant to a compulsory license would not be interpreted as authorized sales.[251] However, this result imposes costs (lost sales to the otherwise willing drug company, procedural costs of procuring a license, and—although relatively small—the costs to generic drug companies of reverse engineering). With a small margin for profits in least developed countries to start with, these inefficiencies could be problematic to the access cause. A better policy would encourage patent holders to manufacture and sell drugs to low-income markets and reserve compulsory licenses for situations where patent holders are unwilling to do so. The position of the FDA as a market gatekeeper means that drugs sold to

that "no [prescription] drug . . . which is manufactured in a State and exported may be imported into the United States unless the drug is imported by the manufacturer of the drug."). Kevin Outterson explains how that law has been used to limit the reimportation of drugs produced in the United States and sent to Canada. Kevin Outterson, *Pharmaceutical Arbitrage: Balancing Access and Innovation in International Prescription Drug Markets*, 5 YALE J. HEALTH POL'Y L. & ETHICS 193, 213–14 (2005) ("The law was ostensibly intended to address safety concerns for the U.S. pharmaceutical supply chain, but its effect is to prevent international pharmaceutical arbitrage or parallel trade." (citations omitted)).

249. *See* 21 U.S.C. § 393(b)(2)(B) (establishing the FDA and giving its mission as "protect[ing] the public health by ensuring that . . . drugs are safe and effective"); *see also* HO, *supra* note 244, at 13–16, 94 (describing similar agencies in other countries).

250. *See* TRIPS, *supra* note 53, art. 31 (allowing and setting parameters for compulsory licensing).

251. In addition, the compulsory licensing provisions of TRIPS require that goods produced pursuant to such a license be primarily for the domestic market, although manufacturers in a given country can also do so for a particular foreign market that does not have manufacturing capabilities. TRIPS, *supra* note 53, art. 31. Similarly, for countries in which inventors have chosen not to seek patents, manufacture and sales of those products would not be infringing, but they would also not be "authorized," and thus sale would not exhaust the patent holder's rights in other countries.

certain countries, or under certain conditions, could be exempted from international exhaustion, thus encouraging drug companies to sell drugs at the cost of production (or below) without concerns about parallel imports.

In addition to access to medicine issues for poor countries, the drug industry may be appropriate for special treatment because it is subject to price controls in so many countries.[252] These market conditions mean that in many places, patent holders have little control over the price at which they introduce products to the market, making the single reward justification for international exhaustion less compelling. Under a regime of price control, the autonomy of the patent holder in choosing to place goods on the market at a given price is no longer so clear. While patent holders may retain the choice of *whether* to sell, they no longer decide a price. For developed or developing countries that engage in price controls, patented medicines are bought at prices much lower than in countries (like the United States) that do not engage in such measures. And unlike the case of compulsory licensing, such sales are authorized. A regime of international exhaustion could be devastating to patent holding companies facing competition from imports in countries that engage in such pricing.[253] The potential for lower prices of imported goods would not be a consequence of comparative advantage in manufacturing or the benefits of increased competition, but a result of disparate regulatory control.[254]

One possible solution would be to exempt pharmaceuticals from parallel importation entirely. Such a move would recognize that although similar levels of patent protection exist in all WTO member countries, other market forces intervene to alter the benefits a patent confers. In this sense, it is the unequal protection afforded to pharmaceutical products by patents that

252. John A. Vernon et al., *The Economics of Pharmaceutical Price Regulation and Importation: Refocusing the Debate*, 32 AM. J.L. & MED. 175, 176 (2006).

253. Europe has seen an increase in parallel trade in pharmaceutical products because the European Union practices regional exhaustion. The combination of regional exhaustion with strong price controls has proved problematic, as parallel traders engage in arbitrage that does not bring down the cost of medicines in countries that pay more for them, but serves only to appropriate profits that would otherwise go to patent holding pharmaceutical companies. As a result, courts have become sympathetic to companies that limit their supply in accordance with the demand of a given market. *See, e.g.*, Case C-53/03, Synetairismos Farmakopoion Aitolias & Akarnanias (Syfait) v. GlaxoSmithKline AEVE, 2005 E.C.R. 4609, 4637 (opining that GlaxoSmithKline AEVE was justified in refusing to meet in full orders it received from Greek pharmaceutical wholesalers in order to limit parallel trade "where the price differential giving rise to the parallel trade is a result of State intervention in the Member State of export to fix the price there at a level lower than that which prevails elsewhere in the Community . . . .").

254. Vernon et al., *supra* note 252, at 182–83.

creates barriers to trade in patented goods. However, as with any rule targeting a particular technology area, this could result in line-drawing problems. The scope of the exemption from international exhaustion could extend to all pharmaceutical products or methods of treating diseases using chemical compounds, but it could also cover treatment with biosimilars. Implementing legislation would have to describe a category that itself is growing due to innovation.

While implementation of a wide-scale exemption could be delegated to the FDA, another possibility for treatment of drugs sold subject to price regulations would draw upon trade mechanisms and the expertise of the agencies that implement them. Under this potential scheme, imports of patented pharmaceuticals from countries engaged in price controls could be treated as potentially dumped goods under trade law, sold in a non-market industry (akin to non-market economies in trade).[255] Under our trade law, foreign manufacturers and importers cannot sell goods in the United States at less than fair value—that is, they cannot "dump" their goods on the U.S. market, because of the harm that would do to domestic industry.[256] If such dumping is found, the Department of Commerce calculates the amount by which the sale is lower than normal value and taxes imports accordingly, raising the prices of the imported goods on the U.S. market so that they are in "fair" competition with domestically produced goods. In the drug industry, drugs imported from single-payer system countries would surely sell for less than domestically produced drugs, but because their U.S. prices would be comparable to those of the home market, they would not usually qualify as dumped. However, there is special treatment of goods that come from a non-market economy, defined as a foreign country that does "not operate on market principles of cost or pricing structures, so that sales of merchandise in such country do not reflect the fair value of the merchandise."[257] Drugs don't function according to market principles of pricing.[258] Therefore, the Department of Commerce would be free to compare their prices in the United States with a "constructed fair market value," which would likely be the market value in the United States. The

255. Dumping refers to selling goods in a foreign market at lower than "fair value." Most often, this is manifested in selling goods for a lower price in a foreign market than in the producer's home market. Luke P. Bellocchi, *The Effects of and Trends in Executive Policy and Court of International Trade (CIT) Decisions Concerning Antidumping and the Non-Market Economy (NME) of the People's Republic of China*, 10 N.Y. INT'L L. REV. 177, 179–80 (1997).

256. JOHN H. JACKSON, *Dumping in International Trade: Its Meaning and Context*, *in* ANTIDUMPING LAW AND PRACTICE 4–5, 11 (1990).

257. 19 U.S.C. § 1677(18)(A) (2012).

258. *See* Vernon et al., *supra* note 252, at 176.

treatment of imports of drugs from single-payer systems as dumped merchandise is one possibility for controlling the potential downsides of a system of international exhaustion.

VI. CONCLUSION

The United States should adopt a rule of international exhaustion in patent law. This would bring the law in line with copyright law and vindicate the practical goals of the *Kirtsaeng* opinion. But more importantly, it would apply theory that was once only considered in the national context to international sales and movements of goods in a way that increases competition and lowers barriers to trade. Modern trade law operates under a recognition that greater global welfare comes from increased competition and freedom for downstream innovators, retailers, and consumers. Although patent law has adopted some of the formalities of modern trade theory, it has so far clung to the protection of national markets for patent holders. This change would involve increased costs to patent holders, but it is unlikely that those increases will be as drastic as some argue. In addition, the change may encourage other forms of price discrimination that better track differences in consumer preferences. An increase in other forms of price discrimination brings into relief questions—already important in domestic law—about the appropriate level of control patent holders should have over downstream sales. These are undoubtedly important questions. However, the answers should not depend on geographical borders. A move to international exhaustion would bring patent law in line with trade theory and allow for free trade in patented goods.