

Texas Law Review

EXCHANGING INFORMATION WITHOUT INTELLECTUAL PROPERTY
Michael J. Burstein

SOLVING THE PATENT SETTLEMENT PUZZLE
Einer Elhauge & Alex Krueger

BOOK REVIEWS
Ashutosh A. Bhagwat
Mary Coombs
Frederick T. Davis
Peter Edelman
Christopher Slobogin
Timothy Zick

SETTING EXAMPLES, NOT SETTLING:
TOWARD A NEW SEC ENFORCEMENT PARADIGM

BLOWING THE WHISTLE ON CIVIL RIGHTS:
ANALYZING THE FALSE CLAIMS ACT AS AN ALTERNATIVE
ENFORCEMENT METHOD FOR CIVIL RIGHTS LAWS

Texas Law Review

A national journal published seven times a year

Recent and Forthcoming Articles of Interest

Visit www.texasrev.com for more on recent articles

SPEAKING TRUTH TO FIREPOWER: HOW THE FIRST AMENDMENT DESTABILIZES THE SECOND

Gregory P. Magarian

November 2012

STATUTES IN COMMON LAW COURTS

Jeffrey A. Pojanowski

February 2013

Individual issue rate: \$15.00 per copy

Subscriptions: \$47.00 (seven issues)

Order from:

School of Law Publications
University of Texas at Austin
727 East Dean Keeton Street
Austin, Texas USA 78705
(512) 232-1149

<http://www.utexas.edu/law/publications>

Texas Law Review *See Also*

Responses to articles and notes found in this and other issues are available at www.texasrev.com/seealso

IMPROVING FORENSIC SCIENCE THROUGH STATE OVERSIGHT: THE TEXAS MODEL

The Honorable Juan Hinojosa

Receive notifications of all *See Also* content—sign up at www.texasrev.com.

TEXAS LAW REVIEW ASSOCIATION

OFFICERS

ERIC NICHOLS
President-Elect

NINA CORTELL
President

AMELIA A. FRIEDMAN
Executive Director

JAMES A. HEMPHILL
Treasurer

HON. DIANE P. WOOD
Immediate Past President

BOARD OF DIRECTORS

R. DOAK BISHOP
JAMES A. COX
ALISTAIR B. DAWSON
KARL G. DIAL
GARY L. EWELL
STEPHEN FINK

DIANA M. HUDSON
DEANNA E. KING
JEFFREY C. KUBIN
D. MCNEEL LANE
LEWIS T. LECLAIR
JOHN B. MCKNIGHT
ELLEN PRYOR

CHRIS REYNOLDS
DAVID M. RÖDI
REAGAN W. SIMPSON
HON. BEA ANN SMITH
STEPHEN L. TATUM
MARK L.D. WAWRO

SCOTT J. ATLAS, *ex officio Director*
PARTH S. GEJJI, *ex officio Director*

Texas Law Review (ISSN 0040-4411) is published seven times a year—November, December, February, March, April, May, and June. The annual subscription price is \$47.00 except as follows: Texas residents pay \$50.88 and foreign subscribers pay \$55.00. All publication rights are owned by the Texas Law Review Association. *Texas Law Review* is published under license by The University of Texas at Austin School of Law, P.O. Box 8670, Austin, Texas 78713. Periodicals Postage Paid at Austin, Texas, and at additional mailing offices.

POSTMASTER: Send address changes to The University of Texas at Austin School of Law, P.O. Box 8670, Austin, Texas 78713.

Complete sets and single issues are available from WILLIAM S. HEIN & CO., INC., 1285 Main St., Buffalo, NY 14209-1987. Phone: 1-800-828-7571.

Single issues in the current volume may be purchased from the *Texas Law Review* Publications Office for \$15.00 per copy plus shipping. Texas residents, please add applicable sales tax.

The *Texas Law Review* is pleased to consider unsolicited manuscripts for publication but regrets that it cannot return them. Please submit a single-spaced manuscript, printed on one side only, with footnotes rather than endnotes. Citations should conform with *The Greenbook: Texas Rules of Form* (12th ed. 2010) and *The Bluebook: A Uniform System of Citation* (19th ed. 2010). Except when content suggests otherwise, the *Texas Law Review* follows the guidelines set forth in the *Texas Law Review Manual on Usage & Style* (12th ed. 2011), *The Chicago Manual of Style* (16th ed. 2010), and Bryan A. Garner, *A Dictionary of Modern Legal Usage* (2d ed. 1995).

© Copyright 2012, Texas Law Review Association

Editorial Offices: *Texas Law Review*
727 East Dean Keeton Street, Austin, Texas 78705
(512) 232-1280 Fax (512) 471-3282
tr@law.utexas.edu
<http://www.texaslrev.com>

THE UNIVERSITY OF TEXAS SCHOOL OF LAW

ADMINISTRATIVE OFFICERS

WARD FARNSWORTH, B.A., J.D.; *Dean, John Jeffers Research Chair in Law.*
ROBERT M. CHESNEY, B.S., J.D.; *Associate Dean for Academic Affairs, Charles I. Francis Professor in Law.*
WILLIAM E. FORBATH, A.B., B.A., Ph.D., J.D.; *Associate Dean for Research, Lloyd M. Bentsen Chair in Law.*
STEFANIE A. LINDQUIST, B.A., J.D., Ph.D.; *Associate Dean for External Affairs, Charles Alan Wright Chair in Federal Courts.*
EDEN E. HARRINGTON, B.A., J.D.; *Associate Dean for Experiential Education, Dir. of William Wayne Justice Ctr. for Public Interest Law, Clinical Professor.*
KIMBERLY L. BIAR, B.B.A.; *Assistant Dean for Financial Affairs, Certified Public Accountant.*
MICHAEL J. ESPOSITO, B.A., J.D., M.B.A.; *Assistant Dean for Continuing Legal Education.*
KIRSTON FORTUNE, B.F.A.; *Assistant Dean for Communications.*
MICHAEL HARVEY, B.A., B.S.; *Assistant Dean for Technology.*
MONICA K. INGRAM, B.A., J.D.; *Assistant Dean for Admissions and Financial Aid.*
TIM KUBATZKY, B.A.; *Interim Assistant Dean for Development and Alumni Relations.*
DAVID A. MONTOYA, B.A., J.D.; *Assistant Dean for Career Services.*
BRANDI L. WELCH, B.A., J.D.; *Interim Assistant Dean for Student Affairs.*

FACULTY EMERITI

HANS W. BAADÉ, A.B., J.D., LL.B., LL.M.; *Hugh Lamar Stone Chair Emeritus in Civil Law.*
RICHARD V. BARNDT, B.S.L., LL.B.; *Professor Emeritus.*
WILLIAM W. GIBSON, JR., B.A., LL.B.; *Sylvan Lang Professor Emeritus in Law of Trusts.*
ROBERT W. HAMILTON, A.B., J.D.; *Minerva House Drysdale Regents Chair Emeritus.*
DOUGLAS LAYCOCK, B.A., J.D.; *Alice McKean Young Regents Chair Emeritus.*
J.L. LEBOWITZ, A.B., J.D., LL.M.; *Joseph C. Hutcheson Professor Emeritus.*
JOHN T. RATLIFF, JR., B.A., LL.B.; *Ben Gardner Sewell Professor Emeritus in Civil Trial Advocacy.*
MICHAEL M. SHARLOT, B.A., LL.B.; *Wright C. Morrow Professor Emeritus in Law.*
JOHN F. SUTTON, JR., J.D.; *A.W. Walker Centennial Chair Emeritus.*
JAMES M. TREECE, B.A., J.D., M.A.; *Charles I. Francis Professor Emeritus in Law.*
RUSSELL J. WEINTRAUB, B.A., J.D.; *Ben H. & Kitty King Powell Chair Emeritus in Business & Commercial Law.*

PROFESSORS

DAVID E. ADELMAN, B.A., Ph.D., J.D.; *Harry Reasoner Regents Chair in Law.*
DAVID A. ANDERSON, A.B., J.D.; *Fred & Emily Marshall Wulff Centennial Chair in Law.*
MARK L. ASCHER, B.A., M.A., J.D., LL.M.; *Joseph D. Jamail Centennial Chair in Law.*
RONEN AVRAHAM, M.B.A., LL.B., LL.M., S.J.D.; *Thomas Shelton Maxey Professor in Law.*
LYNN A. BAKER, B.A., B.A., J.D.; *Frederick M. Baron Chair in Law, Co-Director of Center on Lawyers, Civil Justice, and the Media.*
MITCHELL N. BERMAN, A.B., M.A., J.D.; *Richard Dale Endowed Chair in Law.*
BARBARA A. BINTLIFF, M.A., J.D.; *Joseph C. Hutcheson Professor in Law, Director of Tarlton Law Library & the Jamail Center for Legal Research.*
LYNN E. BLAIS, A.B., J.D.; *Leroy G. Denman, Jr. Regents Professor in Real Property Law.*
ROBERT G. BONE, B.A., J.D.; *G. Rollie White Teaching Excellence Chair in Law.*
OREN BRACHA, LL.B., S.J.D.; *Howrey LLP and Arnold, White, & Durkee Centennial Professor.*
J. BUDZISZEWSKI, B.A., M.A., Ph.D.; *Professor.*
NORMA V. CANTU, B.A., J.D.; *Professor of Law and Education.*
LOFTUS C. CARSON, II, B.S., M. Pub. Affrs., M.B.A., J.D.; *Ronald D. Krist Professor.*
MICHAEL J. CHURGIN, A.B., J.D.; *Raybourne Thompson Centennial Professor.*
JANE M. COHEN, B.A., J.D.; *Edward Clark Centennial Professor.*
FRANK B. CROSS, B.A., J.D.; *Herbert D. Kelleher Centennial Professor of Business Law.*
WILLIAM H. CUNNINGHAM, B.A., M.B.A., Ph.D.; *Professor.*
JENS C. DAMMANN, J.D., LL.M., Dr. Jur., J.S.D.; *William Stamps Farish Professor in Law.*
JOHN DEIGH, B.A., M.A., Ph.D.; *Professor of Law and Philosophy.*
MECHELE DICKERSON, B.A., J.D.; *Arthur L. Moller Chair in Bankruptcy Law and Practice.*
GEORGE E. DIX, B.A., J.D.; *George R. Killam, Jr. Chair of Criminal Law.*
JOHN S. DZIENKOWSKI, B.B.A., J.D.; *Dean John F. Sutton, Jr. Chair in Lawyering and the Legal Process.*
KAREN L. ENGLE, B.A., J.D.; *Minerva House Drysdale Regents Chair in Law, Co-Director of Bernard and Audre Rapoport Center for Human Rights and Justice.*
KENNETH FLAMM, Ph.D.; *Professor.*
JULIUS G. GETMAN, B.A., LL.B., LL.M.; *Earl E. Sheffield Regents Chair.*
JOHN M. GOLDEN, A.B., J.D., Ph.D.; *Loomer Family Professor in Law.*
STEVEN GOODE, B.A., J.D.; *W. James Kronzer Chair in Trial and Appellate Advocacy, University Distinguished Teaching Professor.*
LINO A. GRAGLIA, B.A., LL.B.; *A.W. Walker Centennial Chair in Law.*
CHARLES G. GROAT, B.A., M.S., Ph.D.; *Professor.*
PATRICIA I. HANSEN, A.B., M.P.A., J.D.; *J. Waddy Bullion Professor.*
HENRY T. C. HU, B.S., M.A., J.D.; *Allan Shivers Chair in the Law of Banking and Finance.*
BOBBY R. INMAN, B.A.; *Professor.*
DEREK P. JINKS, B.A., M.A., J.D.; *The Marrs McLean Professor in Law.*
STANLEY M. JOHANSON, B.S., LL.B., LL.M.; *James A. Elkins Centennial Chair in Law, University Distinguished Teaching Professor.*
CALVIN H. JOHNSON, B.A., J.D.; *Andrews & Kurth Centennial Professor.*
EMILY E. KADENS, B.A., M.A., Dipl. M.A., Ph.D., J.D.; *Baker and Botts Professor in Law.*
SUSAN R. KLEIN, B.A., J.D.; *Alice McKean Young Regents Chair in Law.*

SANFORD V. LEVINSON, A.B., Ph.D., J.D.; *W. St. John Garwood & W. St. John Garwood, Jr. Centennial Chair in Law, Professor of Government.*

VIJAY MAHAJAN, M.S.Ch.E., Ph.D.; *Professor.*

BASIL S. MARKESINIS, LL.B., LL.D., D.C.L., Ph.D.; *Jamail Regents Chair.*

INGA MARKOVITS, LL.M.; *"The Friends of Joe Jamail" Regents Chair.*

RICHARD S. MARKOVITS, B.A., LL.B., Ph.D.; *John B. Connally Chair.*

THOMAS O. MCGARITY, B.A., J.D.; *Joe R. & Teresa Lozano Long Endowed Chair in Administrative Law.*

STEVEN A. MOORE, B.A., Ph.D.; *Professor.*

LINDA S. MULLENIX, B.A., M. Phil., J.D., Ph.D.; *Morris & Rita Atlas Chair in Advocacy.*

STEVEN P. NICHOLS, B.S.M.E., M.S.M.E., J.D., Ph.D.; *Professor.*

ROBERT J. PERONI, B.S.C., J.D., LL.M.; *The Fondren Foundation Centennial Chair for Faculty Excellence.*

H. W. PERRY, JR., B.A., M.A., Ph.D.; *Associate Professor of Law and Government.*

LUCAS A. POWE, JR., B.A., J.D.; *Anne Green Regents Chair in Law, Professor of Government.*

WILLIAM C. POWERS, JR., B.A., J.D.; *President of The University of Texas at Austin, Hines H. Baker & Thelma Kelley Baker Chair, University Distinguished Teaching Professor.*

DAVID M. RABBAN, B.A., J.D.; *Dahr Jamail, Randall Hage Jamail & Robert Lee Jamail Regents Chair, University Distinguished Teaching Professor.*

ALAN S. RAU, B.A., LL.B.; *Mark G. & Judy G. Yudof Chair in Law.*

DAVID W. ROBERTSON, B.A., LL.B., LL.M., J.S.D.; *W. Page Keeton Chair in Tort Law, University Distinguished Teaching Professor.*

JOHN A. ROBERTSON, A.B., J.D.; *Vinson & Elkins Chair.*

WILLIAM M. SAGE, A.B., M.D., J.D.; *Vice Provost for Health Affairs, James R. Dougherty Chair for Faculty Excellence.*

LAWRENCE G. SAGER, B.A., LL.B.; *Alice Jane Drysdale Sheffield Regents Chair.*

JOHN J. SAMPSON, B.B.A., LL.B.; *William Benjamin Wynne Professor.*

CHARLES M. SILVER, B.A., M.A., J.D.; *Roy W. & Eugenia C. MacDonald Endowed Chair in Civil Procedure, Professor of Government, Co-Director of Center on Lawyers, Civil Justice, and the Media.*

ERNEST E. SMITH, B.A., LL.B.; *Rex G. Baker Centennial Chair in Natural Resources Law.*

JAMES C. SPINDLER, B.A., M.A., J.D., Ph.D.; *The Sylvan Lang Professor.*

MATTHEW L. SPITZER, B.A., Ph.D., J.D.; *Hayden W. Head Regents Chair for Faculty Excellence.*

JANE STAPLETON, B.S., Ph.D., LL.B., D.C.L., D. Phil.; *Ernest E. Smith Professor.*

JORDAN M. STEIKER, B.A., J.D.; *Judge Robert M. Parker Endowed Chair in Law.*

MICHAEL F. STURLEY, B.A., J.D.; *Fannie Coplin Regents Chair.*

GERALD TORRES, A.B., J.D., LL.M.; *Bryant Smith Chair in Law.*

GREGORY J. VINCENT, B.A., J.D., Ed.D.; *Professor, Vice President for Diversity and Community Engagement.*

WENDY E. WAGNER, B.A., M.E.S., J.D.; *Joe A. Worsham Centennial Professor.*

LOUISE WEINBERG, A.B., J.D., LL.M.; *William B. Bates Chair for the Administration of Justice.*

OLIN G. WELLBORN, A.B., J.D.; *William C. Liedtke, Sr. Professor.*

JAY L. WESTBROOK, B.A., J.D.; *Benno C. Schmidt Chair of Business Law.*

ABRAHAM L. WICKELGREN, A.B., Ph.D., J.D.; *Bernard J. Ward Professor in Law.*

ZIPPORAH B. WISEMAN, B.A., M.A., LL.B.; *Thos. H. Law Centennial Professor.*

PATRICK WOOLLEY, A.B., J.D.; *Beck, Redden & Secrest Professor in Law.*

ASSISTANT PROFESSORS

MARILYN ARMOUR, B.A., M.S.W., Ph.D.

DANIEL M. BRINKS, A.B., J.D., Ph.D.

JUSTIN DRIVER, B.A., M.A., M.A., J.D.

ZACHARY S. ELKINS, B.A., M.A., Ph.D.

JOSEPH R. FISHKIN, B.A., M. Phil., D. Phil., J.D.

CARY C. FRANKLIN, B.A., M.S.T., D. Phil., J.D.

MIRA GANOR, B.A., M.B.A., LL.B., LL.M., J.S.D.

JENNIFER E. LAURIN, B.A., J.D.

ANGELA K. LITWIN, B.A., J.D.

MARY ROSE, A.B., M.A., Ph.D.

SEAN H. WILLIAMS, B.A., J.D.

SENIOR LECTURERS, WRITING LECTURERS, AND CLINICAL PROFESSORS

ALEXANDRA W. ALBRIGHT, B.A., J.D.; *Senior Lecturer.*

WILLIAM P. ALLISON, B.A., J.D.; *Clinical Professor, Director of Criminal Defense Clinic.*

MARJORIE I. BACHMAN, B.S., J.D.; *Clinical Instructor.*

PHILIP C. BOBBITT, A.B., J.D., Ph.D.; *Distinguished Senior Lecturer.*

KAMELA S. BRIDGES, B.A., B.J., J.D.; *Lecturer.*

CYNTHIA L. BRYANT, B.A., J.D.; *Clinical Professor, Director of Mediation Clinic.*

JOHN C. BUTLER, B.B.A., Ph.D.; *Clinical Associate Professor.*

MARY R. CROUTER, A.B., J.D.; *Lecturer, Assistant Director of William Wayne Justice Center for Public Interest Law.*

TIFFANY J. DOWLING, B.A., J.D.; *Clinical Instructor, Director of Actual Innocence Clinic.*

LORI K. DUKE, B.A., J.D.; *Clinical Professor.*

ARIEL E. DULITZKY, J.D., LL.M.; *Clinical Professor, Director of Human Rights Clinic.*

ELANA S. EINHORN, B.A., J.D.; *Lecturer.*

TINA V. FERNANDEZ, A.B., J.D.; *Lecturer, Director of Pro Bono Program.*

LYNDA E. FROST, B.A., M.Ed., J.D., Ph.D.; *Clinical Associate Professor.*

DENISE L. GILMAN, B.A., J.D.; *Clinical Professor, Co-Director of Immigration Clinic.*

KELLY L. HARAGAN, B.A., J.D.; *Lecturer, Director of Environmental Law Clinic.*

BARBARA HINES, B.A., J.D.; *Clinical Professor, Co-Director of Immigration Clinic.*

HARRISON KELLER, B.A., M.A., Ph.D.; *Vice Provost for Higher Education Policy, Senior Lecturer.*

JEANA A. LUNGWITZ, B.A., J.D.; *Clinical Professor, Director of Domestic Violence Clinic.*

TRACY W. MCCORMACK, B.A., J.D.; *Lecturer, Director of Advocacy Programs.*

ROBIN B. MEYER, B.A., M.A., J.D.; *Lecturer.*

RANJANA NATARAJAN, B.A., J.D.; *Clinical Professor, Director of National Security Clinic.*

JANE A. O'CONNELL, B.A., M.S., J.D.; *Lecturer, Deputy Director of Tarlton Law Library Public Services.*
 ROBERT C. OWEN, A.B., M.A., J.D.; *Clinical Professor.*
 SEAN J. PETRIE, B.A., J.D.; *Lecturer.*
 WAYNE SCHIESS, B.A., J.D.; *Senior Lecturer, Director of Legal Writing.*

STACY ROGERS SHARP, B.S., J.D.; *Lecturer.*

PAMELA J. SIGMAN, B.A., J.D.; *Adjunct Professor, Director of Juvenile Justice Clinic.*

DAVID S. SOKOLOV, B.A., M.A., J.D., M.B.A.; *Distinguished Senior Lecturer, Director of Student Life.*
 LESLIE L. STRAUCH, B.A., J.D.; *Clinical Professor.*
 GRETCHEN S. SWEEN, B.A., M.A., Ph.D., J.D.; *Lecturer.*
 MELINDA E. TAYLOR, B.A., J.D.; *Senior Lecturer, Executive Director of Center for Global Energy, International Arbitration, & Environmental Law.*
 HEATHER K. WAY, B.A., B.J., J.D.; *Lecturer, Director of Community Development Clinic.*
 ELIZABETH M. YOUNGDALE, B.A., M.L.I.S., J.D.; *Lecturer.*

ADJUNCT PROFESSORS AND OTHER LECTURERS

ELIZABETH AEBERSOLD, B.A., M.S.
 WILLIAM R. ALLENSWORTH, B.A., J.D.
 CRAIG D. BALL, B.A., J.D.
 SHARON C. BAXTER, B.S., J.D.
 KARL O. BAYER, B.A., M.S., J.D.
 WILLIAM H. BEARDALL, JR., B.A., J.D.
 JERRY A. BELL, B.A., J.D.
 ALLISON H. BENESCH, B.A., M.S.W., J.D.
 CRAIG R. BENNETT, B.S., J.D.
 JAMES B. BENNETT, B.B.A., J.D.
 MELISSA J. BERNSTEIN, B.A., M.L.S., J.D.
 RAYMOND D. BISHOP, B.A., J.D.
 MURFF F. BLEDSOE, B.A., J.D.
 WILLIAM P. BOWERS, B.B.A., J.D., LL.M.
 HUGH L. BRADY, B.A., J.D.
 STACY L. BRAININ, B.A., J.D.
 ANTHONY W. BROWN, B.A., J.D.
 JAMES E. BROWN, B.A., LL.B.
 TOMMY L. BROYLES, B.A., J.D.
 PAUL J. BURKA, B.A., LL.B.
 W.A. BURTON, JR., B.A., M.A., LL.B.
 ERIN G. BUSBY, B.A., J.D.
 AGNES E. CASAS, B.A., J.D.
 RUBEN V. CASTANEDA, B.A., J.D.
 EDWARD A. CAVAZOS, B.A., J.D.
 JEFF CIVINS, A.B., M.S., J.D.
 LEIF M. CLARK, B.A., J.D.
 ELIZABETH COHEN, B.A., M.S.W., J.D.
 JAMES W. COLLINS, B.S., J.D.
 PATRICIA J. CUMMINGS, B.A., J.D.
 KEITH B. DAVIS, B.S., J.D.
 DICK DEGUERIN, B.A., LL.B.
 RICHARD D. DEUTSCH, B.A., B.A., J.D.
 STEVEN K. DEWOLF, B.A., J.D., LL.M.
 REBECCA H. DIFFEN, B.A., J.D.
 PHILIP DURST, B.A., M.A., J.D.
 BILLIE J. ELLIS, JR., B.A., M.B.A., J.D.
 JAY D. ELLWANGER, B.A., J.D.
 EDWARD Z. FAIR, B.A., M.S.W., J.D.
 JOHN C. FLEMING, B.A., J.D.
 KYLE K. FOX, B.A., J.D.
 DAVID C. FREDERICK, B.A., Ph.D., J.D.
 GREGORY D. FREED, B.A., J.D.
 FRED J. FUCHS, B.A., J.D.
 CHARLES E. GHOLZ, B.S., Ph.D.
 MICHAEL J. GOLDEN, A.B., J.D.
 DAVID HALPERN, B.A., J.D.
 ELIZABETH HALUSKA-RAUSCH, B.A., M.A., M.S., Ph.D.
 JETT L. HANNA, B.B.A., J.D.
 CLINT A. HARBOUR, B.A., J.D., LL.M.
 ROBERT L. HARGETT, B.B.A., J.D.
 MARY L. HARRELL, B.S., J.D.
 JAMES C. HARRINGTON, B.A., M.A., J.D.
 CHRISTOPHER S. HARRISON, Ph.D., J.D.
 JOHN R. HAYS, JR., B.A., J.D.
 P. MICHAEL HEBERT, A.B., J.D.
 STEVEN L. HIGHLANDER, B.A., Ph.D., J.D.
 SUSAN J. HIGHTOWER, B.A., M.A., J.D.
 KENNETH E. HOUP, JR., J.D.

RANDY R. HOWRY, B.J., J.D.
 MONTY G. HUMBLE, B.A., J.D.
 JEFF JURY, B.A., J.D.
 PATRICK O. KEEL, B.A., J.D.
 DOUGLAS L. KEENE, B.A., M.Ed., Ph.D.
 CHARI L. KELLY, B.A., J.D.
 ROBERT N. KEPPLER, B.A., J.D.
 MARK L. KINCAID, B.B.A., J.D.
 AMI L. LARSON, B.A., J.D.
 JODI R. LAZAR, B.A., J.D.
 KEVIN L. LEAHY, B.A., J.D.
 DAVID P. LEIN, B.A., M.P.A., J.D.
 MAURIE A. LEVIN, B.A., J.D.
 ANDRES J. LINETZKY, LL.M.
 JAMES-LOYD LOFTIS, B.B.A., J.D.
 JIM MARCUS, B.A., J.D.
 HARRY S. MARTIN, A.B., M.L.S., J.D.
 FRANCES L. MARTINEZ, B.A., J.D.
 LAURA A. MARTINEZ, B.A., J.D.
 RAY MARTINEZ, III, B.A., J.D.
 LISA M. MCCLEAIN, B.A., J.D., LL.M.
 BARRY F. MCNEIL, B.A., J.D.
 ANGELA T. MELINARAAB, B.F.A., J.D.
 MARGARET M. MENICUCCI, B.A., J.D.
 JO A. MERICA, B.A., J.D.
 RANELLE M. MERONEY, B.A., J.D.
 ELIZABETH N. MILLER, B.A., J.D.
 JONATHAN F. MITCHELL, B.A., J.D.
 DARYL L. MOORE, B.A., M.L.A., J.D.
 EDWIN G. MORRIS, B.S., J.D.
 SARAH J. MUNSON, B.A., J.D.
 MANUEL H. NEWBURGER, B.A., J.D.
 DAVID G. NIX, B.S.E., LL.M., J.D.
 PATRICK L. O'DANIEL, B.B.A., J.D.
 M.A. PAYAN, B.A., J.D.
 MARK L. PERLMUTTER, B.S., J.D.
 ELIZA T. PLATTS-MILLS, B.A., J.D.
 JONATHAN PRATTER, B.A., M.L.S., J.D.
 VELVA L. PRICE, B.A., J.D.
 BRIAN C. RIDER, B.A., J.D.
 ROBERT M. ROACH, JR., B.A., J.D.
 BRIAN J. ROARK, B.A., J.D.
 BETTY E. RODRIGUEZ, B.S.W., J.D.
 JAMES D. ROWE, B.A., J.D.
 MATTHEW C. RYAN, B.A., J.D.
 KAREN R. SAGE, B.A., J.D.
 MARK A. SANTOS, B.A., J.D.
 MICHAEL J. SCHLESS, B.A., J.D.
 AMY J. SCHUMACHER, B.A., J.D.
 SUZANNE SCHWARTZ, B.J., J.D.
 RICHARD J. SEGURA, JR., B.A., J.D.
 DAVID A. SHEPPARD, B.A., J.D.
 HON. ERIC M. SHEPPERD, B.A., J.D.
 RONALD J. SIEVERT, B.A., J.D.
 AMBROSIO A. SILVA, B.S., J.D.
 STUART R. SINGER, A.B., J.D.
 HON. BEA A. SMITH, B.A., M.A., J.D.
 LYDIA N. SOLIZ, B.B.A., J.D.
 STEPHEN M. SONNENBERG, A.B., M.D.

JAMES M. SPELLINGS, JR., B.S., J.D.
DAVID B. SPENCE, B.A., J.D., M.A., Ph.D.
KACIE L. STARR, B.A., J.D.
WILLIAM F. STUTTS, B.A., J.D.
MATTHEW J. SULLIVAN, B.S., J.D.
JEREMY S. SYLESTINE, B.A., J.D.
BRADLEY P. TEMPLE, B.A., J.D.
SHERINE E. THOMAS, B.A., J.D.
TERRY O. TOTTENHAM, B.S., LL.M., J.D.
MICHAEL S. TRUESDALE, B.A., M.A., J.D.
JEFFREY K. TULIS, B.A., M.A., Ph.D.
TIMOTHY J. TYLER, B.A., J.D.
SUSAN S. VANCE, B.B.A., J.D.
LANA K. VARNEY, B.J., J.D.
SRIRAM VISHWANATH, B.S., M.S., Ph.D.
DEBORAH M. WAGNER, B.A., M.A., J.D.

CLARK C. WATTS, B.A., M.D., M.A., M.S., J.D.
WARE V. WENDELL, A.B., J.D.
RODERICK E. WETSEL, B.A., J.D.
THEA WHALEN, B.A., J.D.
DARA J. WHITEHEAD, B.A., M.S.
RANDALL B. WILHITE, B.B.A., J.D.
TIMOTHY A. WILKINS, B.A., M.P.P., J.D.
DAVID G. WILLE, B.S.E.E., M.S.E.E., J.D.
ANDREW M. WILLIAMS, B.A., J.D.
MARK B. WILSON, B.A., M.A., J.D.
HON. PAUL L. WOMACK, B.S., J.D.
LUCILLE D. WOOD, B.A., J.D.
DENNEY L. WRIGHT, B.B.A., J.D., LL.M.
LARRY F. YORK, B.B.A., LL.B.
DANIEL J. YOUNG, B.A., J.D.

VISITING PROFESSORS

OWEN L. ANDERSON, B.A., J.D.
ANTONIO H. BENJAMIN, LL.B., LL.M.
PETER F. CANE, B.A., LL.B., D.C.L.
JOSHUA DRESSLER, B.A., J.D.
ROBIN J. EFFRON, B.A., J.D.

VICTOR FERRERES, J.D., LL.M., J.S.D.
PETER M. GERHART, B.A., J.D.
LARRY LAUDAN, B.A., M.A., Ph.D.
GRAHAM B. STRONG, B.A., J.D., LL.M.

Texas Law Review

Volume 91

Number 2

December 2012

PARTH S. GEJI
Editor in Chief

BENJAMIN S. MORGAN
Managing Editor

MOLLY M. BARRON
Chief Articles Editor

AMELIA A. FRIEDMAN
Administrative Editor

LAUREN K. ROSS
Chief Notes Editor

RALPH C. MAYRELL
Book Review Editor

LISA D. KINZER
Chief Online Content Editor

TYSON M. LIES
Research Editor

BRITTANY R. ARTIMEZ
ALESE L. BAGDOL
WILLIAM P. COURTNEY
MONICA E. GAUDIOSO
Articles Editors

ALEXANDER G. HUGHES
Managing Online Content Editor

WILLIAM J. MCKINNON
MARTHA L. TODD
COLIN M. WATTERSON
COLLIN R. WHITE
Articles Editors

MONICA R. HUGHES
ROSS M. MACDONALD
MICHAEL N. SELKIRK
Notes Editors

MICHAEL ABRAMS
BRIAN J. BAH
JULIA C. BARRETT
BRADEN A. BEARD
DAWSON A. BROTEMARKLE

ERIN L. GAINES
DANIEL D. GRAVER
JONATHAN LEVY
NATHANIEL H. LIPANOVICH
KYLE E. MITCHELL
CORBIN D. PAGE
Associate Editors

CHRISTOPHER S. PATTERSON
ADAM R. PERKINS
KATHRYN G. RAWLINGS
STEPHEN STECKER
JAMES T. WEISS

Members

MICHELLE K. ARISHITA
KATHRYN W. BAILEY
CECILIA BERNSTEIN
MICHAEL R. BERNSTEIN
MATTHEW J. BRICKER
CAITLIN A. BUBAR
KRISTIN C. BURNETT
CHARLES D. CASSIDY
SAMANTHA CHEN
CHASE E. COOLEY
JASON A. DANOWSKY
MICHAEL C. DEANE
MARIE E. DELAHOUSAYE
DAVID D. DOAK
ALLISON L. FULLER
REBECCA L. GIBSON
JOSHUA S. GOLD
SEAN M. HILL
ALEXANDRA C. HOLMES
ROBERT P. HUGHES

LAURA C. INGRAM
SAMUEL F. JACOBSON
HANNAH L. JENKINS
COURTNEY H. JOHNSON
ELIZABETH M. JOHNSON
MICHAEL C. KELSO
MELANIE M. KISER
JEFFREY P. KITCHEN
KELSIE A. KRUEGER
ARIELLE K. LINSEY
JONATHAN D. LIROFF
ROCCO F. MAGNI
YANIV M. MAMAN
THOMAS K. MATHEW
DINA W. MCKENNEY
RYAN E. MELTZER
JOHN K. MORRIS
JACOB MOSS
DAVID A. NIEDRAUER

MARTIN OBERST
MATTHEW M. OLSON
JACKSON A. O'MALEY
SPENCER P. PATTON
JAMES D. PETERS
CHRISTA G. POWERS
JAMES R. POWERS
JENNIFER N. RAINEY
ALEZA S. REMIS
AMANDA D. ROBERTS
A. ELIZABETH ROMEFELT
BRETT S. ROSENTHAL
BRENT M. RUBIN
JONATHAN E. SARNA
JOHN W. STRIBLING
WILLIAM C. VAUGHN
VINCENT M. WAGNER
LECH K. WILKIEWICZ
JAMIE L. YARBROUGH
E. ALEXINE ZACARIAS

PAUL N. GOLDMAN
Business Manager

MITCHELL N. BERMAN
JOHN S. DZIENKOWSKI
Faculty Advisors

TERI GAUS
Editorial Assistant

Texas Law Review

Volume 91, Number 2, December 2012

ARTICLES

- Exchanging Information Without Intellectual Property
Michael J. Burstein 227
- Solving the Patent Settlement Puzzle
Einer Elhauge & Alex Krueger 283

BOOK REVIEWS

- Henry Friendly: The Judge, the Man, the Book
Mary Coombs 331
- On Becoming a Great Judge: The Life of Henry J. Friendly
Frederick T. Davis 339
- Henry Friendly: As Brilliant as Expected but Less Predictable
Peter Edelman 345
- all reviewing* David M. Dorsen's
HENRY FRIENDLY: GREATEST JUDGE OF HIS ERA
- Assembly Resurrected
Ashutosh A. Bhagwat 351
- Recovering the Assembly Clause
Timothy Zick 375
- both reviewing* John D. Inazu's
LIBERTY'S REFUGE: THE FORGOTTEN FREEDOM OF
ASSEMBLY
- What Is the Essential Fourth Amendment?
Christopher Slobogin 403
- reviewing* Stephen J. Schulhofer's
MORE ESSENTIAL THAN EVER: THE FOURTH
AMENDMENT IN THE TWENTY-FIRST CENTURY

NOTES

Setting Examples, Not Settling: Toward a New SEC
Enforcement Paradigm

Ross MacDonald

419

Blowing the Whistle on Civil Rights: Analyzing the False
Claims Act as an Alternative Enforcement Method for Civil
Rights Laws

Ralph C. Mayrell

449

Articles

Exchanging Information Without Intellectual Property

Michael J. Burstein^{*}

Contracting over information is notoriously difficult. Nearly fifty years ago, Kenneth Arrow articulated a “fundamental paradox” that arises when two parties try to exchange information. To complete such a transaction, the buyer of information must be able to place a value on the information. But once the seller discloses the information, the buyer can take it without paying. The conventional solution to this disclosure paradox is intellectual property. If the information is protected by a patent or a copyright, then the seller can disclose the information free in the knowledge that the buyer can be enjoined against making, using, or selling it without permission. This account of information exchange forms the basis for an increasingly popular argument in favor of strong and broad intellectual property rights for the purpose of overcoming the disclosure paradox and thereby facilitating the development and commercialization of ideas.

That argument, however, rests on assumptions about the nature of information that are neither theoretically nor empirically justified. This Article explains that, contrary to the conventional account of the disclosure paradox, information is not always nonexcludable and is not always a homogeneous asset. Instead, information is complex and multifaceted, subject to some inherent limitations but also manipulable by its holders. These characteristics give rise to a range of strategies for engaging in information exchange, of which intellectual property is only one. Information holders can use the characteristics of information itself as well as contractual and norms-based mechanisms and other legal or business strategies to achieve exchange. And examples drawn from fields as diverse and disparate as software and biotechnology show that entrepreneurs and inventors use these strategies alone or in combination to

^{*} Assistant Professor of Law, Benjamin N. Cardozo School of Law, Yeshiva University. I thank Yochai Benkler, Rachel Barkow, Margaret Chon, Kevin Collins, Rochelle Dreyfuss, Terry Fisher, Brett Frischmann, Justin Hughes, Mark Lemley, Daryl Levinson, Fiona Murray, Ben Roin, Christopher Sprigman, Susannah Tobin, and Melissa Wasserman for helpful comments and conversations. I am also grateful to participants in the 2012 Gruter Institute Conference on Law and Human Behavior, the 2012 Patent Conference at Boston College Law School, the “Law and Entrepreneurship Mini-Conference” at the 2011 Law & Society Association Annual Meeting, and workshop participants at Cardozo, Drexel, Harvard, Illinois, Nebraska, Ohio State, Penn State, St. John’s, Stanford, Utah, and Virginia for valuable discussions. Rachel Sachs and Danielle Shultz provided outstanding research assistance. Several examples in this Article are drawn from a small number of pilot field interviews with entrepreneurs and investors in the Boston area who have asked to remain anonymous. I am grateful to these individuals for sharing their insights.

effectively link their ideas with capital and development skills, often without intellectual property appearing to play a significant role in the transaction.

Intellectual property is therefore not necessary to promote robust markets for information and is, in fact, just as contingent and context-specific a solution to the paradox as the alternatives described here. At the very least, then, there is reason to doubt that commercialization theories founded upon information exchange provide a stand-alone justification for intellectual property. This Article urges caution in policy interventions that seek to respond to the disclosure paradox and sets the stage for future empirical research to better understand the dynamics of information-exchange strategies and the social welfare costs and benefits that may accompany them.

Introduction.....	228
I. The Conventional Account of Intellectual Property and Information Exchange.....	235
A. The Commercialization Imperative	237
B. Commercialization and Information Exchange	241
C. Questioning Commercialization Theory.....	246
II. A New Framework for Understanding and Overcoming the Disclosure Paradox	247
A. The Economics of Information Goods.....	247
1. <i>Excludability</i>	248
2. <i>Heterogeneity</i>	255
B. Alternative Solutions to the Disclosure Paradox	258
1. <i>Intellectual Property</i>	258
2. <i>Contracts</i>	262
3. <i>Norms</i>	267
4. <i>Alternative Sources of Appropriability</i>	270
III. Using Policy Tools to Promote Information Exchange	274
A. Costs and Benefits.....	276
B. Dynamic Interactions.....	279
C. The Need for Empirical Research.....	280
Conclusion	282

Introduction

Contracting over information is notoriously difficult. Fifty years ago, Kenneth Arrow articulated a “fundamental paradox” that arises when two parties try to exchange information.¹ In order to complete such a transaction,

1. Kenneth J. Arrow, *Economic Welfare and the Allocation of Resources for Invention*, in THE RATE AND DIRECTION OF INVENTIVE ACTIVITY: ECONOMIC AND SOCIAL FACTORS 609, 615 (Nat'l Bureau of Econ. Research ed., 1962).

the buyer of information must be able to place a value on the information and determine how much she is willing to pay.² But once the seller discloses the information, the buyer is in possession of the subject of the trade and no longer has any reason to pay for it.³ This problem has come to be known as the “disclosure paradox” or the “information paradox.”⁴ The conventional legal solution to the paradox is a grant of intellectual property rights.⁵ If information is subject to a patent or a copyright, then it can be disclosed without fear that it will be taken without compensation. Any potential buyer who tries to make, use, or sell the information without permission can be enjoined against doing so through legal process.⁶

This account of information exchange forms the basis for an increasingly popular argument in favor of broad and strong intellectual property rights. That argument proceeds roughly as follows: Exchanging information is critical to innovation because the initial act of creation or invention is only the first step in bringing a product to market.⁷ Inventors must usually recruit capital and partners with the skills to develop and then to commercialize their inventions.⁸ If the disclosure paradox interferes with entrepreneurs’ ability to contract for capital or other resources, and intellectual property solves the disclosure paradox, then the scope of

2. *Id.*

3. *Id.*

4. See, e.g., Shyamkrishna Balganesh, “Hot News”: *The Enduring Myth of Property in News*, 111 COLUM. L. REV. 419, 433 (2011) (describing “Arrow’s information paradox” wherein “[a] potential licensee has no way of evaluating the information/intangible until it is disclosed to him; yet, upon such disclosure he has little reason to want to pay for it”); Jonathan M. Barnett, *Intellectual Property as a Law of Organization*, 84 S. CAL. L. REV. 785, 794 (2011) (“Arrow drew attention to this sensitive juncture—postinvention but precommercialization—by describing a dilemma that has since become known as ‘Arrow’s paradox’ or the ‘disclosure paradox.’”); Margaret Chon, *Sticky Knowledge and Copyright*, 2011 WIS. L. REV. 177, 198 (noting “the typical assumption of Arrow’s information-disclosure paradox: that is, the problem is that knowledge is not easily disclosed”); Mark A. Lemley, *The Myth of the Sole Inventor*, 110 MICH. L. REV. 709, 748 (2012) (“Arrow’s Information Paradox suggests that parties may find it difficult to contract to disclose information in the absence of a property right over that information.”). Cooter and Edlin refer to the phenomenon as the “double trust dilemma.” Robert D. Cooter & Aaron Edlin, *Law and Growth Economics: A Framework for Research* 16 (Berkeley Program in Law & Econ., Working Paper Series, 2011), available at <http://escholarship.org/uc/item/50t4d0kt>.

5. See, e.g., Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265, 277–78 (1977) (“The patent creates a defined set of legal rights known to both parties at the outset of negotiations. . . . [T]he owner can [therefore] disclose such information protected by the scope of the legal monopoly.”); Robert P. Merges, *A Transactional View of Property Rights*, 20 BERKELEY TECH. L.J. 1477, 1485 (2005) (arguing that parties may rely on property rights to solve the problem created by the disclosure paradox because property rights “operate[] effectively even when contracts are difficult to enforce”). Throughout this Article I use the term “intellectual property” to refer to the legal conferral of exclusive rights over information. I exclude from this definition the underlying substance of the information protected by those rights.

6. 17 U.S.C. § 502 (2006); 35 U.S.C. § 283 (2006); *eBay Inc. v. MercExchange, L.L.C.*, 547 U.S. 388, 391–93 (2006).

7. See *infra* notes 35–41 and accompanying text.

8. See, e.g., Ted Sichelman, *Commercializing Patents*, 62 STAN. L. REV. 341, 348–54 (2010) (providing an example of the commercialization process in the software industry).

intellectual property should expand to encompass whatever information will be socially valuable to exchange. Indeed, although the traditional justification for intellectual property is that it provides necessary incentives for new works of invention or creation,⁹ an increasing number of theorists focus on the commercialization of those products as a stand-alone justification for intellectual property.¹⁰

There can be little doubt that commercialization is of critical importance to innovation and economic growth.¹¹ Facilitating linkages between creators or inventors and potential sources of development, improvement, and capital is increasingly being recognized as an important policy lever for promoting innovation.¹² But reaching even the narrow conclusion that intellectual property may help join ideas and capital by solving the disclosure paradox in some circumstances requires a more thorough understanding of the disclosure paradox and the range of potential solutions that parties may employ to overcome it than the literature currently offers. This Article explores the paradox and its potential solutions in detail, a necessary first step toward validating both descriptive and normative accounts of the role of intellectual property in information exchange, and it casts doubt on commercialization theory as a stand-alone justification for expanding intellectual property.

9. See, e.g., Mark A. Lemley, *Ex Ante Versus Ex Post Justifications for Intellectual Property*, 71 U. CHI. L. REV. 129, 129–30 (2004) (drawing a distinction between “traditional economic justification” and “new justifications . . . focus[ed] not on the incentive to create new ideas, but on what happens to those ideas after they have been developed”); *infra* notes 23–28 and accompanying text (citing economics literature).

10. See, e.g., Michael Abramowicz, *The Danger of Underdeveloped Patent Prospects*, 92 CORNELL L. REV. 1065, 1065 (2007) (arguing that current patent law may not protect inventions long enough to make commercialization attractive and proposing an auction system to extend patents to remedy this deficiency); Kitch, *supra* note 5, at 275–80 (justifying “the need for a system of property rights in information” by “a scarcity of resources that may be employed to use information” rather than by lack of incentives to generate information (emphasis added)); F. Scott Kieff, *Property Rights and Property Rules for Commercializing Inventions*, 85 MINN. L. REV. 697, 705 n.27 (2001) (“This Article offers a view of the patent system that is tied to commercialization, rather than to inventing.”); Sichelman, *supra* note 8, at 341 (arguing that traditional patent rights fail to encourage substantial commercialization of inventions and proposing a new “commercialization” patent to rectify this defect).

11. See OFFICE OF TECH. ASSESSMENT, U.S. CONG., *INNOVATION AND COMMERCIALIZATION OF EMERGING TECHNOLOGIES* iii (1995), available at <http://www.fas.org/ota/reports/9539.pdf> (“Technological innovation is essential to the future well-being of the United States. The ability of the nation to sustain economic growth . . . depends, in many ways, on its success in developing and commercializing new products, processes, and services.”); Robert Cooter et al., *The Importance of Law in Promoting Innovation and Growth*, in *RULES FOR GROWTH: PROMOTING INNOVATION AND GROWTH THROUGH LEGAL REFORM* 1, 4 (Kauffman Found. ed., 2011) (arguing that research and development spending is not likely to translate into new production and thus economic growth without commercialization); Cooter & Edlin, *supra* note 4, at 14 (“Newly discovered ideas seldom have economic value until they are developed . . .”).

12. Cooter and Edlin, for example, place the development of innovations at the core of their theory of law and growth economics. In their view, “[m]inimizing the double trust problem”—the disclosure paradox—“is central to increasing the pace of innovation.” Cooter & Edlin, *supra* note 4, at 17.

More specifically, I demonstrate that the conventional account of the disclosure paradox and its legal solution rests on assumptions that are neither theoretically nor empirically justified. It is based on a stylized model of information that does not reflect the reality of the economic good that parties seek to exchange. And it largely ignores the possibility that alternative mechanisms for facilitating information exchange exist and may present a different social welfare calculus than intellectual property. Drawing on the literatures in management, information science, and law, I develop a framework for evaluating the range of potential solutions to the disclosure paradox and populate that framework with examples of such solutions in operation.¹³ I conclude that proponents of a commercialization theory of intellectual property that is focused on the costs of information exchange consistently underappreciate the range of potential strategies by which parties may enable commercially significant exchange and the ways in which those strategies interact within complex business, cultural, and legal environments.¹⁴

There may be situations where intellectual property is both an effective and the optimal means to facilitate the exchange of valuable information, but there are also circumstances in which either or neither condition will obtain. Intellectual property should be the preferred solution to the disclosure paradox only when it is the best among alternatives. The social welfare costs and benefits of intellectual property must therefore be compared on a case-by-case basis with the costs and benefits of other available solutions. At the very least, our policy discourse ought not to begin with intellectual property as a default rule. Because intellectual property is only one of a number of highly contingent potential solutions to the disclosure paradox, I urge caution

13. A note on methodology is appropriate here. My argument in this Article is largely theoretical. I draw examples from the existing literature and from a small number of pilot interviews solely to demonstrate that the alternative strategies I describe as a matter of theory actually exist in practice. My examples are offered as “proof of concept” rather than as support for conclusions about the prevalence or frequency with which any particular strategy for exchanging information is employed. That is the subject of my next article.

14. Indeed, most discussion of the paradox in the legal literature is limited to an acknowledgment that it exists and that it may be solved through intellectual property. See, e.g., CRAIG ALLEN NARD, *THE LAW OF PATENTS* 27 (2008) (“Absent a property right, the inventor will likely be reticent to disclose information for fear of inducing competition.”); Oren Bar-Gill & Gideon Parchomovsky, *Law and the Boundaries of Technology-Intensive Firms*, 157 U. PA. L. REV. 1649, 1654 (2009) (“Absent legal protection, the information holder is in a bind: in order to sell the information, she must disclose it to the potential buyer, but once she does, she has nothing left to sell.”); Dan L. Burk & Brett H. McDonnell, *The Goldilocks Hypothesis: Balancing Intellectual Property Rights at the Boundary of the Firm*, 2007 U. ILL. L. REV. 575, 585 (“By publicly disclosing technical information, while protecting it by exclusivity, patents circumvent the Arrow paradox.”); Paul J. Heald, *A Transaction Costs Theory of Patent Law*, 66 OHIO ST. L.J. 473, 475 n.16 (2005); Christopher S. Yoo, *Copyright and Public Good Economics: A Misunderstood Relation*, 155 U. PA. L. REV. 635, 658 (2007) (“[G]iving follow-on authors a degree of copyright protection offers a solution to Arrow’s information paradox.”). Jonathan Barnett acknowledges the possibility that other mechanisms exist that may solve the disclosure paradox, but he does not give them significant weight. Barnett, *supra* note 4, at 800–02.

in policy interventions that seek to promote markets in information and set the stage for further empirical research to shed light on when one or another such intervention may be appropriate.

Consider the following example:¹⁵ Biotechnology companies (biotechs) specialize in early-stage research and development of pharmaceuticals. Large-scale clinical testing and manufacturing of pharmaceuticals, however, requires the skills and financial resources of a large pharmaceutical company.¹⁶ It is very common, therefore, for biotechs to seek to license the compounds that they have under development. Information must be exchanged in order for these transactions to take place. The two parties must identify one another as possessing mutually beneficial products or skills. They must then learn enough about one another's products or skills to set the terms of the licensing arrangement.

In these negotiations, a biotech often will approach a potential development and commercialization partner and give an informal presentation about the compound it is developing. In this presentation, the biotech will disclose some data about the compound: the therapeutic area and potential market, the biological targets with which the compound interacts, the compound's pharmacological characteristics, and perhaps some information gleaned from preclinical testing that is relevant to conversations about the potential business opportunity. This presentation is effectively a sales pitch. The biotech will reveal this information to multiple potential partners in search of the right fit. But the biotech will not reveal the chemical structure of the compound itself.

When two companies become interested in pursuing the opportunity further, they will enter into a confidential disclosure agreement (CDA). That agreement typically restricts each party to using the confidential information solely to evaluate whether to enter into a business relationship. With the CDA in place, the parties will engage in further disclosures. The newly disclosed information will include more closely held data about the compound's efficacy or other potential commercial advantages. Yet it will generally still not include the structure of the compound or toxicity data (i.e., information about potential problems).

As the parties move further along in their negotiations, they will sign a "term sheet" that outlines the contours of the potential business deal. They will then engage in significant further disclosures in the course of conducting due diligence. At that point, the biotech will disclose raw efficacy and toxicity data. Even here, there may be some disclosure of the structure, but that disclosure will be only to a limited number of people or a third party

15. This account is drawn from interviews with the CEO and General Counsel of a Boston-area biotech firm, as well as from a review of documentary evidence they provided.

16. For an overview of the pharmaceutical research and development process, see Benjamin N. Roin, *Unpatentable Drugs and the Standards of Patentability*, 87 TEXAS L. REV. 503, 510-11 & nn.21-23 (2009).

“clean team” that will evaluate it independently of the two parties. Finally, when the parties negotiate a contract based on the term sheet, the biotech will disclose the structure of the compound.

This example fundamentally challenges the conventional understanding of the disclosure paradox and the role that intellectual property plays in its resolution. In the classic account, the parties negotiate over the (uncertain) value of the molecule. The biotech must reveal the molecule for the parties to bargain over its commercial worth. But once the biotech discloses the structure, the pharmaceutical company can develop the molecule on its own without paying for it.¹⁷ Intellectual property is therefore thought to be of paramount importance in the pharmaceutical industry.¹⁸ Yet intellectual property is noticeably absent from the story told above, even though the setting represents one of the strongest candidates for conformity to the economic model of the disclosure paradox. That is because although the molecule is covered by a patent, that patent does not effectively protect the molecule at this stage of development. Indeed, in the early stages of pharmaceutical research, competitors may be able to design around any applicable patents. According to the conventional theory, the absence of effective patent protection means that the transaction cannot occur.¹⁹

But the transaction does occur, for several reasons. First, the biotech can disclose information *about* the compound without revealing the compound itself. That information carries significantly less risk of misappropriation yet is still commercially useful enough to form a basis for bargaining and exchange. Second, the parties rely significantly on reputation effects. Consolidation in the pharmaceutical industry has resulted in a small number of firms that have the capability to do large-scale clinical development and drug marketing. These firms compete heavily with one another for the rights to develop compounds that originate in biotech companies. As a result, their reputations as good-faith negotiating partners are critical to securing future deal flow. Third, these reputational effects are reinforced by formal contracts. CDAs are almost never litigated.²⁰ Instead, they are used as signals to the reputation market that the relationship between the two companies is becoming deeper. In the pre-CDA interactions, the biotech is responsible for protecting its own sensitive information, and the pharmaceutical company generally does not incur any reputational loss for

17. See *infra* notes 155–58 and accompanying text (describing self-disclosing characteristics of pharmaceutical products).

18. See, e.g., JAMES BESSEN & MICHAEL J. MEURER, PATENT FAILURE: HOW JUDGES, BUREAUCRATS, AND LAWYERS PUT INNOVATORS AT RISK 88–89 (2009) (“The canonical example of the free-riding problem is traditional drug development . . .”).

19. See, e.g., Bar-Gill & Parchomovsky, *supra* note 14, at 1654 (“As Kenneth Arrow famously observed, information that is not afforded legal protection cannot be bought or sold on the market.”).

20. Indeed, there are serious questions about whether nondisclosure agreements are effectively enforceable at all. See *infra* notes 188–90 and accompanying text.

the use or sharing of information disclosed in such settings. Once a CDA is signed, however, that is a signal that the firms have undertaken a heightened duty of confidentiality to one another, and a pharmaceutical firm that misappropriates information at that stage is likely to suffer reputational harm. The potential for harm is even more serious after a term sheet is signed. And a firm that cheats on a deal after contract is likely to find itself cut off from many future deals. Finally, the entire negotiation takes place against the backdrop of a significant first-mover advantage on the part of the biotech firm. Because drug development is time-consuming and expensive,²¹ a biotech company with a head start of several years is at a significant advantage. While it is true that a potential pharmaceutical company partner may be able to appropriate some of the information provided to it in the course of negotiations, as a practical matter that company would be far behind in the development process if it struck out on its own. That commercial reality provides a powerful incentive to deal rather than to defect.

This example and others described in this Article suggest that intellectual property may not be playing the role in facilitating information exchange that the conventional account of the paradox predicts. Indeed, it suggests that intellectual property may be one of several mechanisms that overlap and interact in complex ways. It highlights both the contingency of the intellectual property solution to the paradox and the utility of strategies based on information-flow design, contract, and norms.

To the extent that commercialization theory is founded upon the conventional account of the disclosure paradox, there is reason to doubt that it provides a stand-alone justification for intellectual property. At the very least, the expansion of intellectual property to facilitate exchange is likely to be justified in a far narrower range of circumstances than commercialization theorists predict. Public policy aimed at facilitating robust markets for the exchange of information goods therefore must take full account of the social welfare costs and benefits of all of the various solutions to the paradox.

My argument proceeds as follows. Part I briefly surveys and critiques “commercialization theory,” the argument that intellectual property is justified and should be strengthened on the ground that it promotes the development and commercialization of inventions or creations. On one account, this theory is effectively the classic story of incentives to invent just pushed forward in the innovation cycle. Just as intellectual property may be necessary to recoup the costs of invention, so too may it be necessary to recoup the costs of commercialization. But to the extent that commercialization theory aims at a distinct economic function, it is primarily pitched as a solution to the disclosure paradox.²² Here, the theory suffers

21. Roin, *supra* note 16, at 510–11.

22. *See infra* subpart I(B).

from an overly thin account of the problem it is trying to solve and the solution. Relying primarily on insights from the theory of the firm, commercialization theorists assume that information can be successfully propertized and therefore made into a ready product for exchange. But these insights depend on an insufficiently nuanced theory of information.

Part II begins by examining and complicating two assumptions about information that drive the conventional account of the disclosure paradox. First, information is not always nonexcludable. It has various degrees of opacity that depend in part on its inherent characteristics and in part on how information holders choose to communicate it (or not) to the world. Second, information is not homogeneous. It is not always a stock tip. Instead, it is a multilayered, continuous asset that can simultaneously communicate value in different ways.

These complex characteristics of information give rise themselves to a number of strategies for minimizing or overcoming the disclosure paradox through information-flow design. They also enable a variety of alternative approaches to the paradox. Some are based in intellectual property, while others are based in contracts, in norms of exchange, or in alternative legal or business strategies. The remainder of Part II explains why these solutions to the paradox are theoretically plausible and it offers real-world examples of each to demonstrate that information holders actually utilize them in some circumstances.

Part III draws several implications from this analysis. It argues that intellectual property is not always necessary for the exchange of information and is, in fact, just as contingent and circumstance-specific a solution to the disclosure paradox as the alternatives described in Part II. These solutions each have social welfare costs and benefits that are likely to be similarly situation-specific. Intellectual property is also likely to interact with other mechanisms in complex and overlapping ways. Indeed, if intellectual property works as an overlay on already existing disclosure strategies, then there may be a doubling up of social welfare costs without a concomitant doubling of social benefits. In all events, these tangled consequences suggest that the optimality of any particular policy solution is ultimately an empirical question.

I. The Conventional Account of Intellectual Property and Information Exchange

The traditional economic justification for intellectual property is that it provides needed incentives for the invention or creation of intellectual works.²³ Inventions or creative works require significant investment to

23. See, e.g., Peter S. Menell & Suzanne Scotchmer, *Intellectual Property Law*, in 2 *HANDBOOK OF LAW AND ECONOMICS* 1473, 1476–78 (A. Mitchell Polinsky & Steven Shavell eds., 2007); SUZANNE SCOTCHMER, *INNOVATION AND INCENTIVES* 38 (2004) (“Intellectual property protection gives innovators an incentive to invest in new knowledge.”); WILLIAM M. LANDES &

produce. But once they come into existence, they may be copied freely by others.²⁴ Intellectual property, by “securing for limited [t]imes to [a]uthors and [i]nventors the exclusive [r]ight to their respective [w]ritings and [d]iscoveries,” allows inventors or creators to charge supercompetitive prices during the period of exclusivity.²⁵ The ability to exclude others allows inventors and creators to recoup the costs of their initial investment.²⁶ In turn, this is thought to create an *ex ante* incentive to engage in the creative work in the first place.²⁷ In the traditional utilitarian view, then, intellectual property is a policy response to a specific public goods problem.²⁸

This incentive, however, entails significant social costs.²⁹ For one thing, the ability to price intellectual goods above marginal cost results in deadweight loss.³⁰ This static inefficiency is compounded by a dynamic inefficiency. Because intellectual goods are themselves inputs into further production, exclusion limits the ability of follow-on innovators to create new works.³¹ Intellectual property therefore involves a social welfare tradeoff: Society purchases the dynamic benefits of incentives to innovate at the cost of deadweight loss from monopoly pricing and the dynamic inefficiency that results from inhibiting downstream research. The standard incentive thesis suffers from another weakness: There is little empirical evidence that patents

RICHARD A. POSNER, *THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW* 294–300 (2003) (“The standard rationale of patent law is that it is an efficient method of enabling the benefits of research and development to be internalized, thus promoting innovation and technological progress.”).

24. More precisely, information-based goods are thought to be both nonrivalrous and nonexcludable, making them classic public goods. Nonrivalry means that one person’s use of a good does not preclude use by any other person. Nonexcludability means that no person can be excluded from using the good. SCOTCHMER, *supra* note 23, at 34.

25. U.S. CONST. art. I, § 8, cl. 8.

26. Menell & Scotchmer, *supra* note 23, at 1478.

27. SCOTCHMER, *supra* note 23, at 38.

28. See Menell & Scotchmer, *supra* note 23, at 1476–79 (justifying intellectual property as a solution to the market’s inability to incentivize innovation for nonrival public goods like knowledge and creative works).

29. See Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 TEXAS L. REV. 1031, 1058–59 (2005) (describing five types of social costs of intellectual property: “First, intellectual property rights distort markets away from the competitive norm, and therefore create static inefficiencies in the form of deadweight losses. Second, intellectual property rights interfere with the ability of other creators to work, and therefore create dynamic inefficiencies. Third, the prospect of intellectual property rights encourages rent-seeking behavior that is socially wasteful. Fourth, enforcement of intellectual property rights imposes administrative costs. Finally, overinvestment in research and development is itself distortionary.”).

30. *Id.* at 1059; see also Menell & Scotchmer, *supra* note 23, at 1477; SCOTCHMER, *supra* note 23, at 36–37.

31. Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEXAS L. REV. 989, 996–97 (1997); see also Suzanne Scotchmer, *Standing on the Shoulders of Giants: Cumulative Research and the Patent Law*, 5 J. ECON. PERSP. 29, 29–30 (1991) (stating “the cumulative nature of research poses problems” in intellectual property law as patents prevent innovators from building upon the works of others).

provide an incentive for the creation of works that would not have come into existence if the patent system did not exist in the first place.³²

These problems have led commentators and policy makers to search for alternative bases for the patent system. These efforts are both descriptive and normative in nature. Some seek to explain current features of the patent system; others seek to justify those features or to alter the patent system in ways that are justified by their social welfare effects.³³ Chief among these efforts is an attempt to look past the initial act of invention to ask what effects a system of intellectual property has on subsequent efforts to develop and commercialize that invention.³⁴

A. *The Commercialization Imperative*

Economists since Schumpeter have recognized that there is a difference between “invention” and “innovation.”³⁵ The act of invention or creation is the first step in bringing an intellectual product into the world. Invention is “the act of conceiving the design for a new and non-obvious technological product or process.”³⁶ Innovation, by contrast, is more than the conception of a new idea. It is “the search for and the discovery, development,

32. Barnett, *supra* note 4, at 793–94 & n.15. Most of the evidence against the incentive theory comes in the form of industry surveys that suggest that innovators do not rely on patents to protect their investments in research and development (R&D). See, e.g., Richard C. Levin et al., *Appropriating the Returns from Industrial Research and Development*, 1987 BROOKINGS PAPERS ON ECON. ACTIVITY 783, 796 (relying on survey data to conclude that patents have “limited effectiveness . . . as a means of appropriation”); WESLEY M. COHEN ET AL., PROTECTING THEIR INTELLECTUAL ASSETS: APPROPRIABILITY CONDITIONS AND WHY U.S. MANUFACTURING FIRMS PATENT (OR NOT) (Nat’l Bureau of Econ. Res., Working Paper No. 7552, 2000) (“Based on a survey questionnaire administered to 1478 R&D labs in the U.S. manufacturing sector in 1994, we find that firms typically protect the profits due to invention with a range of mechanisms Of these mechanisms, however, patents tend to be the least emphasized by firms in the majority of manufacturing industries, and secrecy and lead time tend to be emphasized most heavily.”); Cf. Michael Abramowicz & John F. Duffy, *The Inducement Standard of Patentability*, 120 YALE L.J. 1590, 1594 (2011) (“[I]f the innovation would be created and disclosed even without patent protection, denying a patent on the innovation costs society nothing (because the innovation would be developed anyway) and saves society from needlessly suffering the well-known negative consequences of patents . . .”).

33. See, e.g., Barnett, *supra* note 4, at 787 (offering “an alternative account of the patent system . . . that examines how patents influence innovation behavior by influencing organizational behavior”); Clarisa Long, *Patent Signals*, 69 U. CHI. L. REV. 625 (2002) (explaining the social value of patents as mechanisms to signal valuable information); Gideon Parchomovsky & R. Polk Wagner, *Patent Portfolios*, 154 U. PA. L. REV. 1, 1 (2005) (articulating theory of patent value based on aggregation of individual patents).

34. Lemley calls this distinction the difference between *ex ante* and *ex post* incentives, where *ex ante* refers to the incentives that exist before the initial act of creation or invention, and *ex post* refers to the incentives following that act. Lemley, *supra* note 9, at 130.

35. JOSEPH A. SCHUMPETER, CAPITALISM, SOCIALISM, AND DEMOCRACY 84 (2d ed. 1947); see also RICHARD R. NELSON & SIDNEY G. WINTER, AN EVOLUTIONARY THEORY OF ECONOMIC CHANGE 263 (1982); Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 89 VA. L. REV. 1575, 1660–61 & n.321 (2003) (following Schumpeter’s distinction between invention and innovation).

36. Sichelman, *supra* note 8, at 366.

improvement, adoption and commercialization of new processes, products, and organizational structures and procedures.”³⁷ Invention is the genesis of a new idea. Innovation is the process of bringing that idea to practical life.

There are several ways to describe the multitude of actions that inventors and others must take to bring a new idea to commercial fruition. The process usually requires the inventor first to put the idea into practice—to write a draft, record a demo, design a device, build a prototype. The inventor or creator must then demonstrate its worth. She must then figure out how to produce and distribute the product and determine whether there is a market for it and how to gain access to that market. In one view, the steps comprising “innovation” include identifying a problem to be solved, developing a working prototype, market testing and marketing, distribution, and follow-on improvements.³⁸

More generally, innovation can be thought of as comprising three distinct sets of activities: conception, development, and marketing.³⁹ *Conception* is the discovery of an idea. Ideas rarely have stand-alone economic value. Instead, they gain value when they are *developed*. Development therefore requires resources—capital and skills—to take the bare idea and operationalize it; that is, to determine how the idea will become embodied in a product or a process that has economic value.⁴⁰ Finally, those with a product in hand must still bring that product to market. They must produce it for sale, distribute it, and market it.⁴¹

37. Thomas M. Jorde & David J. Teece, *Innovation, Cooperation, and Antitrust*, in ANTITRUST, INNOVATION, AND COMPETITIVENESS 47, 48 (Thomas M. Jorde & David J. Teece eds., 1992).

38. Sichelman, *supra* note 8, at 348–54.

39. Cooter & Edlin, *supra* note 4, at 14–15; Bar-Gill & Parchomovsky, *supra* note 50, at 398–99. The details of the activities that innovators must undertake to bring their ideas through development and marketing will vary, of course, with the particular industry in which they are working. For several snapshots of the process in different industries, see ASHISH ARORA ET AL., *MARKETS FOR TECHNOLOGY: THE ECONOMICS OF INNOVATION AND CORPORATE STRATEGY* 45–89 (2001).

40. It is often said that development is the point at which an idea becomes patentable. *See, e.g.*, Bar-Gil & Parchomovsky, *supra* note 33, at 398 (noting that traditional patent law “denie[s] independent property rights in ideas,” but “grant[s] full property protection to ideas embedded in inventions”). This view finds support in black-letter patent law that draws a distinction between “conception” and “reduction to practice,” where only the latter is patentable. *See, e.g.*, *Ariad Pharm., Inc. v. Eli Lilly & Co.*, 598 F.3d 1336, 1352 (Fed. Cir. 2010) (en banc) (noting that actual or constructive reduction to practice, but not mere conception, may be sufficient to satisfy the description requirement of 35 U.S.C. § 112). At the same time, however, a competing and equally longstanding principle of patent law is that the inventor need not create a particular embodiment in order to receive a patent. *See id.* Patent law therefore appears to blur the line between conception and development, at least as those terms are defined as a matter of economic theory above. In the analysis that follows, I take the position that the choice when to protect an innovation as a matter of law is endogenous; that is, intellectual property can attach earlier or later in the process that I describe above.

41. A note on terminology is appropriate here. I shy away from the term “commercialization” in the description of economic functions above because it means different things to different people. To some, commercialization is only the step that I call “marketing.” *E.g.*, Bar-Gill & Parchomovsky, *supra* note 33, at 398. To others, commercialization “writ large” includes “any

This process is costly.⁴² Each of these activities requires financial resources. In some industries, the cost of development and marketing far outstrips the cost of conception. Partly as a result of these costs (but partly for other reasons) a great many inventions go without commercialization.⁴³ In such cases, society loses the benefits of invention. Promoting commercialization is therefore an important goal of innovation policy.⁴⁴

Edmund Kitch famously advanced the argument that intellectual property could be used to provide incentives not only for the initial act of creation or invention of an intellectual work, but for the subsequent development of that work as well.⁴⁵ Kitch analogizes patents to mining claims.⁴⁶ In his view, if a patentee is given broad control over a particular area of technology, the patentee will have the incentive to manage the development of that technology to maximize its social value, just as a private landowner has the incentive to maximize the value of her land.⁴⁷ In this way, broad patents give the owner the ability efficiently to “coordinate the search for technological and market enhancement of the patent’s value.”⁴⁸ Kitch also advocates early patenting, which provides the patent holder with the ability to coordinate subsequent development, a point to which I will return in subpart I(B).⁴⁹ Although Kitch’s argument is primarily concerned with improvements to the original patented technology, it directly addresses the commercialization concern described above. If commercialization is just as expensive and subject to free riding as the initial act of invention, then a broad patent will serve to internalize those costs in the patent holder and allow her to coordinate the development and marketing of the patented invention.

Following Kitch’s work, several scholars have advocated more directly for taking the costs of commercialization into account in setting patent policy. Scott Kieff, for example, makes the argument that strong property

activity following the initial invention that leads to a commercially available product or service—including developing, testing, manufacturing, sales, and service of the initial invention, *as well as* the invention and subsequent development of improvements.” Sichelman, *supra* note 8, at 354. I use the term “commercialization” to refer to both the development and marketing functions described above.

42. See Sichelman, *supra* note 8, at 371–72 n.184 (remarking that the cost of development and marketing greatly outweighs pre-invention expenses in many industries).

43. See *id.* at 362–65 (surveying empirical data).

44. See OFFICE OF TECH. ASSESSMENT, *supra* note 11, at iii (stressing the importance of successful development and commercialization of technological innovations for the future well-being of the United States).

45. Kitch, *supra* note 5. As Kieff points out, concerns about commercialization were voiced during the period leading up to and including the drafting of the 1952 Patent Act. Kieff, *supra* note 10, at 739–44. Kitch’s analysis is, however, the pioneering law and economics analysis of the incentives that the patent system may offer to potential developers and marketers of inventions.

46. Kitch, *supra* note 5, at 271–75.

47. *Id.*

48. *Id.* at 276.

49. *Id.* at 271, 277–78; *infra* subpart I(B).

rights are needed “to facilitate investment in the complex, costly, and risky commercialization activities required to turn nascent inventions into new goods and services.”⁵⁰ Kieff grounds his theory upon the same free-rider problem that plagues the initial development of new technology.⁵¹ Kieff argues that this problem also can hinder the commercialization of that technology.⁵² The investment in commercialization may be just as freely appropriable as the investment in the initial invention.⁵³ His solution is strong, property-rule-based intellectual property. Extending intellectual property rights and protecting them through a strong property rule will ensure that sufficient incentives continue through the commercialization process.⁵⁴ Michael Abramowicz similarly addresses the problem of patent “underdevelopment,” which he argues occurs “when a patentee decides to abandon a patent that the patentee would have commercialized if longer patent protection were available.”⁵⁵ Abramowicz focuses on the patent term length and observes that many patents expire before commercialization can take place. His solution, therefore, is to extend the patent term so that exclusivity continues through commercialization and second entrants have less ability to misappropriate the commercialization efforts of first entrants.⁵⁶

Of course, the logic of providing incentives for commercialization can extend beyond the patent system as it currently exists. Ted Sichelman critiques earlier commercialization theorists on the ground that early and broad patenting can bring about suboptimal levels of innovation and commercialization activity.⁵⁷ He instead approaches the commercialization problem more directly with a proposal for “commercialization patents” that would operate solely in the post-invention phase of innovation to produce a limited incentive to commercialize.⁵⁸ Along similar lines, Abramowicz and Duffy propose a new form of intellectual property protection for “market experimentation”—efforts to determine the size and extent of markets for new products.⁵⁹

Theories of intellectual property that place commercialization rather than invention at their core have been the subject of extensive critiques. Those critiques take two related forms. The first questions whether

50. Kieff, *supra* note 10, at 703.

51. *Id.* at 708–10.

52. *Id.* at 710.

53. *Id.* at 708–09.

54. *Id.* at 717–27.

55. Abramowicz, *supra* note 10, at 1073.

56. *Id.* at 1071–72. Abramowicz proposes that patent term extensions be doled out via an auction mechanism to limit patentees’ incentives to delay commercialization in the hope of gaining an extension of their period of exclusive rights. *Id.*

57. Sichelman, *supra* note 8, at 381–89.

58. *Id.* at 402.

59. Michael Abramowicz & John F. Duffy, *Intellectual Property for Market Experimentation*, 83 N.Y.U. L. REV. 337, 340 (2008).

incentives are really needed for commercialization. Mark Lemley takes this approach. He argues that “ex post” theories of intellectual property are “jarringly counterintuitive in a market economy” because we ordinarily suppose that efficiency in marketing and distribution arises from competition, not from exclusive rights.⁶⁰ The second questions whether the additional costs of broadening protection beyond what is necessary for the initial production of an intellectual good are worth the additional social benefits, if any, that accompany expanded intellectual property rights. Merges and Nelson, for example, argue that excessive patent scope leads to *less* development and commercialization and offer a series of case studies as evidence.⁶¹

Without engaging the broader debate about whether incentives are necessary for more than the initial act of creation or invention, I note that this strand of commercialization theory does not offer an independent justification for intellectual property. To be sure, these commercialization theorists have successfully focused attention on a more nuanced model of the innovation process than that which underlies the classical incentive or reward theory.⁶² But they have not identified an economically different function for intellectual property. The theory that commercialization efforts may be freely appropriable by others, and therefore need to be incentivized *ex ante* through a system of exclusive rights, is functionally indistinguishable from the theory that creative or inventive activity may be freely appropriable by others and therefore needs to be incentivized through a system of exclusive rights. In many ways, the “commercialization dilemma”⁶³ is a version of the same public goods problem that is thought to hamper inventive or creative activity in the first instance. It just occurs later in time. Or, to be more precise, it occurs later in the innovation process.

B. Commercialization and Information Exchange

There is another aspect to post-invention activity, however, that is different economically from the provision of *ex ante* incentives.⁶⁴ Development and commercialization not only are expensive, but they also require parties to communicate with one another. After conception, for

60. Lemley, *supra* note 9, at 135; *see also* Lemley, *supra* note 4, at 739–40 (“[W]e don’t normally need supracompetitive returns or the prospect of exclusivity just to encourage someone to take an existing invention to market.”).

61. *See generally* Robert P. Merges & Richard R. Nelson, *On the Complex Economics of Patent Scope*, 90 COLUM. L. REV. 839 (1990); *see also* Lemley, *supra* note 4, at 740–41 (explaining that inventors usually are not the best commercializers for a variety of reasons).

62. *See, e.g.*, Ted Sichelman, *Taking Commercialisation Seriously*, 33 EUR. INTEL. PROP. REV. 200, 200 (2011) (arguing for deeper and more consistent consideration of commercialization in economic and legal analyses of intellectual property).

63. Barnett, *supra* note 4, at 793–94.

64. *See* Tim Wu, *Intellectual Property, Innovation, and Decentralized Decisions*, 92 VA. L. REV. 123, 133 (2006) (noting that reducing transaction costs is a static rather than a dynamic benefit of intellectual property).

example, an inventor who seeks resources and skills for development must convince sources of financing or potential development partners that it is worth their effort to commit resources to the invention. To do this, of course, she must disclose sufficient information about the invention to enable her partners to make a decision regarding their resources. This process repeats itself, on perhaps a different scale and with different actors, once a fully developed invention needs to be marketed.

The disclosure paradox potentially inhibits this communication. An inventor seeking funds or development expertise may be reluctant to disclose information about her invention for fear that the recipients of the information can take it for themselves. On the other side of the transaction, the funders or developers will be unwilling to commit money or resources to the project unless or until they can assess its value. Arrow observed this dynamic and deemed it a “fundamental paradox”: the value of information “for the purchaser is not known until he has the information, but then he has in effect acquired it without cost.”⁶⁵ More recently, Cooter and his collaborators have described this phenomenon as a “double trust dilemma”: “To develop an innovation, the innovator must trust the investor not to steal his idea, and the investor must trust the innovator not to steal his capital.”⁶⁶ The double trust dilemma figures prominently in Cooter’s and Edlin’s account of the relationship between law and economic growth. They argue that overcoming the dilemma is critical to increasing the pace of innovation, which in turn is a key determinant of economic growth.⁶⁷

Some commercialization theorists recognize this problem and posit intellectual property as a solution. But their accounts of how intellectual property solves the problem are incomplete. The logic of the property rights solution is straightforward enough: The disclosure paradox arises because information is nonexcludable.⁶⁸ Once disclosed, it is generally difficult to prevent others from using the information. To the extent that intellectual property makes information excludable⁶⁹—by allowing the holder of a patent or a copyright to seek injunctive and monetary relief against those who would use the information—it provides a mechanism by which an inventor or creator can simultaneously disclose and protect her idea. Arrow himself recognized that “[w]ith suitable legal measures, information may become an appropriable commodity.”⁷⁰ In somewhat more detail, Merges explains that property rights create “the most effective form of precontractual liability,”⁷¹

65. Arrow, *supra* note 1, at 615.

66. ROBERT D. COOTER & HANS-BERND SCHÄFER, SOLOMON’S KNOT: HOW LAW CAN END THE POVERTY OF NATIONS 27 (2012).

67. Cooter & Edlin, *supra* note 4, at 13, 17.

68. See *infra* note 105 and accompanying text.

69. I cast doubt upon the ability of intellectual property to ensure perfect excludability of protected information in section II(B)(1), *infra*.

70. Arrow, *supra* note 1, at 615.

71. Merges, *supra* note 5, at 1488.

allowing parties to disclose information that is protected through other (noncontract) legal mechanisms. As Merges explains, property rights in information serve as a “protective cloak” during precontractual negotiations, enabling parties to disclose valuable information while still holding their negotiating partners liable for any attempts to appropriate that information before a contract is completed.⁷² If negotiations do fail, infringement actions are available to recover the value of the information disclosed.⁷³ Kitch similarly invokes the disclosure paradox and observes that a patent can “create[] a defined set of legal rights known to both parties at the outset of negotiations.”⁷⁴ That is, the disclosure of the invention in the patent instrument itself⁷⁵ solves the problem of negotiation in the face of asymmetric information: Both parties know the content of the intellectual good they are bargaining for. With this symmetrical knowledge, the parties can bargain over the “information protected by the scope of the legal monopoly.”⁷⁶

Kieff expands on Kitch’s argument by allowing for the possibility of coordination among multiple actors rather than by a single rights holder.⁷⁷ Kieff posits two mechanisms by which intellectual property can accomplish that coordination. First, intellectual property can serve as a “beacon,” “drawing together . . . many complementary users.”⁷⁸ Kieff explains that the threat of an injunction when intellectual property is protected by a strong property rule facilitates this effect. Threatened with possible injunctive relief, “diverse complementary users of the asset” have an incentive to find each other.⁷⁹ Once they do, Kieff posits that a “bargaining effect” facilitates

72. *Id.* at 1496.

73. Merges cites a second mechanism by which property rights facilitate transactions: They enable information holders to choose from a wider variety of enforcement options should the relationship go awry. This “enforcement flexibility” “enhance[s] the position of property holders when contractual disputes break out” by giving the rights holders a choice of different remedies and different forums. The availability of such a choice increases the confidence of potential information sellers. *Id.* at 1488.

74. Kitch, *supra* note 5, at 278.

75. See 35 U.S.C. § 112 (2006) (requiring disclosure of the patented invention).

76. Kitch, *supra* note 5, at 278.

77. See F. Scott Kieff, *Coordination, Property, and Intellectual Property: An Unconventional Approach to Anticompetitive Effects and Downstream Access*, 56 EMORY L.J. 327, 328 (2006) (arguing that property rule enforcement can lead to “coordination among entrepreneurs, inventors, and venture capitalists to facilitate commercialization of new ideas”).

78. *Id.* at 333–34; see also *infra* note 261 and accompanying text (noting that patents may potentially lower transaction costs by standardizing exchange).

79. *Id.* at 346. Of course, this reasoning requires at least two assumptions about the operation of the patent system. First, that the information disclosed in the patent document is sufficient to inform interested parties that they may want to engage with the patent holder. *But see infra* section II(B)(1). Second, that the information contained in the patent, even if adequate to convey the scope of the invention, is regularly communicated to the potential universe of competitors or collaborators. *But see* BESSEN & MEURER, *supra* note 18, at 54–68 (explaining why and how patents fail to provide adequate notice of the subject matter that they cover).

transactions among those attracted to the patent.⁸⁰ The latter effect refers to a solution to the disclosure paradox.⁸¹

A number of scholars drawing upon insights from the theory of the firm have explained how a grant of property rights in information could facilitate transactions over the protected information. Ronald Coase famously articulated the choice of production structure as being between markets and hierarchies.⁸² When transaction costs are low, production can be mediated through freely operating markets and contractual exchange.⁸³ When, on the other hand, transactions costs become prohibitively high, Coase predicted that firms would develop to bring the production process under the control of a central “hierarchy” free from the vagaries of market exchange.⁸⁴ Subsequent work has fleshed out the conditions under which production can be expected to take place through markets or within firms. Oliver Williamson and others have focused on the perils of contracting, noting in particular that it is impossible to write complete contingent contracts—contracts that specify the obligations of the parties in every state of the world.⁸⁵ In light of this difficulty, contracting parties often must determine how to minimize the threat that a party will behave opportunistically, attempting to benefit at the expense of the other.⁸⁶ Theorists of the firm have developed two approaches to this problem. Economists working in the tradition of transaction cost economics assert that parties can either attempt to erect contractual mechanisms to reduce the threat of opportunism, or they may bring the threat in house by vertically integrating.⁸⁷ Others working in the property rights theory tradition have identified a third option—the allocation of residual property rights over the subject of the contract.⁸⁸ As Merges describes, “transactors can work around contractual incompleteness by assigning a property right before entering into a contract.”⁸⁹

These insights can apply to transactions in information. The disclosure paradox acts as a kind of transaction cost, preventing parties from completing

80. *Id.* at 334, 346.

81. *Id.* at 414.

82. R.H. Coase, *The Nature of the Firm*, 4 *ECONOMICA* 386, 387–88 (1937).

83. *Id.* at 390–92.

84. *Id.* at 392–94.

85. OLLIVER E. WILLIAMSON, *THE ECONOMIC INSTITUTIONS OF CAPITALISM* 30–32 (1985); OLIVER HART, *FIRMS, CONTRACTS, AND FINANCIAL STRUCTURE* 23–24 (1995).

86. Oliver E. Williamson, *The Economics of Organization: The Transaction Cost Approach*, 87 *AM. J. SOC.* 548, 554 (1981).

87. *See, e.g.*, WILLIAMSON, *supra* note 85, at 90 (explaining that the degree of “asset specificity” “explain[s] vertical integration”); Benjamin Klein et al., *Vertical Integration, Appropriable Rents, and the Competitive Contracting Process*, 21 *J.L. & ECON.* 297, 298 (1978) (“Following Coase’s framework, this problem [the possibility of opportunistic behavior] can be solved in two possible ways: vertical integration or contracts.”).

88. *See* Merges, *supra* note 5, at 1484–85 (citing HART, *supra* note 85, among others who demonstrated that “transactors can work around contractual incompleteness by assigning a property right before entering into a contract”).

89. *Id.* at 1485.

market transactions.⁹⁰ Parties can minimize the threat that the buyer of information will act opportunistically upon the disclosure of the information he seeks to buy so long as the seller's information is protected through a property right. Several writers posit that intellectual property helps to minimize the transaction costs of interfirm transfers by solving the disclosure paradox.⁹¹ It is a short step from that observation to the argument that where such transfers would be economically efficient but for the presence of transaction costs, intellectual property rights in information that is the subject of exchange promote efficiency.⁹²

The theory of the firm suggests that in the absence of other solutions to transaction costs, firms will vertically integrate.⁹³ By this logic, the absence of property rights in information that firms need to transfer should lead those firms to integrate in order to accomplish the transaction. Arora and Merges demonstrate how strong intellectual property rights "make it possible for technology-intensive inputs to be supplied by separate firms," and therefore "contribute[] to the viability of these specialized firms as standalone entities."⁹⁴ Bar-Gill and Parchomovsky similarly argue that intellectual property plays a key role in defining the boundaries of the firm. In their model, nonprotectable innovation will take place within vertically integrated firms, while the advent of legal protection for intellectual property allows firms to achieve gains from trade during the innovative process.⁹⁵ Assuming that smaller firms tend to be more dynamic and innovative, the development of such firms may be efficiency-promoting.⁹⁶

This line of argument proposes an alternative economic rationale for intellectual property. It is aimed not at providing incentives for invention or commercialization but at reducing the costs of exchanging critical information. It also supports—sometimes explicitly and sometimes implicitly—the argument that intellectual property should be granted early in the innovation process and should be broad and strong so as to encourage the development of efficient industry structures.

90. Burk & McDonnell, *supra* note 14, at 587.

91. *Id.* at 587–90; Merges, *supra* note 5, at 1513–19; Heald, *supra* note 14, at 476; Bar-Gill & Parchomovsky, *supra* note 14, at 1653–54.

92. Burk & McDonnell, *supra* note 14, at 613–17; Bar-Gill & Parchomovsky, *supra* note 14, at 1654–55.

93. Coase, *supra* note 82, at 395–97.

94. Ashish Arora & Robert P. Merges, *Specialized Supply Firms, Property Rights and Firm Boundaries*, 13 *INDUS. & CORP. CHANGE* 451, 452 (2004).

95. Oren Bar-Gill & Gideon Parchomovsky, *Intellectual Property Law and the Boundaries of the Firm* 4 (Harvard Law Sch. John M. Olin Ctr. for Law, Econ. & Bus. Discussion Paper Series, Paper 480, 2004), available at http://lsr.nellco.org/harvard_olin/480. Bar-Gill and Parchomovsky assume that trade is not possible absent intellectual property rights. See *id.* at 1 (“[I]nformation that is not afforded legal protection cannot be bought or sold on the market.”). In subsequent work, Bar-Gill and Parchomovsky relax this assumption. Bar-Gill & Parchomovsky, *supra* note 14, at 1652. Barnett makes a similar argument that intellectual property rights are determinants of industry structure, which, in turn, determines the efficiency of innovation. Barnett, *supra* note 4, at 790–93.

96. Arora & Merges, *supra* note 94, at 451–52; Barnett, *supra* note 4, at 819–21.

C. *Questioning Commercialization Theory*

The studies described above identify an economic rationale for intellectual property distinct from both the traditional reward or incentive theory and the incentivize-to-commercialize theory I describe above. Rather than a dynamic benefit to be traded off against static social welfare losses, it is an independent static benefit of intellectual property. That is, by reducing transaction costs, intellectual property can induce the efficient exchange of information goods between purchasers and sellers. If the magnitude of this benefit is significant enough, then it represents a strong argument for the expansion of intellectual property. Indeed, most of the scholars described above advocate for stronger or broader intellectual property protection for the purpose of encouraging transactions in information. Taken to its logical conclusion, this argument suggests that intellectual property should expand backwards into the innovative process, where the problems of information exchange are particularly acute. If Cooter and Edlin are right that the interface between conception and development is the point in the innovation process that is most subject to inhibition by virtue of the disclosure paradox, then intellectual property should protect ideas.⁹⁷

But the writers described above seldom consider the full social welfare costs of their proposals.⁹⁸ To be certain, it is difficult to disentangle the various social welfare costs and benefits of intellectual property, especially when a given policy intervention is likely to affect more than one aspect of the calculus. Expanding intellectual property in early-stage inventions because it is thought to overcome the disclosure paradox will also result in changes to intellectual property's incentive effects and to the dynamic social welfare costs described above.

That said, if overcoming the disclosure paradox is to represent a stand-alone justification for intellectual property, it must at least satisfy two tests. First, the policy solution must be addressed toward a problem that is accurately described and of sufficient importance to warrant policy intervention. Second, the intellectual property solution must be the best among alternatives. If there are other, less socially costly, solutions that can be implemented, then, all else being equal, they should be preferred to intellectual property.

The existing literature mostly elides these two standards. Most commentators assume that the conventional account of the disclosure paradox is correct and that intellectual property solves the paradox.⁹⁹ In particular, they assume that the economic description of information that

97. Bar-Gill and Parchomovsky do, in fact, propose a limited entitlement of ideas for the purpose of encouraging a thicker marketplace for the exchange of such ideas. Bar-Gill & Parchomovsky, *supra* note 39, at 397. They do not advocate for outright patent or copyright protection for ideas and acknowledge that such proposals would be too socially costly. *Id.*

98. See *infra* part III(A).

99. For a representative sampling of such statements in the literature, see *supra* note 14.

underlies the conventional account is accurate,¹⁰⁰ and they largely fail to consider potential alternative solutions to the paradox other than intellectual property.¹⁰¹ In the next Part, I complicate each of those assumptions. Doing so reveals that further empirical work is needed before we can state that the conditions for adopting expanded intellectual property as a solution to the disclosure paradox are met.

II. A New Framework for Understanding and Overcoming the Disclosure Paradox

As Part I explained, there is an increasingly popular argument that seeks to justify strong and broad intellectual property rights because of their utility in overcoming the disclosure paradox. But that argument makes several assumptions about the nature of the paradox and its solutions that do not comport with the lived experience of information exchange. This Part therefore takes on those assumptions and demonstrates that they are neither theoretically nor empirically justified. Information is a far more complicated economic good than most commercialization theorists acknowledge. The extent to which the disclosure paradox actually disrupts information exchange depends on just how appropriable the information is. That characteristic—appropriability—is partly inherent in the information and partly manipulable by its holders. This more nuanced understanding of information supports a range of potential strategies for engaging in exchange, of which intellectual property is only one. Yet the existing literature largely discounts the efficacy and prevalence of these alternatives for exchanging information.

A. *The Economics of Information Goods*

The conventional understanding of the paradox relies on a highly stylized account of information. In particular, it assumes that information is nonexcludable and homogeneous.¹⁰² The former assumption is that once information is revealed, it is impossible to prevent others from using it.¹⁰³ The latter assumption is that information is a unitary good; it is revealed or concealed in its entirety.¹⁰⁴ Under these assumptions, the disclosure paradox is easy to explain. Take, for example, a valuable stock tip. Anyone who is exposed to the revealed information can act on it. And the original holder of the information, in choosing whether or not to disclose it, must generally

100. See *infra* subpart II(A).

101. See *infra* subpart II(B).

102. See, e.g., Arrow, *supra* note 1, at 614–15 (assuming that “[t]he cost of transmitting a given body of information is frequently very low” and that “a given piece of information is by definition an indivisible commodity”).

103. See *id.* (stating that any purchaser of information “can destroy the monopoly [of the information seller], since he can reproduce the [purchased] information at little or no cost”).

104. *Id.* at 615.

disclose the entire tip or none of it at all. Neither of these characteristics, however, accurately reflects the lived experience of information exchange. Instead, excludability is highly variable. It depends on the nature of the information and the parties' choices about how to communicate that information. And information usually is not a unitary good like a stock tip. It is a multilayered asset around which parties can self-consciously structure communications and relationships.

1. *Excludability*.—Economists and legal scholars often refer to information as either excludable or nonexcludable.¹⁰⁵ But excludability refers more precisely to the *costs* of exclusion.¹⁰⁶ Those costs are not binary. They occupy a spectrum. When the benefit of a good is the *information* conveyed in or about that good, the costs of exclusion actually can be highly variable. The costs of exclusion of information depend in part on the inherent characteristics of that information and in part on choices that information holders can make in shaping the environment in which their information interacts with the world.

Purely nonexcludable information can be imagined as free-floating facts and concepts that can be plucked out of the ether whenever someone encounters them. In this mental picture, the cost of exclusion is infinite.¹⁰⁷ Legal mechanisms are then thought to bring the cost of exclusion down by “fixing” the information in an identifiable *res* through the application of legal entitlements.¹⁰⁸ But information as it exists in the world—and, importantly, as it is exchanged between parties—is not so ethereal as the description above suggests. Instead, information is contained in “artifacts.”¹⁰⁹ Sometimes these artifacts are intangible—the information is contained in the

105. See, e.g., Menell & Scotchmer, *supra* note 23, at 1477 (“[I]n its natural state . . . knowledge is . . . ‘nonexcludable.’ That is, even if someone claims to own the knowledge, it is difficult to exclude others from using it.”); Lemley, *supra* note 29, at 1050–51 (“Information is what economists call a pure ‘public good,’ which means both that its consumption is nonrivalrous . . . and that it is not something from which others can easily be excluded.”).

106. See RICHARD CORNES & TODD SANDLER, *THE THEORY OF EXTERNALITIES, PUBLIC GOODS, AND CLUB GOODS* 6 (2d ed. 1996) (“Goods whose benefits can be withheld costlessly by the owner or provider display excludable benefits. Benefits that are available to all once the good is provided are termed nonexcludable.”).

107. Inversely, the cost of communication or transmission of the information is zero. See *infra* note 131 and accompanying text.

108. See Arrow, *supra* note 1, at 615 (“With suitable legal measures, information may become an appropriable commodity.”). Many property theorists also take this approach to conceiving information. See, e.g., Balganes, *supra* note 2, at 433 (“Two things become central then to the effective functioning of a licensing market: (1) the *ex ante* characterization of the entitlement as a property right, and (2) the law’s attaching it to an identifiable *res*, albeit a notional one.”); Henry E. Smith, *Intellectual Property as Property: Delineating Entitlements in Information*, 116 *YALE L.J.* 1742, 1755 (2007) (describing intellectual property rights as “a thing to be the object of exclusive rights as against the world”).

109. See CARLISS Y. BALDWIN & KIM B. CLARK, 1 *DESIGN RULES: THE POWER OF MODULARITY* 2 (2000) (explaining “artifacts” and design theory’s concern with their production). Design theory is largely concerned with the production of artifacts. *Id.*

minds of natural persons, in the operation of organizations, or in the structure of laws or institutions.¹¹⁰ Sometimes, however, they are quite tangible. Information may be contained in books, drawings, blueprints, computer code, datasets, and products. Different artifacts communicate information in different ways, and at different costs. Take, for example, information about how a simple machine might work. The information can be in the mind of the machine's inventor, where it can only be accessed through interaction with the inventor. He can set it down in a plan or a manual, where it can be accessed by reading. Or he can produce the machine, in which case the information about how it operates may or may not be revealed by inspecting the machine itself.

The excludability of information depends at least in part on the artifact in which it is contained. Patent law scholars have recognized that some inventions are "self-disclosing" or "self-revealing" while others are not.¹¹¹ Self-disclosing inventions, in Katherine Strandburg's formulation, allow "competitors . . . immediately [to] appropriate inventive ideas and begin commercial competition almost as soon as an inventor brings a patented product to market."¹¹² Many mechanical inventions have this characteristic—the paper clip, say, or a particular type of screw or fastener. The value-creating characteristics of the invention are apparent on its face once it is in use in the world. Others therefore can freely appropriate that value once they encounter the invention. Self-disclosing inventions are not limited to mechanical products. Pharmaceutical or chemical products may have this characteristic, as may some business methods.¹¹³ Other inventions are "impossible to discern by evaluating the product," such as the formula for Coca-Cola.¹¹⁴ Chemical processes that produce particular products may fall into this category as well.¹¹⁵ Of course, these categories are not binary. There are some inventions from which valuable information may be gleaned with effort—that is, they may be reverse engineered.¹¹⁶ Software code often has that characteristic.¹¹⁷ The object code sold to customers does not reveal the source code that would enable duplication, but that latter information

110. See *id.* (outlining intangible artifacts); see also *infra* note 149 and accompanying text (providing an example of an intangible artifact in computer design).

111. Alan Devlin, *The Misunderstood Function of Disclosure in Patent Law*, 23 HARV. J.L. & TECH. 401, 405 (2010); Mark A. Lemley, *The Surprising Virtues of Treating Trade Secrets as IP Rights*, 61 STAN. L. REV. 311, 338–41 (2008); Katherine J. Strandburg, *What Does the Public Get? Experimental Use and the Patent Bargain*, 2004 WIS. L. REV. 81, 104–18.

112. Strandburg, *supra* note 111, at 105.

113. *Id.*

114. Lemley, *supra* note 111, at 338.

115. *Id.* at 338–39.

116. See Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575, 1582–91 (2002) (describing legal and economic perspectives on reverse engineering).

117. Lemley, *supra* note 111, at 339.

sometimes can be gleaned through reverse engineering.¹¹⁸ In all events, the cost of exclusion depends in no small part on the manner in which information may be accessed from the artifacts that contain it.

The same reasoning applies to information contained in intangible artifacts. Economics and management scholars have long recognized that some knowledge is to be found not in transferable artifacts, but in persons.¹¹⁹ Most broadly, this “tacit knowledge” is information that has not been set down or codified.¹²⁰ More specifically, the term sometimes applies to information that is costly, difficult, or impossible to codify. In this narrower sense, tacit knowledge is perhaps more accurately described as “know-how.”¹²¹ To return to the example of the simple machine above, when the knowledge about how to work the machine resides solely in the mind of the inventor, it is “tacit” in the sense that it is uncoded. Should the inventor write an instruction manual, he would convert some of his tacit knowledge to articulated or codified knowledge. But there is perhaps some aspect of the machine’s working that is impossible to articulate; that is the accumulated “complex set of knowledge bases, competencies, and skills”¹²² that a person with expertise in a particular art comes to possess over time. Regardless of the precise definition of tacit knowledge, which can at times be elusive,¹²³ the important point is that tacit knowledge is at least partially excludable. Tacit knowledge, as Eric von Hippel notes, is “sticky”—it is “costly to acquire, transfer, and use.”¹²⁴ Sticky information can be transferred only if the costs of codification are incurred or if the person in possession of the information engages in social interaction with others who might want to acquire and use the information.¹²⁵

118. *Id.*

119. Michael Polanyi is widely credited with first articulating this concept of “tacit knowledge” in *THE TACIT DIMENSION* (1966). Nelson and Winter extend the concept to include knowledge contained not only in individuals, but also in organizations. NELSON & WINTER, *supra* note 35, at 76, 115–17.

120. Ashish Arora, *Contracting for Tacit Knowledge: The Provision of Technical Services in Technology Licensing Contracts*, 50 J. DEV. ECON. 233, 234 (1996) (“As the name suggests, tacit knowledge represents those components of technology that are not codified into blueprints, manuals, patents and the like.”); *see also* ARORA ET AL., *supra* note 39, at 95 (citing distinction between “tacit and codified dimensions of knowledge”).

121. Ashish Arora, *Licensing Tacit Knowledge: Intellectual Property Rights and the Market for Know-How*, 4 ECON. INNOVATION & NEW TECH. 41, 42–43 (1995); Chon, *supra* note 4, at 187.

122. ARORA ET AL., *supra* note 39, at 95.

123. *See* Robin Cowan et al., *The Explicit Economics of Knowledge Codification and Tacitness*, 9 INDUS. & CORP. CHANGE 211, 211–13 (2000) (describing a “considerable amount of semantic and taxonomic confusion” associated with “tacit knowledge”).

124. Eric von Hippel, *“Sticky Information” and the Locus of Problem Solving: Implications for Innovation*, 40 MGMT. SCI. 429, 429 (1994).

125. A separate branch of the literature addresses the social rather than economic dimension of tacit knowledge. *See, e.g.*, HARRY COLLINS, *TACIT AND EXPLICIT KNOWLEDGE* 11 (2010) (describing knowledge that requires individual social relationships or immersion in society to transfer); Chon, *supra* note 4, at 191–95 (describing both interpersonal and cultural aspects of knowledge transmission).

Generalizing from these observations—that information can be contained in artifacts, including individuals and organizations, that have different excludability characteristics—Sidney Winter articulates a taxonomy of information goods.¹²⁶ Winter writes that information goods can be classified along six dimensions: tacit and articulable; not teachable and teachable; not articulated and articulated; not observable in use and observable in use; complex and simple; and elements of a system and independent.¹²⁷ In this taxonomy, each attribute pair represents two poles. Information that lies closer to the pole represented by the first description above is harder or costlier to transfer; information that lies closer to the opposite pole is easier or less costly to transfer.¹²⁸ Each pairing represents a continuum.¹²⁹ Information may be easier or harder to transfer depending on where on each of the continuums the information lies.

It is worth pausing for a moment to return to the disclosure paradox. Recall that in Arrow's model, information is perfectly nonexcludable.¹³⁰ At the very least, the foregoing discussion demonstrates that this is not an accurate assumption to make. Information may be partially excludable, depending on the form that it takes as it exists in the world. This means that the costs of communicating that information are not always zero.¹³¹ Misappropriation of information therefore does not happen automatically upon exposure. Instead, nonzero communication costs mean that the disclosure paradox will not operate in all circumstances as the conventional account suggests. A potential development partner or venture capitalist who is shown a prototype of a device may not be able to determine from inspection how the device works. Some information may be transferred—information about what the device is or what it does; but other information will not necessarily be appropriated by the potential buyer—information about how to replicate the device and make it work. So long as the value of the latter is higher than the value of the former, disclosure by a seller of some information to a buyer does not imply that the buyer “has in effect acquired [the information] without cost.”¹³²

126. Sidney G. Winter, *Knowledge and Competence as Strategic Assets*, in *THE COMPETITIVE CHALLENGE: STRATEGIES FOR INDUSTRIAL INNOVATION AND RENEWAL* 159, 170–73 (David J. Teece ed., 1987).

127. *Id.*

128. *Id.*

129. *Id.*; see also Cristiano Antonelli, *The Business Governance of Localized Knowledge: An Information Economics Approach for the Economics of Knowledge*, 13 *INDUS. & INNOVATION* 227, 229–31, 237 tbl.1 (2006) (articulating an alternative framework).

130. Arrow, *supra* note 1, at 615.

131. See James Bessen, *From Knowledge to Ideas: The Two Faces of Innovation* 3 (Boston Univ. Sch. of Law Working Paper No. 10-35, 2012), available at <http://www.bu.edu/law/faculty/scholarship/workingpapers/2010.html> (arguing that communication costs fluctuate depending on economic factors).

132. Arrow, *supra* note 1, at 615.

Biotechnology companies often take advantage of the difficulty in transferring sticky knowledge in the early stages of negotiations for early-stage platform technologies. These are technologies that are primarily used as research tools.¹³³ When such technologies are in the early stages of development, they are typically not yet the subject of patent protection. But their development often requires partnerships or infusions of capital. Because they are research tools, some aspects of their effective use are tacit. The scientists who work with the tools know how to use and optimize them. As one biotech entrepreneur explained, he allows potential development or financial partners free access to his labs. These partners can see the technology in operation yet cannot use or replicate it themselves without the tacit knowledge of its developers. But the lab tours offer enough information about the invention to at least determine mutual interest. The parties then can negotiate for the transfer of the tacit knowledge.¹³⁴

In addition to assuming that the costs of communication are zero, Arrow's model also assumes that communication costs are exogenously fixed.¹³⁵ It is certainly true that some aspects of information goods are likely to be inherent in the goods.¹³⁶ Highly "tacit" information in Winter's taxonomy, for instance, is simply not capable of being "articulated" in symbols. (Though it may be transferable by teaching.)¹³⁷ Similarly, in the realm of tangible artifacts, information may be capable of embodiment in certain artifacts but not in others.

But the fact that some aspects of the informational content of a good may be unchangeable does not justify an assumption that *all* information characteristics of a good are immutable. Winter was among the first to point out that the structure of information is the result of economic choices that those in possession of the information can make.¹³⁸ It is often an endogenous choice. As Winter puts it, "The degree of articulation of anything that is articulable is partially controllable."¹³⁹ At times, information holders can

133. See Douglas Lichtman, *Property Rights in Emerging Platform Technologies*, 29 J. LEGAL STUD. 615, 615 & n.1 (2000) (describing and defining "platform technology").

134. Cf. Chon, *supra* note 4, at 196 ("The stickiness of such knowledge is something that can be used in a deliberate way to ensure that it is not diffused or that it is diffused only under controlled conditions such as the licensing of inventions.").

135. See Bessen, *supra* note 131, at 3 (noting Arrow's model assumed "exogenously low communication costs").

136. See Winter, *supra* note 126, at 174 ("[I]ntrinsic differences among knowledge bases and other circumstances of different areas of technology and organization are important determinants of where newly developed assets tend to fall along the taxonomic dimensions identified above."); Chon, *supra* note 4, at 189 (differentiating between tacit knowledge by choice and tacit knowledge due to communication costs).

137. See Winter, *supra* note 126, at 171–72.

138. See *id.* at 174 ("There do exist important opportunities for affecting the positions that particular knowledge development take on these dimensions."); ARORA ET AL., *supra* note 39, at 96 ("[T]he extent to which knowledge is codified, or more generally, the extent to which it is easy to transfer, is an economic decision rather than an inherent property of knowledge.").

139. Winter, *supra* note 126, at 174.

choose to articulate or codify their information or not. Similarly, information holders can choose to embody their information goods in self-disclosing artifacts or not. These choices of course impact the extent to which information can be transferred.

A small literature in both economics and law has attempted to understand the nature of the choice to make information more or less transferable.¹⁴⁰ It starts from the premise that converting less transferable knowledge to a more transferable form—for instance, codifying previously tacit knowledge—is costly.¹⁴¹ It requires developing a means to codify the information—to convert it from knowledge contained in individuals' minds to knowledge communicable through artifacts, and then actually doing so.¹⁴² The economic question, then, is under what circumstances might a firm undertake to incur the costs of making knowledge more transferable. Winter posits that a firm will do so when the benefits of voluntary transfers outweigh the potential costs of involuntary transfers; that is, when it is more beneficial to a firm to be able to engage in information exchange than to guard against misappropriation.¹⁴³ Bessen models the decision to formalize knowledge where the costs of doing so are nonzero in a variety of circumstances, and finds that “it does not pay to formalize knowledge unless the market is sufficiently large to recoup formalization costs.”¹⁴⁴

This literature takes the stickiness of knowledge as an impediment to transfer that must be overcome in order for contracting over knowledge to occur. There are two complicating factors, however, that shed further light on the nature of the disclosure paradox: First, information holders do not face a binary choice to codify their information or not. Instead, the range of options available to information holders is much wider. The decision whether or not to *codify* information is really a decision about how to *structure* information. Consider, for example, the concept of “modularity” that is often invoked in software design (and in design theory more broadly). Modularity is a design principle that seeks to decompose a complex system into parts—or “modules”—that are highly independent yet can work together.¹⁴⁵ An architect designing a complex system achieves modularity in part by drawing a sharp distinction between visible information and hidden

140. See, e.g., ARORA ET AL., *supra* note 39; Bessen, *supra* note 131; Dan L. Burk, *The Role of Patent Law in Knowledge Codification*, 23 BERKELEY TECH. L.J. 1009, 1012–17 (2008); Cowan et al., *supra* note 123; Winter, *supra* note 126.

141. See Bessen, *supra* note 131, at 9–14 (detailing the costs of communicating technical knowledge); Burk, *supra* note 140, at 1013–16 (discussing the costs associated with the codification of knowledge).

142. See Burk, *supra* note 140, at 1013–14 (discussing the costs involved in creating and implementing a scheme to codify information); Cowan et al., *supra* note 123, at 247–48 (explaining the ways in which codification of knowledge can sometimes decrease its communicability).

143. Winter, *supra* note 126, at 173–80.

144. Bessen, *supra* note 131, at 3.

145. See BALDWIN & CLARK, *supra* note 109, at 63 (defining modularity); Smith, *supra* note 108, at 1761–63 (explaining the utility of modularity in dealing with complex systems).

information.¹⁴⁶ Only the visible information is required for the modules to cooperate.¹⁴⁷ Information specific to the workings of the module itself can remain hidden from the other modules.¹⁴⁸ The designer of a computer operating system, for example, can keep most of the details of the system's internal processes secret, while revealing to the world the set of commands that allow programs to interface with it.¹⁴⁹ Similarly, information holders can design the artifacts that embody their information to make some aspects excludable and other aspects freely available.

Second, the conventional account of the disclosure paradox suggests that there is a one-to-one correspondence between the decision to codify and the decision to transfer. Winter writes that “[f]eatures that restrain involuntary transfer tend to inhibit voluntary transfer; likewise, actions undertaken to facilitate voluntary transfer may well facilitate involuntary transfer also.”¹⁵⁰ Likewise, the literature modeling the economic choice to codify information assumes that the choice to codify is made when the possessor of the information wants to transfer it.¹⁵¹ But once the choice of information structure is understood not to be binary, the relationship between information structure and transfer becomes more complicated.

The impediment to transfer that the disclosure paradox describes is not cost. It is appropriability. The paradox suggests that parties will be unable to transfer information when it is in a form that renders it freely appropriable by others. What is needed, therefore, is some kind of *optimum* level of appropriability that allows for (a) sufficient information to be transferred to link ideas with capital and development partners while (b) ensuring that enough value remains in the original information holder so that she still has an incentive to disclose.

This theoretical optimum can be achieved through the use of nonbinary information management techniques described above. Most simply, parties can engage in selective disclosure. If parties are able to partition their information so as to reveal some but not all of the relevant information to counterparties, then it is possible to facilitate exchange while simultaneously guarding against misappropriation. But the discussion above suggests that parties can manipulate not only the plain amount of information that they reveal to others, but also the form that their information takes. Biotech companies thus choose to leave certain information tacit not to keep it to themselves, but actually to facilitate transfer by overcoming the disclosure

146. See BALDWIN & CLARK, *supra* note 109, at 72–76.

147. *Id.* at 73.

148. *Id.*

149. See, e.g., *United States v. Microsoft Corp.*, 253 F.3d 34, 53 (D.C. Cir. 2001) (explaining operation of “Application Program Interfaces” or “APIs” that expose some but not all software operating routines to potential developers).

150. Winter, *supra* note 126, at 174.

151. See, e.g., Arora, *supra* note 121; Bessen, *supra* note 131.

paradox.¹⁵² Software developers use modularity to shield some information from potential partners so that they can overcome the disclosure paradox and engage in constructive transfers of commercially valuable information. In each of these examples, the information holder relies upon the partial excludability of information and the ability to manipulate the information content of the artifacts at her disposal to achieve *some* level of disclosure and some level of forbearance. It is not always the case that decisions to make information less transferable will induce less transfer. Instead, utilizing relatively less transferable forms of artifacts that nevertheless convey sufficient information to enable exchange actually can induce *more* transfer by overcoming the disclosure paradox.

As the discussion above indicates, the excludability characteristics of information are far from binary. This means that the disclosure paradox does not always prevent the successful sharing of an information good. The good may itself be partially excludable, allowing the potential buyer to access enough information to estimate its worth while allowing the seller to retain sufficient value; or the information holder can design the information-conveyance mechanism in such a way as to enable disclosure while guarding against misappropriation.

2. *Heterogeneity*.—The conventional account of the disclosure paradox conceives of information as a homogeneous asset.¹⁵³ In this view, information is discrete. It is singular. It is the stock tip described above, which the holder either knows or not, can act upon or not, and can disclose or not.¹⁵⁴ But very little information has the characteristics of a stock tip. More often, information is multilayered and continuous. More particularly, different types of information about a particular intellectual product may be relevant in different circumstances and contexts of exchange. Information is heterogeneous.

This phenomenon is perhaps best illustrated by the example of small-molecule pharmaceutical development described above.¹⁵⁵ Most drugs are single compounds.¹⁵⁶ A single compound corresponds to a single product. The structure of the compound is the critical information behind the product—it defines the product's pharmacological properties. The structure also is highly self-revealing.¹⁵⁷ Once a drawing or chemical formula that

152. See *supra* notes 133–34 and accompanying text.

153. See, e.g., Arrow, *supra* note 1, at 615 (noting that information is “indivisible”).

154. See *supra* subpart II(A).

155. See *supra* text accompanying notes 15–20.

156. Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 89 VA. L. REV. 1575, 1590, 1617 (2003).

157. It is true that most pharmaceutical compounds are protected by patents. But patents provide only incomplete protection from competitive misappropriation. This is particularly true during negotiations between small biotechs and large pharmaceutical companies. Because these negotiations take place in the preclinical or early-stage clinical phases of pharmaceutical testing, it

reveals the structure of the compound is shown to potential partners, those partners know all they need to know to reproduce the pharmaceutical.¹⁵⁸ The disclosure paradox should operate according to Arrow's model in this circumstance to block even the initial contact between the biotech that is developing the compound and the pharmaceutical company with which it seeks a partnership for development and commercialization. But while the structure is of course the driver of value in the market for approved pharmaceuticals, its disclosure may not be necessary to assess its value as an input into development and commercialization processes. Instead, at the licensing stage, the most commercially relevant information might be data *about* the compound: its efficacy, its pharmacological characteristics, and so forth. Commercially useful information short of the core intellectual asset may thus be disclosed in the course of a negotiation. Indeed, in the pharmaceutical example, the negotiation may be all but completed by the time the structure is revealed.¹⁵⁹

A similar phenomenon can be observed between software innovators and potential sources of funding.¹⁶⁰ The core intellectual asset that a software developer has is her code.¹⁶¹ But she need not disclose the code to convey commercially relevant information to potential funders. Instead, the early meetings between entrepreneurs and investors focus on what the software can do, what the potential underserved need might be, what the competitive landscape for the application might be, and similar questions. That information enables potential funders and partners to evaluate the business opportunity without appropriating the core information asset. Only later in the negotiation will the code be revealed.

As a practical matter, then, both biotechnology and software entrepreneurs will begin discussions with potential investors and partners by revealing information *about* their product or idea, but not the structure of the product or the details of the idea itself.¹⁶² They are able to do this because

is possible for a large pharmaceutical company that has access to the structure of a promising compound to innovate around the patent protecting that compound. *See infra* notes 171–75 and accompanying text.

158. That knowledge does not, however, guarantee that a potential competitor could complete testing, Food and Drug Administration (FDA) approval, and marketing of the compound first. The seller here therefore retains some first-mover advantage, which may itself be a means to guard against misappropriation. *See infra* notes 240–41 and accompanying text.

159. *See supra* text accompanying notes 15–20.

160. This example is drawn from interviews with several Boston-area entrepreneurs and venture capital investors.

161. Like the pharmaceutical molecules described above, software code may be subject to formal intellectual property protection, but that protection is inevitably incomplete. Most source code is copyrighted, but it is often a relatively straightforward task to produce similar functionality using code that is not directly copied from the copyright holder.

162. A similar illustration of the multifaceted nature of information can be seen in the literature on “patent-paper pairs,” which seeks to explain why scientists reveal information about a research project simultaneously in academic publications and patent applications. The explanation turns on the fact that scientific research produces both academically useful and commercially useful

information is multilayered. To generalize from the examples above, imagine a series of concentric circles. In the innermost circle lies the “core” information asset. The definition of the core asset depends on the particular technological and business context. One can reasonably posit, however, that it is at least the asset that the holder would be most fearful of releasing to the public. Most likely, this is because it represents the bulk of the value to the holder. To the pharmaceutical company, the structure of the molecule it is developing into a drug is the core information asset. To a software developer, it may be the code for the software.¹⁶³

Beyond this core lies “second-order” information that can be used to describe some relevant characteristics of the asset.¹⁶⁴ This information is directly related to the characteristics of the core asset. In the case of the pharmaceutical molecule, it may refer to the molecule’s physical characteristics other than its structure: its pharmacological properties, the diseases that it targets, and so forth. In the case of software code, this direct information may include what the code does or a description of its operation at a somewhat higher level of abstraction. Beyond this second-order information lies other higher-order information. The further one gets from the core, the more attenuated this information becomes. In the pharmaceutical example, this higher-order information may be the data about the drug’s performance in preclinical testing; in software, it may be information about the market opportunity. But even highly attenuated information still conveys knowledge about the core asset.

In this way, entrepreneurs can design their information flows to enable meaningful commercial exchange without revealing the core information

information, and that the two types of information can often be separated from one another. Joshua S. Gans et al., *Contracting Over the Disclosure of Scientific Knowledge: Intellectual Property and Academic Publication 4* (Apr. 8, 2011) (unpublished manuscript), available at <http://ssrn.com/abstract=1559871>.

163. It is important to note here the contingency of the word “may.” It is also possible that there are other sources of value for a software developer that ultimately are more important than the code. See *infra* note 243 and accompanying text.

164. This taxonomy bears some resemblance to that in R. Polk Wagner, *Information Wants to Be Free: Intellectual Property and the Mythologies of Control*, 103 COLUM. L. REV. 995, 1003–10 (2003). It is, however, different in both concept and purpose. Wagner articulates three types of information that vary primarily in their appropriability: Type I, which is protected by intellectual property; Type II, which comprises directly related works or improvements; and Type III, which represents spillovers or generative information related to the intellectual property. Wagner is concerned, however, with appropriability as a matter of positive law, while my concern is with the communicability of commercially useful information. Closer perhaps is the concept of information spillovers described in Brett M. Frischmann & Mark A. Lemley, *Spillovers*, 107 COLUM. L. REV. 257 (2007). Finally, this concept is similar to the problem—common to both copyright and patent law—of identifying the correct “level of abstraction” to define the scope of information to be protected by an exclusive right. See BRETT M. FRISCHMANN, *INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES* 288–92 (2012) (discussing the difficulty in drawing a line between idea and expression); Tun-Jen Chiang, *The Levels of Abstraction Problem in Patent Law*, 105 NW. U. L. REV. 1097, 1100–01 (2011) (describing how patent scope varies with the level of abstraction of description).

asset. This is true even when that asset is highly self-revealing. The information holder who is unable to rely on inherent or designed excludability may nonetheless still engage in exchange of information *about* her information.

B. Alternative Solutions to the Disclosure Paradox

The discussion above suggests that parties seeking to exchange information may in some cases rely upon characteristics of the information itself to accomplish transactions or, perhaps more frequently, can design their information flows in such a way as to enable commercially meaningful communication while simultaneously guarding against misappropriation. The nature of information itself therefore gives rise to strategies for overcoming the disclosure paradox that are based on manipulating information flows. The characteristics of information described above complicate the intellectual property solution and also enable a series of strategies that are routinely overlooked or dismissed in the existing literature.

1. Intellectual Property.—As Merges and others have observed, intellectual property may in certain circumstances play a role in overcoming the disclosure paradox.¹⁶⁵ But positive law intellectual property regimes have limitations. While intellectual property may facilitate disclosure in some circumstances, it may be inadequate in others. Understanding the complex nature of information helps to determine circumstances in which intellectual property may or may not help to overcome the disclosure paradox.

The basic logic of the disclosure paradox suggests that legal intervention is necessary for otherwise freely appropriable information to become less appropriable and therefore subject to exchange.¹⁶⁶ Arrow understood, however, that these legal measures were necessarily limited: “[N]o amount of legal protection,” he wrote just a paragraph before explaining the disclosure paradox, “can make a thoroughly appropriable commodity of something so intangible as information. . . . Legally imposed property rights can provide only a partial barrier, since there are obviously enormous difficulties in defining in any sharp way an item of information and differentiating it from other similar sounding items.”¹⁶⁷ Arrow’s

165. See *supra* notes 68–76 and accompanying text.

166. See Arrow, *supra* note 1, at 615 (“With suitable legal measures, information may become an appropriable commodity.”).

167. *Id.* Sivaramjani Thambisetty similarly argues that patents do not provide an adequate solution to the disclosure paradox because they are not in fact “the sharp exclusive right that is central to [Arrow’s] thesis.” Sivaramjani Thambisetty, *Patents as Credence Goods*, 27 OXFORD J. LEGAL STUD. 707, 707 (2007). Thambisetty does not, however, question the need for intellectual property to resolve the disclosure paradox; his argument is confined to criticizing the current implementation of patent law on the ground that it fails to resolve the paradox. *Id.* at 707–09.

observation is consistent with a more nuanced conception of the information that is produced by and is necessary for innovation.

When the scope of intellectual property rights corresponds with the scope of information sought to be disclosed, then intellectual property may indeed solve the disclosure paradox. This is most likely to occur with respect to inventions that are relatively easy to “claim” through modern intellectual property regimes.¹⁶⁸ When claiming is effective, there is a one-to-one correspondence between the scope of protection of the patent and the invention. In this case, the invention can be disclosed and will be entirely protected from misappropriation by the scope of the patent.

But there are a variety of circumstances in which this one-to-one correspondence will break down.¹⁶⁹ For one thing, intellectual property may underprotect the information good that needs to be exchanged. For goods that are highly self-disclosing, revelation of the core information asset in the patent may facilitate design-around. That is, a potential buyer once exposed to the information can attempt to implement the invention covered by the patent with changes that remove the new effort from the patent’s coverage. Designing around is a familiar phenomenon in patent law, and is often thought to represent a social welfare benefit.¹⁷⁰ But a rational information holder faced with the possibility that disclosing her information may lead to easy design-around will still be reluctant to disclose even if the information is protected by a patent.

The extent to which design-around poses a continuing danger to information holders who have intellectual property protection depends on several factors, including the timing of the negotiation over the information and the ability to draft broader patent claims.¹⁷¹ In pharmaceuticals, for instance, negotiations over the rights to develop a compound often occur relatively early in the product-development cycle. At this stage, patent doctrine may prevent overly broad claims.¹⁷² At the same time, because

168. In the patent system, for example, the “claim” represents the “metes and bounds” of the invention. A rich literature details some of the difficulties associated with modern claiming, not the least of which is that it is highly uncertain. *See, e.g.,* BESSEN & MEURER, *supra* note 18, at 56–62 (discussing various processes for interpreting vague claims). More specifically for present purposes, claiming often proves to be both under- and over-inclusive. Dan L. Burk & Mark A. Lemley, *Fence Posts or Sign Posts? Rethinking Patent Claim Construction*, 157 U. PA. L. REV. 1743, 1750, 1765 (2009).

169. *Cf. Lemley, supra* note 4, at 740 (noting that patents usually do not correspond one-to-one with relevant product markets).

170. *See, e.g., Lemley, supra* note 4, at 753 n.248 (listing courts and commentators that have recognized the value in design-around).

171. Conventional wisdom is that patent drafters attempt to draft claims as broadly as possible, but their ability to do so depends on the technology and the relevant doctrine in the area. Burk & Lemley, *supra* note 168, at 1762–63.

172. More specifically, the enablement doctrine limits the extent to which pharmaceutical companies may patent small molecules whose efficacy remains uncertain. *See, e.g., In re '318 Patent Infringement Litig.*, 583 F.3d 1317, 1327 (Fed. Cir. 2009) (holding that early-stage research failed to support a patent application for small-molecule drug treatment).

many small molecules may have similar biological effects, it is possible for a competitor, upon learning the focus of a company's research, to pursue its own research on a similar molecule that falls outside the scope of the patent. This goes a long way toward explaining why, in the biotech–pharmaceutical example with which this Article began, the patent that protects the molecule does not appear to play a role in the process of exchanging information. This is so despite the conventional wisdom that pharmaceuticals are the paradigmatic industry in which patents promote innovation.¹⁷³ Although patents may offer protection in the product market for pharmaceutical products, they appear to play a very different role in the market for development and commercialization rights. In software, where the evidence that patents play a significant role in the product market is much more attenuated,¹⁷⁴ it is not surprising that design-around is particularly easy as well.¹⁷⁵

The alternative scenario in which patents fail to solve the disclosure paradox completely is when they underdisclose. The disclosure provided by a patent is limited. The Patent Act requires a patentee to provide, in addition to the “claims” described above,¹⁷⁶ “a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same.”¹⁷⁷ Patentees can often draft their patent disclosures, however, in such a way as to keep significant—and significantly useful—information to themselves.¹⁷⁸ A skilled patent lawyer will draft the disclosure of a patent to meet the bare minimum requirements of the law without disclosing any information that can usefully be held back as a trade secret.¹⁷⁹ Even to the extent that patents do disclose useful information, there are a variety of

173. See, e.g., BESSEN & MEURER, *supra* note 18, at 138–46 & fig.6.5 (concluding that positive returns to patent prosecution and litigation exist only in the chemical and pharmaceutical industries); DAN L. BURK & MARK A. LEMLEY, *THE PATENT CRISIS AND HOW THE COURTS CAN SOLVE IT* 80–81 (2009) (noting the importance of patents in the pharmaceutical industry).

174. See, e.g., BURK & LEMLEY, *supra* note 173, at 40, 43, 47 (explaining features of IT industries that render patent protection less relevant for innovation).

175. See Colleen V. Chien, *Predicting Patent Litigation*, 90 TEXAS L. REV. 283, 291 (2011) (“Short life cycles and the ability to design around patents in the IT sector contribute to what Henry Chesbrough characterizes as a ‘weak appropriability’ regime in which it is more difficult for innovators to exclusively benefit from their innovations.” (quoting HENRY CHESBROUGH, *EMERGING SECONDARY MARKETS FOR INTELLECTUAL PROPERTY: US AND JAPAN COMPARISONS* 31 (2006))).

176. See 35 U.S.C. § 112 ¶ 2 (2006) (“The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.”).

177. *Id.* § 112 ¶ 1.

178. See Devlin, *supra* note 111, at 403 (noting that patents often fail to convey meaningful information); Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539, 563 (2009) (suggesting that the patent system encourages writers to underdisclose); Note, *The Disclosure Function of the Patent System (or Lack Thereof)*, 118 HARV. L. REV. 2007, 2025–26 (2005) (same).

179. *The Disclosure Function of the Patent System (or Lack Thereof)*, *supra* note 178, at 2026.

reasons to believe that they are insufficient as communication devices for information exchange. Patent documents are usually written by and for lawyers rather than by and for scientists or business people; as such, they often fail to communicate the relevant technical data in the most usable fashion.¹⁸⁰

Putting these observations together yields the conclusion that the exchange of commercially useful information often requires parties to go beyond patents. As Arora observes, “most of the theoretical literature on licensing assumes that all technical knowledge is contained in patents or formulae,” but “efficient technology transfer usually also requires the transfer of know-how.”¹⁸¹ Even to the extent, then, that patents facilitate the transfer of *some* useful knowledge, that transfer often must be accompanied by the simultaneous transfer of additional knowledge that is not the subject of intellectual property protection. It is not enough to share the details of a machine. You also need to share the inventor’s insight into how it *works*. That brings back the same problems in transferring tacit knowledge described in subpart II(A). That knowledge is costly to transfer and its transfer is subject to opportunistic behavior.¹⁸²

Economists have identified a role for patents in this transfer, but it is not the role that is assumed by the conventional account of the disclosure paradox. A patent can be thought of as one component of a package of knowledge that also includes know-how. Successful technology transfer requires transferring *all* components of the package.¹⁸³ But because the patent creates legal excludability, a license to use the subject matter of the patent can be withdrawn. One contracting strategy, therefore, is to use the complementarity between the excludable asset (the patent) and the nonexcludable asset (the know-how) to induce efficient contracting. The patent is effectively used as a “hostage” that can be withdrawn if payment is not made for the know-how; likewise, the buyer of the know-how can

180. See Sean B. Seymore, *The Teaching Function of Patents*, 85 NOTRE DAME L. REV. 621, 625–27 (2010) (describing why patent documents are not read more widely by scientists and business people).

181. Arora, *supra* note 121, at 41.

182. See *supra* notes 119–25 and accompanying text. Recall that the specific double opportunism associated with transferring tacit knowledge is that “[o]nce the know-how is transferred, the buyer may try to avoid paying for it, since it would be difficult to force her to unlearn what she has been taught. On the other side, given the cost of transferring know-how, the licensor may be tempted to skimp on the know-how provided.” ARORA ET AL., *supra* note 39, at 118.

183. See Peter Lee, *Transcending the Tacit Dimension: Patents, Relationships, and the Industrial Organization of Technology Transfer*, 100 CALIF. L. REV. (forthcoming 2012) (manuscript at 21) (explaining that “even where patent disclosure satisfies statutory and doctrinal requirements, it is often lacking” and that technology transfer therefore must include “useful knowledge about patented inventions [that] remains tacit”), available at <http://ssrn.com/abstract=2019335>.

postpone at least part of the payment until the information has been transferred.¹⁸⁴

Patents therefore can play a variable role in the exchange of valuable information. Sometimes they may facilitate transfer of the entire sum of useful knowledge. At other times, they may fall short. And sometimes they may be used in conjunction with other strategies. The ultimate conclusion, however, is that the multifaceted nature of information makes the use of a patent to overcome the disclosure paradox contingent.

2. *Contracts*.—The difficulties of contracting for the sale of information lie at the heart of the conventional account of the disclosure paradox. In a world in which information is a simple asset, opportunism will effectively prevent a contract for sale and will also prevent the parties from striking a separate contract for secrecy. But understanding that information is a complex, multifaceted asset reveals a range of contracting strategies by which parties may effectively accomplish exchange. Key to these strategies is that—consistent with the complexity of the information that parties seek to exchange—they take on features of privately agreed-to *governance* mechanisms rather than simple contracts.

The disclosure paradox is, at its heart, a problem of contract. A contract for the sale of information cannot be completed because of the threat of dual opportunism.¹⁸⁵ The parties generally cannot strike a one-time bargain for the sale of information because the seller fears the buyer can take the information without paying if she divulges first, and the buyer cannot value the information without disclosure.¹⁸⁶ Other tools of contract theory, including “earnouts” and other mechanisms contingent upon a determination of the value of the information following disclosure, also are generally ineffective.¹⁸⁷

184. ARORA ET AL., *supra* note 39, at 116–17 (“[E]fficient contracts for the exchange of technology can be written by exploiting the complementarity between know-how and any other technology input that the licensor can use as a ‘hostage.’”); *see also* Arora, *supra* note 120, at 234–35 (proposing stronger intellectual property to facilitate contracting in this manner).

185. Arrow, *supra* note 1, at 614–16. These difficulties also are predicted by the transaction cost economics literature. *See, e.g.*, WILLIAMSON, *supra* note 85; *see also* Merges, *supra* note 5, at 1495–504 (explaining the property rights solution to the paradox as a means to establish precontractual liability).

186. *See* Cooter & Edlin, *supra* note 4, at 16 (describing this so-called “double trust dilemma”); Barnett, *supra* note 4, at 797–98 (noting that unwillingness to enter transactions reflects an “underlying drafting constraint”). Barnett generalizes from these difficulties to conclude that “contractual solutions cannot reliably overcome the disclosure paradox.” *Id.* at 797. My analysis goes beyond Barnett’s by relaxing his assumptions about the nature of the good to be traded. *Cf. id.* at 797–98 (“Suppose the typical scenario in which an inventor has formulated an idea and wishes to sell it to a large integrated firm.”).

187. *See* Barnett, *supra* note 4, at 798–99 (outlining issues arising from “earnout” provisions). Several economists have modeled scenarios in which certain contractual mechanisms may facilitate the exchange of appropriable information. *See generally, e.g.*, James J. Anton & Dennis A. Yao, *The Sale of Ideas: Strategic Disclosure, Property Rights, and Contracting*, 69 REV. ECON. STUD. 513 (2002) (arguing that partial disclosure plus bond might overcome transactional problems in

The parties usually cannot overcome this difficulty through the use of a nondisclosure agreement, for a number of reasons. First, nondisclosure agreements themselves may fall victim to the disclosure paradox. Without knowing the information that the agreement might seek to protect, a buyer will generally be unwilling to subject herself to potential liability for violating the terms of the agreement. The problem is that the buyer may *already* know the information. In that case, a buyer who signs a nondisclosure agreement and only then learns of the subject matter the agreement covers is exposed to liability.¹⁸⁸ This explains the conventional wisdom that most venture capitalists or Hollywood studios routinely refuse to sign nondisclosure agreements.¹⁸⁹ These sources of capital hear hundreds if not thousands of pitches in a year. If they signed nondisclosure agreements prior to hearing every new idea, they would likely be exposed to massive liability when the ideas inevitably overlapped in some fashion, large or small.¹⁹⁰

But the utility of contracts changes when the subject of exchange is viewed not as a singular stock tip but as a more complicated asset. Most importantly, the exchange of information often requires more than a single interaction. Multiple exchanges are sometimes necessary as a result of the inherent characteristics of the information. Tacit information that cannot be readily codified, for example, can only be transferred through multiple interactions among the parties to the exchange.¹⁹¹ Alternatively, parties can structure the flow of information around their core assets to enable staged disclosure.¹⁹² In all events, the need for multiple interactions expands

technology contracts); Joshua S. Gans & Scott Stern, *The Product Market and the Market for "Ideas": Commercialization Strategies for Technology Entrepreneurs*, 32 RES. POL'Y 333 (2003) (identifying a range of commercialization strategies based on excludability and asset complementarity). I put these models aside for several reasons. First, there is no evidence that they are used in practice. Second, to the extent that they rely on the use of bonding mechanisms, they presuppose some independent wealth in the idea holder. See, e.g., Anton & Yao, *supra*.

188. See Barnett, *supra* note 4, at 798 ("No idea buyer will covenant against use since the idea buyer may already possess the idea, in which case it would be exposed to expropriation by the idea seller."); Bar-Gill & Parchomovsky, *supra* note 39, at 405 (noting that buyers are unlikely to sign a nondisclosure agreement without receiving substantial disclosure from the seller beforehand); Lemley, *supra* note 111, at 337 (same).

189. Lemley, *supra* note 111, at 337.

190. See Barnett, *supra* note 4, at 798; Lemley, *supra* note 111, at 337 & n.109 (noting that "[b]oth venture capitalists and Hollywood executives . . . are notoriously unwilling to sign nondisclosure agreements before reading business plans or movie scripts"); Bar-Gill & Parchomovsky, *supra* note 14, at 1678 ("Powerful parties . . . often refuse to sign NDAs and instead demand that the disclosing party sign a legal document that releases the powerful party from all liability if the information is somehow disclosed."). Anton and Yao model the circumstances under which an information seller will waive confidentiality rights—in effect a reverse-NDA. They conclude that such waivers help persuade skeptical buyers to participate in the exchange. James J. Anton & Dennis A. Yao, *Attracting Skeptical Buyers: Negotiating for Intellectual Property Rights*, 49 INT'L ECON. REV. 319, 319 (2008).

191. See *supra* note 125 and accompanying text.

192. See *supra* notes 155–59 and accompanying text.

significantly the range of contractual mechanisms that can help facilitate the transfer of information.

Indeed, contracts for the sale of information more closely resemble governance mechanisms than simple transactions.¹⁹³ Because the exchange either requires or can be structured as a series of interactions, contractual governance structures can be erected that support this relationship. Notably, these governance structures do not contemplate vertical integration of the sort typically posited as the alternative to market-based exchange in the absence of reliable solutions to the disclosure paradox.¹⁹⁴

As an example, recall from the previous discussion that the ability to withhold tacit knowledge allows holders of biotech platform technologies freely to disclose the nature of those technologies without fear of misappropriation.¹⁹⁵ The contractual work that remains facilitates the exchange of deeper know-how once the parties have determined that they are interested in further dealings. In 1997, Millennium Pharmaceuticals, at that time a leading biotechnology company with technology centered on genomic analysis, entered into an agreement with the agricultural products giant Monsanto.¹⁹⁶ That deal was the result of an initial negotiation similar to that described above. Monsanto employees toured Millennium facilities as the parties conducted due diligence, learning about the kinds of platform technologies that Millennium possessed, and determining which technologies were potentially of interest.¹⁹⁷ The subsequent contract established a new entity called Cereon, structured as a wholly owned subsidiary of Monsanto.¹⁹⁸ Millennium agreed to provide support to Cereon in utilizing Millennium's platform technologies in return for royalty payments.¹⁹⁹ In order to guard against appropriation of the technology beyond the scope of the agreement, the parties put in place a set of complicated monitoring and

193. I use the term "governance" as it is used in the transaction cost economics tradition to refer to "the ex post support institutions of contract." WILLIAMSON, *supra* note 82, at 29 (emphasis omitted). The questions that branch of contract theory asks include: "What institutions are created with what adaptive, sequential decision-making and dispute settlement properties?" *Id.* I follow Gilson, Sabel, and Scott in adapting this view to the particular problems of contracting in the face of significant uncertainty and information asymmetries. Ronald J. Gilson et al., *Contracting for Innovation: Vertical Disintegration and Interfirm Collaboration*, 109 COLUM. L. REV. 431, 433 n.1 (2009).

194. *Cf.* Barnett, *supra* note 4, at 803–05; Burk & McDonnell, *supra* note 14, at 587–88.

195. *See supra* notes 133–34 and accompanying text.

196. Millennium Pharm., Inc., Current Report (Form 8-K) 3 (Nov. 4, 1997).

197. *See supra* note 15.

198. *See supra* note 196; Millennium Pharm., Inc., Amendment No. 1 to Current Report (Form 8-K/A) 2 (Jan. 30, 1998); *Monsanto Company IPO Overview*, NASDAQ (Oct. 18, 2000), <http://www.nasdaq.com/markets/ipos/company/monsanto-co-new-76144-2630> ("Cereon is our wholly owned subsidiary.").

199. *See* Millennium Pharm. Inc., Amendment No. 1 to Current Report (Form 8-K/A) 42, 49 (Jan. 30, 1998) (discussing the terms of Monsanto's royalty payments to Millennium).

governance mechanisms.²⁰⁰ These mechanisms included joint committees that would meet at regular intervals and a procedure for resolving disputes.²⁰¹ In short, they governed not the terms of the information itself, but the manner in which the parties would interact over the course of the information exchange. The initial exchange was enabled by Millennium's ability to withhold know-how; the contractual terms then specified the conditions for future exchange.

These contracts are similar in nature to the contracts in disaggregated supply chains that Gilson, Sabel, and Scott refer to as "contract[s] for innovation."²⁰² The problem that Gilson, Sabel, and Scott address is different from but analogous to the problem of contracting around the disclosure paradox. They begin with two observations: that supply chains across a wide variety of industries have been disaggregated, and that the pace of technological innovation compels these disaggregated suppliers to collaborate closely to bring new products to market.²⁰³ In the face of significant uncertainty about the final shape that these products will take, buyers and suppliers do more than just enter into arm's-length supply arrangements (or simply vertically integrate). Instead, the transactions that take place among disaggregated firms "involve novel forms of collaboration" and "carefully organized exchanges of information designed to identify and utilize possibilities for innovation."²⁰⁴ The contracts that underlie these relationships establish "elaborate governance mechanisms in lieu of the more familiar risk-allocation provisions of conventional contracts"—and often little else²⁰⁵—through which the parties engage in mutual information sharing and product development over the course of several years.²⁰⁶ Gilson, Sabel, and Scott describe these governance mechanisms as "a rich braiding of formal and informal terms that deters opportunism during the collaborative/learning phase of the contract."²⁰⁷ The contracting challenge that Gilson, Sabel, and Scott confront is how parties can make asset-specific investments to develop new products collaboratively in the face of uncertainty about both one another's capabilities and the final product. The parties overcome the threat of opportunism in such situations by engaging in

200. *See id.* at 23–35, 41–42, 54 (establishing joint committees and teams responsible for coordinating the research program and disclosing information between parties as well as establishing a duty of cooperation between parties).

201. *Id.* at 23–26, 56–60.

202. *See* Gilson, Sabel & Scott, *supra* note 193, at 436.

203. *Id.* at 431.

204. *Id.* at 436–37.

205. *Id.* at 449; *see id.* at 460 (describing an exemplar agreement between John Deere and a supplier that does not specify any supply orders).

206. *Id.* at 472–73.

207. *Id.* at 473; *see also* JOHN HAGEL III & JOHN SEELY BROWN, THE ONLY SUSTAINABLE EDGE: WHY BUSINESS STRATEGY DEPENDS ON PRODUCTIVE FRICTION AND DYNAMIC SPECIALIZATION 91–95 (2005) (describing mechanisms for building "dynamic trust" in the context of loosely coupled process networks).

a collaborative process that both builds trust—and therefore enables the exchange of increasingly sensitive and detailed information about each party’s technical knowledge and capabilities—and raises the switching costs of finding another partner, thereby discouraging defection.²⁰⁸

Parties seeking to transfer complex information face some similar impediments. Unlike contracts for collaborative product development, contracts for the exchange of information may contemplate a single project. But like the Gilson, Sabel, and Scott contracts, they require the development of mechanisms to promote trust and limit opportunism. The exchange of sensitive information requires trust on both sides. Governance mechanisms that elaborate the terms by which parties will structure an ongoing relationship provide a contractual foundation for building that trust over time.

One can also see the “braiding” of legally enforceable obligations with informal obligations in the arrangements that parties seeking to exchange information may make. Returning to the example of pharmaceutical licensing,²⁰⁹ recall that the negotiations between large pharmaceutical manufacturers and biotechs are carried out in stages. In the first stage, the parties engage in disclosure of information without any contractual protections. Should the parties prove interested in further disclosures, however, they typically will sign a NDA. The NDA creates binding legal obligations, though litigation over these agreements is rare. These NDAs are signed more for the signal they send to the parties and to outsiders about the seriousness of the ongoing negotiation than for the actual contractual protection provided. Similarly, when the parties have reached basic agreement on the contours of the deal and are ready to conduct in-depth disclosures and exchange of information as part of their mutual due diligence, they will sign a “term sheet.” This term sheet may or may not be a binding contract, but it again signals that the negotiations have reached a stage where serious disclosures are being made. At each stage of the process, the public signaling provided by the parties’ willingness to sign a contract operates to increase that party’s liability not in litigation, but in the court of public opinion in the relevant norm community.²¹⁰ In this manner, the parties braid together contract-based mechanisms and informal norms based on trust and reputation signaling to accomplish a deepening exchange of information over time.

208. See Gilson et al., *supra* note 193, at 472 (“The contracting problem is to craft a structure that (1) induces efficient, transaction-specific investment by both parties; (2) establishes a framework for iterative collaboration and adjustment of the parties’ obligations under conditions of continuing uncertainty . . . ; and (3) limits the risk of opportunism that could undermine the incentive to make relation-specific investments in the first place.”).

209. See *supra* note 15 and accompanying text.

210. See *infra* note 232 and accompanying text (discussing reputational harms as a mechanism for inducing disclosure).

3. *Norms.*—Legal scholars have long understood that norms as well as law play a significant role in shaping private behavior.²¹¹ In the production of intellectual goods, a well-developed literature seeks to understand what incentives individuals have to innovate in the absence of intellectual property.²¹² Norms can support and regulate the exchange of information as well as its production. As the previous Part demonstrated, parties have some ability self-consciously to structure the information flows around their products and ideas. These flows of information are often shaped by norms in the industries and communities of which information holders are a part.

Take, for example, the classic comparison of technology clusters in Silicon Valley and Route 128 in Massachusetts.²¹³ Saxenian was the first to explain that the relative success of Silicon Valley was attributable to that area's comparatively efficient transfer of useful knowledge between and among firms.²¹⁴ In Saxenian's account, subsequently followed by Gilson and Hyde, the critical driver of economic performance in Silicon Valley was an industrial organization that encouraged the free flow of information between firms. This allowed firms to develop an industrial market structure particularly conducive to innovation. As Gilson writes, Silicon Valley entrepreneurs "moved between companies, founded start-ups, supplied former employers, purchased from former employees, and in the course of

211. See, e.g., ROBERT C. ELLICKSON, *ORDER WITHOUT LAW: HOW NEIGHBORS SETTLE DISPUTES* 282–83 (1991) (concluding that norms influence private behavior more than law in "some spheres of life").

212. There are at least two strands to this literature. The first explores the mechanisms that underlie alternative production systems that are based neither in markets nor hierarchies. The seminal contribution to understanding commons-based peer production is YOCHAI BENKLER, *THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM* (2006). The second strand explores intellectual property's "negative space," that is, areas of intellectual production that succeed in the absence of intellectual property. See, e.g., Dotan Oliar & Christopher Sprigman, *There's No Free Lunch (Anymore): The Emergence of Intellectual Property Norms and the Transformation of Stand-Up Comedy*, 94 VA. L. REV. 1787, 1859–62 (2008) (discussing informal social norms that protect stand-up comedians' material); Emmanuelle Fauchart & Eric von Hippel, *Norms-Based Intellectual Property Systems: The Case of French Chefs*, 19 ORG. SCI. 187, 188 (2008) (discussing implicit social norms that protect French chefs' recipes). Unlike the former, the discussion here is concerned primarily with exchange rather than production, though the two admittedly go hand-in-hand at times; unlike the latter, the discussion here is concerned not with proprietary norms but with the norms that encourage and support exchange.

213. See, e.g., ANNALEE SAXENIAN, *REGIONAL ADVANTAGE: CULTURE AND COMPETITION IN SILICON VALLEY AND ROUTE 128*, at 1–4 (1994) (describing the differences in productive organization between Silicon Valley and Route 128); Ronald J. Gilson, *The Legal Infrastructure of High Technology Industrial Districts: Silicon Valley, Route 128, and Covenants Not to Compete*, 74 N.Y.U. L. REV. 575, 586–94 (1999) (same); see generally ALAN HYDE, *WORKING IN SILICON VALLEY: ECONOMIC AND LEGAL ANALYSIS OF A HIGH-VELOCITY LABOR MARKET* (2003) (arguing that the culture of start-ups in Silicon Valley is the "key influence" on factors that distinguish it from Route 128).

214. See SAXENIAN, *supra* note 213, at 34–37 (explaining that Silicon Valley was "distinguished by the speed with which technical skill and know-how diffused within a localized industrial community" and that the diffusion of knowledge "enhanced the viability of Silicon Valley start-ups"); see also Gilson, *supra* note 213, at 586–94 (summarizing and agreeing with Saxenian's basic account).

their careers developed personal and professional relationships that cut across companies and competition.²¹⁵ In Massachusetts's high-tech corridor along Route 128, by contrast, firm mobility was low and the flow of information was much more tightly controlled.²¹⁶

Critically, the regulation of information flows in these two cases was determined by a combination of norms and law. Gilson argued famously that legal rules drove norms.²¹⁷ In his view, the unenforceability of covenants not to compete in employment contracts as a matter of California state law marked a critical legal difference with Massachusetts that allowed the norms of employee mobility and easy information exchange to flourish.²¹⁸ Hyde, by contrast, argued that the norms shaped the applicable law.²¹⁹ In all events, the interaction of a complex set of cultural and legal institutions determined—in two different geographies—whether and to what extent valuable knowledge was shared and shaped the resulting economic effect.

The story of Silicon Valley and Route 128 illustrates important ways in which norms can affect information sharing. I highlight three that may be of particular importance in overcoming the disclosure paradox: norms of reciprocity, attribution, and reputation. These norms support the exchange of information by serving as limitations on opportunism.

In many communities of technologists and entrepreneurs, there is a strong norm favoring free exchange of information based not on altruism or idealism, but on a calculation that reciprocity is to everyone's advantage. Venture capitalists, for example, describe the value of "being in the mix." Industry participants who share information about their businesses generate interest among investors and potential partners. Similarly, idea sharing among the entrepreneurial community leads to opportunities for collaboration or other joint efforts that may yield important business advantages. Overprotection of intellectual assets in that environment actually operates as a competitive disadvantage.²²⁰

Management scholars have described at least two aspects of this norm in greater detail. The first is the need for learning in addition to innovation. Cohen and Levinthal explain that investment in R&D is useful to firms not only to generate new information, but to allow firms to "identify, assimilate, and exploit knowledge from the environment."²²¹ Learning, in other words, is just as important as innovation. Firms derive a benefit, they argue, from engaging in research and development despite the fact that the knowledge

215. Gilson, *supra* note 213, at 590.

216. *Id.* at 591–92.

217. *Id.* at 578.

218. *Id.* at 609.

219. HYDE, *supra* note 213, at 15–24.

220. See Gans & Stern, *supra* note 187, at 343–45 (describing conditions necessary for development of reputation-based markets for idea exchange).

221. Wesley M. Cohen & Daniel A. Levinthal, *Innovation and Learning: The Two Faces of R&D*, 99 *ECON. J.* 569, 569 (1989).

generated may be partially—or even mostly—appropriable by others because such engagement improves firms’ “absorptive capacity.”²²² The need to build absorptive capacity is directly related to the complexity and transferability of information in the relevant technological area.²²³ In areas marked by inherently tacit or difficult-to-transfer knowledge,²²⁴ generating spillovers helps a firm build its own capacity to take advantage of others’ spillovers.²²⁵ The incentive to be “in the mix” is therefore correlated with the need to accomplish more difficult transfers of information.

Powell adds to this analysis by demonstrating that networks of learning, in which information is freely exchanged among participants in the network, develop in response to the need to understand and absorb widely dispersed and quickly evolving information.²²⁶ “When there is a regime of rapid technological development, research breakthroughs are so broadly distributed that no single firm has all the internal capabilities necessary for success.”²²⁷ In that environment, “the locus of innovation is found in a network of interorganizational relationships” that require reciprocity for ongoing collaboration.²²⁸ Firms that attempt to restrain the flow of knowledge often will find themselves excluded from the network by operation of the reciprocity norm. A Silicon Valley firm, for example, that acquires a reputation for suing its employees when they take knowledge elsewhere will find it hard to recruit and retain talent.²²⁹

At times, this norm of reciprocity is supported by a norm of attribution, at least in cases where the valuable currency that needs protection is credit for one’s work. Academic discourse is a critical example here. Norms of sharing have long been part of the scientific and academic process.²³⁰ But

222. *Id.* at 593–94.

223. *See id.* at 593 (suggesting that “the characteristics of knowledge that affect the ease of firm learning” influence the degree of investment in research and development).

224. *See supra* notes 136–37 and accompanying text.

225. *See* Frischmann & Lemley, *supra* note 164, at 268–69 (describing a “virtuous cycle” created by spillovers that increases the overall investment in research and development).

226. *See* Walter W. Powell et al., *Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology*, 41 ADMIN. SCI. Q. 116, 143 (1996) (explaining that networks form to access relevant knowledge that is widely dispersed and rapidly expanding); *see also* Jason Owen-Smith & Walter W. Powell, *Knowledge Networks as Channels and Conduits: The Effects of Spillovers in the Boston Biotechnology Community*, 15 ORG. SCI. 5, 6 (2004) (explaining that networks improve rates of learning and access to knowledge).

227. Powell et al., *supra* note 226, at 117.

228. *Id.* at 119.

229. *See* SAXENIAN, *supra* note 213, at 41 (noting that Silicon Valley was far less litigious than other parts of the country); *see also* Michael J. Madison et al., *Constructing Commons in the Cultural Environment*, 95 CORNELL L. REV. 657, 696–97 (2010) (describing a similar phenomenon in patent pools).

230. *See, e.g.*, Rebecca S. Eisenberg, *Proprietary Rights and the Norms of Science in Biotechnology Research*, 97 YALE L.J. 177, 180–84 (1987) (highlighting the norms of community and sharing in scientific research); Arti Kaur Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 NW. U. L. REV. 77, 88–94 (1999) (noting that traditional scientific norms promote freely available information).

ideas and information are the stock-in-trade among academics. To protect the valuable asset associated with being the first to generate or publicize information, academics have long relied on a norm of attribution.²³¹ Attribution (and its counterpart, a strong antiplagiarism norm) effectively allows academics to capture value from their contributions to the literature—in the form of enhanced reputation, career prospects, etc.—while simultaneously disclosing their intellectual output to the broader community.

Finally, these norms also are supported by reputational constraints. It is well understood, for example, that venture capital firms overcome the disclosure paradox in part by relying on their reputations.²³² These firms require access to private information in order to complete financing deals; their access to such information depends critically on their reputations as repeat players. A firm that divulges private information is not likely to find many entrepreneurs seeking financing from it in the future. There is no reason to believe that venture capital is *sui generis* in this regard; reputational effects can and do play a role in information exchange more broadly.²³³ Indeed, reputation is a critical part of the operation of licensing deals between pharmaceutical and biotechnology companies. The reputation effect arises because consolidation in the pharmaceutical industry has left relatively few large firms capable of carrying out the development and marketing necessary to commercialize the products of biotechnological research. These few firms are therefore the primary “customers” of biotech firms seeking to license their potential targets. At each stage of the negotiation over the potential licensing of a biotechnology-based compound, the likelihood of reputational harm to a pharmaceutical company that misappropriates sensitive information increases. At each step of the process, the additional reputational risk that the pharmaceutical company takes on increases the ability of the biotechnology company to make further disclosures.

4. *Alternative Sources of Appropriability.*—Certain features of the broader business and legal environment can also support the strategies described above. These mechanisms operate in the background, insofar as they provide the parties with additional assurance that they can retain some value despite their disclosures. They therefore form an important part of the story about how transactions in information can take place, even in the absence of intellectual property rights.

231. Catherine L. Fisk, *Credit Where It's Due: The Law and Norms of Attribution*, 95 GEO. L.J. 49, 81–85 (2006); see also Oliar & Sprigman, *supra* note 212, at 1829–30 (describing the attribution norm in stand-up comedy).

232. See Bar-Gill & Parchomovsky, *supra* note 14, at 1689 & n.156; Ronald J. Gilson, *Engineering a Venture Capital Market: Lessons from the American Experience*, 55 STAN. L. REV. 1067, 1085–87 (2003) (discussing the benefits of an effective reputation market to support the transfer of discretion between an entrepreneur and venture capital fund).

233. See *supra* notes 20–21 and accompanying text.

There is a significant economic literature that demonstrates that intellectual property is not the only mechanism by which a party can appropriate the gains from its investment in R&D.²³⁴ Innovators can and do rely on a host of other methods to ensure that they can receive an adequate return on their investment. These mechanisms can substitute for intellectual property not only with respect to the generation of *ex ante* incentives to engage in innovative activity, but also in solving the *ex post* expropriation problem that comprises the disclosure paradox.

In his classic work, David Teece explains that innovators have numerous sources of “appropriability”—the “ability to capture the profits generated by an innovation.”²³⁵ These sources vary with the market structure of an industry, business strategy of a firm, and the legal environment in which both operate.²³⁶ While patents often play an important part in firms’ strategies to appropriate the gains from research and development, they rarely allow for perfect appropriability;²³⁷ they are not, therefore, the sole means by which firms profit from innovation.

Teece highlights two alternative sources of appropriability. The first is the first-mover advantage. When an innovator is the first to market, she occupies the entire market for a time.²³⁸ During that time of *de facto* exclusivity, the innovator may directly recoup much of her investment.²³⁹ The innovator may also be able to execute strategies that preserve long-term competitive advantage during the time when the market is relatively uncompetitive. Building a brand name and customer loyalty, for example, or developing a competitive advantage with respect to supplies or manufacturing, could produce appropriable rents for many years after competitors enter the market.²⁴⁰ The second alternative source of appropriability is the ability of owners of complementary assets to leverage their ownership over such assets to charge supracompetitive prices even for unprotected innovations.²⁴¹ Innovators following this strategy rely not on the

234. DAVID J. TEECE, *MANAGING INTELLECTUAL CAPITAL: ORGANIZATIONAL, STRATEGIC, AND POLICY DIMENSIONS* (2000); Levin et al., *supra* note 32; COHEN ET AL., *supra* note 32.

235. David J. Teece, *Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy*, 15 RES. POL’Y 285, 287 (1986).

236. *Id.*

237. *See supra* notes 166–68 and accompanying text.

238. *See Teece, supra* note 235, at 286 (noting that a “first-to-market advantage” can be “translated into a sustained competitive advantage which either creates a new earnings stream or enhances an existing one”).

239. *Id.*; *see also* Roger A. Kerin et al., *First-Mover Advantage: A Synthesis, Conceptual Framework, and Research Proposition*, 56 J. MKTG. 33, 34 (1992) (citing Eric von Hippel, *Appropriability of Innovation Benefit as a Predictor of the Functional Locus of Innovation* (Nat’l Sci. Found., Working Paper No. 1084-79, 1984), available at http://pdf.aminer.org/000/326/964/perceived_net_benefit_as_a_measure_of_is_success_and.pdf) (stating that the first mover may be in a “position to gain higher profits than would be possible in a competitive marketplace”).

240. TEECE, *supra* note 234, at 30, 121–22.

241. ARORA ET AL., *supra* note 39, at 116–17; Teece, *supra* note 235, at 288–90.

innovation for their competitive advantage, but on their unique ability to control use of the innovation through other means.

Each of these alternative mechanisms for appropriating the gains from research and development can also support information exchange by enabling parties to retain value derived from their information even after disclosure. In biotechnology, for example, disclosure of the structure of a molecule to a pharmaceutical company does not automatically divest the biotech of competitive advantage. It is already several years farther along the path towards development and marketing. Given the lengthy and complicated FDA approval process, a competitor in possession even of the structure of the molecule may have difficulty catching up.²⁴²

Or consider the sources of value in software.²⁴³ Both entrepreneurs and venture capital investors agree that the value of a potential startup is determined primarily not by the idea motivating the business but by the ability of the putative company to execute the idea. Early-stage venture capitalists may see up to 1,000 companies in a year, and make investments in twenty to thirty of them. Among these business proposals, there will be much overlap and repetition. The likelihood is that a venture capitalist will see multiple iterations of the same idea. The source of value creation in that industry, however, is not primarily in the idea. Rather, it is in the execution. Venture capitalists certainly are interested in creative solutions to problems that represent good market opportunities, but most of their due diligence time is spent evaluating the entrepreneur and her team, and determining whether she can effectively bring the idea to fruition. Because the idea itself is of relatively lower value compared with the complementary assets that the entrepreneur and her team bring to the table, the entrepreneur can potentially disclose the idea to potential investors or collaborators and rely upon her superior skills to prevent misappropriation.

Industrial structure can also provide a source of appropriability. Anton and Yao demonstrate that under certain conditions an information holder may still profit from her disclosure of the information prior to coming to terms. Specifically, they model a scenario in which a financially weak inventor discloses the information to a potential partner, and extracts surplus by threatening to disclose the invention to the partner's competitors.²⁴⁴ If the inventor has sufficient financial resources, she may be able to negotiate a contract *ex ante* by bargaining some of those resources should the idea prove unworkable.²⁴⁵

Finally, some degree of appropriability also can be provided by legal doctrines other than positive law, property rights-style intellectual property.

242. See *supra* note 158.

243. See *supra* note 160.

244. James J. Anton & Dennis A. Yao, *Expropriation and Inventions: Appropriable Rents in the Absence of Property Rights*, 84 AM. ECON. REV. 190, 191–92 (1994).

245. *Id.* at 191, 203.

Trade secrecy is the most likely candidate to replicate the functions of intellectual property, especially insofar as it grants certain limited entitlements to the holders of information that cannot be protected through conventional patent or copyright.²⁴⁶ As Mark Lemley points out, the property-like aspects of trade secrecy can help overcome Arrow's paradox in much the same way that patent or copyright can.²⁴⁷ Even in the absence of an explicit NDA, courts can infer a confidential relationship in certain circumstances, and thereby hold one party liable for misappropriation of a trade secret.²⁴⁸ Some states also provide direct protection for the exchange of ideas under the rubric of "idea submission law."²⁴⁹ Although the details vary by state, these doctrines generally create liability for the misappropriation of ideas divulged in the course of soliciting development, when such ideas are sufficiently concrete and novel. Although the various doctrines that states apply are inconsistent with one another and inconsistently applied,²⁵⁰ they too form the basis for an argument that *ex post* liability may confer enough protection to sustain a negotiation for the sale of information.

Some authors have been skeptical of trade secrecy's efficacy in promoting exchange. Bar-Gill and Parchomovsky, for example, criticize the use of trade secrecy on the ground that it is not a right in rem, but merely in personam.²⁵¹ But in personam rights protected through liability rules are the traditional tools for ensuring the smooth operation of commercial exchange.²⁵² So the question with respect to exchange of information is whether liability-rule treatment will depart in meaningful ways from the

246. Lemley, *supra* note 111, at 338–41. Trade secrets are generally defined broadly to include a wide variety of confidential and valuable business information. See UNIF. TRADE SECRETS ACT § 1(4) (amended 1985), 14 U.L.A. 538 (2005) (defining trade secrets as "information . . . that: (i) derives independent economic value . . . from not being generally known to . . . other persons who can obtain economic value from its disclosure or use, and (ii) is the subject of efforts that are reasonable under the circumstances to maintain its secrecy"); RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 39 (1995) (defining trade secrets as "any information that can be used in the operation of a business . . . and that is sufficiently valuable and secret to afford an actual or potential economic advantage over others").

247. Lemley, *supra* note 111, at 336–37.

248. *Id.* at 337.

249. See Bar-Gill & Parchomovsky, *supra* note 14, at 1681–84 (discussing state common law doctrines designed to protect ideas in certain circumstances); Arthur R. Miller, *Common Law Protection for Products of the Mind: An "Idea" Whose Time Has Come*, 119 HARV. L. REV. 705, 718–32 (2006) (advocating more robust protection for ideas); see generally Mary LaFrance, *Something Borrowed, Something New: The Changing Role of Novelty in Idea Protection Law*, 34 SETON HALL L. REV. 485 (2004) (discussing the evolution of idea protection doctrine in New York and New Jersey).

250. See Bar-Gill & Parchomovsky, *supra* note 14, at 1681 (observing that state judicial efforts to "afford protection to ideas" have "resulted in a largely inconsistent and incoherent body of law"); Miller, *supra* note 249, at 718 (noting that state law doctrines "have met heavy resistance, with scholars criticizing their variegated and unpredictable application").

251. See Bar-Gill & Parchomovsky, *supra* note 14, at 1677–78.

252. Cf. Guido Calabresi & A. Douglas Melamed, *Property Rules, Liability Rules, and Inalienability: One View of the Cathedral*, 85 HARV. L. REV. 1089, 1110 (1972) (explaining that liability rules are often used over property rules in order to achieve efficient valuation in a market).

commercial norm. In light of the more detailed conception of information described above, there is at the very least reason to think an appropriately tailored *ex post* remedy for wrongful precontractual use of information may help support contracting even in the absence of *ex ante* property rights.

This Part has demonstrated several flaws with the conventional understanding of the disclosure paradox. Because that understanding is founded upon unrealistic assumptions about the nature of information, it leads to an overly simplistic solution. As an asset and the subject of commercial exchange, information often is neither wholly nonexcludable nor entirely homogeneous. The variegated nature of information gives rise to a number of strategies for ensuring its exchange that the existing literature underappreciates. Sometimes the characteristics of the information itself allow for it to be exchanged without significant threat of appropriation. At other times, parties may employ a range of techniques including, but not limited to, intellectual property protection to disclose information without giving up all of its value. Ultimately, the precise circumstances in which one or another technique may be useful will vary with the characteristics of the information the parties seek to exchange and the legal and business environment in which they seek to exchange it.

III. Using Policy Tools to Promote Information Exchange

As Part II has demonstrated, the conventional account of the disclosure paradox is, at best, a significant oversimplification of the process of exchanging information. Even in the area where one would most expect to see intellectual property playing a core role in facilitating the exchange of highly self-revealing information—pharmaceuticals—there exist both theoretical reasons to believe that intellectual property is not as necessary as many have suggested and at least anecdotal evidence that parties can utilize a variety of non-intellectual-property-based strategies for accomplishing exchange. Indeed, despite the fact that the core asset may be protected by intellectual property, parties still rely on these alternatives.²⁵³ Intellectual property therefore may not be playing the role traditionally ascribed to it; that is, it may not always be *sufficient* for information exchange. And Part II also offered examples where information exchange takes place in the absence altogether of intellectual property. That suggests that intellectual property may not always be *necessary* for information exchange. The utility of intellectual property in facilitating information exchange therefore is just as contingent on specific technological and economic circumstances as that of the other methods described in Part II.

253. This is not to say that intellectual property is useless in pharmaceuticals. In this analysis, I have focused solely on the effects of intellectual property in the market for research inputs rather than in the market for finished products. There is significant evidence to suggest that in fact intellectual property remains highly useful in pharmaceuticals and biotech. *See supra* note 173 and accompanying text.

These conclusions cast doubt upon a core argument in favor of expanding intellectual property. Recall from Part I that the unique economic function that underlies commercialization theory is the linking of ideas and capital or skills. Commercialization theory justifies intellectual property on the ground that it facilitates the development and commercialization of early-stage inventions.²⁵⁴ It does so, in this telling, by solving the disclosure paradox. But if intellectual property does not solve the disclosure paradox in all cases—if, indeed, neither the disclosure paradox nor the intellectual property solution operates as the commercialization theorists predict—then commercialization cannot be a stand-alone justification for intellectual property.²⁵⁵

Two notes of qualification are appropriate here. First, my normative claim is limited. The theory and evidence presented in Part II support the conclusion that commercialization theory rests on assumptions that likely are not justified. It does not support—and I do not draw from it—the conclusion that intellectual property *never* operates to promote commercial exchange or that the commercialization rationale *never* justifies a particular change to intellectual property policy. My argument instead is that commercialization theory cannot justify expanded intellectual property without qualification or in all circumstances. The extent to which a particular change is justified will depend on a complicated social welfare calculus that I begin to sketch in only the broadest of terms in subparts III(A) and III(B) below.

Second, my analysis is limited to the commercialization rationale for intellectual property. I recognize, however, that the policy tools of positive intellectual property law operate across the theories that scholars use to justify those tools. Changes made (or not) with one theory in mind will necessarily impact the operation of the intellectual property system as it relates to other views or theories. Expanding or contracting the scope of intellectual property to achieve a particular policy objective justified by the commercialization theory will have an impact on broader incentives to innovate, and vice versa.

Putting these observations together, the argument for caution presented here is strongest with respect to proposals that seek to introduce intellectual property into areas where it has not previously existed solely on the ground that doing so would facilitate exchange of the newly protected subject matter. In other words, we should be especially cautious about protecting ideas on the ground that doing so will enable a market for their exchange. So too with respect to the more commonly made argument that intellectual property

254. See *supra* subpart I(A).

255. Intellectual property may, of course, be justified on other grounds. See *supra* notes 23–24 and accompanying text. I do not question those grounds for the purpose of this Article. Nevertheless, to the extent that particular arguments for expanding or augmenting intellectual property depend on the commercialization theory alone, the argument in this Part urges caution.

should be broadened and strengthened for a variety of early-stage inventions and creations.²⁵⁶

That said, the question remains what, if anything, policymakers can do to promote robust markets for information exchange. After all, effective exchange of information for the purpose of development and commercialization is critical to innovation.²⁵⁷ The remainder of this Part lays out some of the considerations that may ultimately guide any policy analysis. I do not make the claim here that the mechanisms described in Part II are better or worse than intellectual property as a matter of social welfare. There is simply not enough data to draw any conclusions about the relative social welfare benefits of the various mechanisms that parties can use to minimize or overcome the disclosure paradox. The social welfare analysis is complicated and ultimately turns on the particular technological, legal, and business circumstances surrounding the proposed exchange. Determining the conditions under which one or another policy tool may be socially optimal therefore requires a deeper empirical understanding of the dynamics of information exchange across different industries and geographies.

A. *Costs and Benefits*

In the basic social welfare calculus, any given policy tool should be adopted if its benefits exceed its costs. As described earlier, the proponents of commercialization theory have identified a static benefit to intellectual property—it reduces the transaction costs associated with exchanging information.²⁵⁸ Part II demonstrates that this benefit may not be as significant as many believe.²⁵⁹ Yet there are likely situations in which intellectual property really is a necessary condition for information exchange. And even shy of that, there will be circumstances in which intellectual property offers a less costly solution to the disclosure paradox than other methods of exchange. Intellectual property allows for standardization of commercial exchange, for example, while contract- and norms-based methods require more costly customization of the interaction.²⁶⁰

But even accepting the benefits of the intellectual property solution as a given, they must still be weighed against the costs. The social welfare costs of intellectual property are well understood. The classic economic analysis of intellectual property posits a tradeoff between static costs and dynamic benefits.²⁶¹ The static cost arises from the fact that the exclusive right

256. *Cf. supra* notes 68–93 and accompanying text.

257. *See supra* notes 66–67 and accompanying text.

258. *See supra* note 97 and accompanying text.

259. *See supra* section II(B)(1).

260. *See Kieff, supra* note 77, at 333–34 (arguing that IP regimes are “fairly effective in facilitating the coordination among complementary users of the IP-protected subject matter that can help get it commercialized”).

261. Wu, *supra* note 64, at 131.

provided by intellectual property allows the rights holder to price the intellectual good above marginal cost.²⁶² Deadweight loss results.²⁶³ Usually this deadweight loss is offset by the dynamic benefit of incentives to create the good in the first place.²⁶⁴ With intellectual property, intellectual products may be priced inefficiently, but there will be more of them. This is the classic “access–incentive” tradeoff.²⁶⁵

But intellectual property also entails a further dynamic cost. That cost arises because information is not only an end product, but also an input into future innovation.²⁶⁶ As a result, innovation is cumulative. It is not a one-time activity that produces a new product. It often is an ongoing process of improvement. New innovators build on and improve upon what has come before.²⁶⁷ Intellectual property can interfere with this process in several ways.²⁶⁸ First, as Arrow himself recognized, “The preinvention monopoly power acts as a strong disincentive to further innovation.”²⁶⁹ It generally is easier for a monopolist to rely on monopoly rents than to engage in further product development, as might be necessary in a competitive market.²⁷⁰ Second, intellectual property gives the initial inventor or creator control over potential improvements and new uses of her work.²⁷¹ That “leaves improvers vulnerable to bargaining breakdown, strategic behavior, or valuation error.”²⁷² Simply put, intellectual property allows the rights holder to deny downstream innovators or improvers access to the original work. Finally, a variety of mechanisms may raise the cost of potential improvements. When making new products requires the use of a large number of inputs, each of which is independently protected by intellectual property, the cost of aggregating the rights to engage in downstream production may be prohibitively high.²⁷³ This is the “anticommons” problem that often is

262. *Id.*

263. *Id.*

264. *Id.*

265. For a succinct description of the tradeoff, see *id.* at 131–32 & fig.2.

266. See BENKLER, *supra* note 215, at 37–38 (“The other crucial quirkiness is that information is both input and output of its own production process.”); FRISCHMANN, *supra* note 164, at 270–75.

267. See *supra* note 31 and accompanying text.

268. See, e.g., Lemley, *supra* note 29, at 1060–62 (detailing five categories of costs of intellectual property rights); Merges & Nelson, *supra* note 61, at 870 (arguing that broad patents could discourage much useful future innovation).

269. Arrow, *supra* note 1, at 620.

270. See generally CLAYTON M. CHRISTENSEN, *THE INNOVATOR’S DILEMMA: WHEN NEW TECHNOLOGIES CAUSE GREAT FIRMS TO FAIL* (1997). There is significant controversy in the literature over the question whether monopoly or competition better spurs innovation. There is significant evidence, however, that competition works better in industries marked by significant cumulative innovation. See, e.g., Merges & Nelson, *supra* note 61, at 884–97 (discussing the impact of competition or lack thereof on industries with cumulative innovation).

271. Lemley, *supra* note 29, at 1042.

272. *Id.* at 1060.

273. See Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, in 1 *INNOVATION POLICY AND THE ECONOMY* 119, 121 (Adam B. Jaffe et al. eds.,

thought to arise in biotechnology.²⁷⁴ Relatedly, when patent claims are broad, multiple patents may purport to cover the technology, giving rise to a “patent thicket.”²⁷⁵

Importantly for the purposes of this study, the magnitude of dynamic social welfare losses is likely to be particularly high when intellectual property protection is conferred upon early-stage innovations or ideas. That is because such early-stage products are much more likely to be inputs into downstream research.²⁷⁶ They are therefore more susceptible to the pathologies described above. Perhaps unsurprisingly, then, intellectual property traditionally has declined to protect mere ideas.²⁷⁷ Bar-Gill and Parchomovsky make a strong case against departing from that tradition, arguing that the costs of doing so far outweigh the potential benefits.²⁷⁸

Of course, the policy tools other than intellectual property have their own social welfare profiles as well. For present purposes the benefits may be assumed to be roughly similar—reducing the transaction costs of exchanging information. From a static perspective, each of the mechanisms described in subparts II(A) and II(B) involve some restriction on the availability of information that would otherwise be priced at marginal cost,²⁷⁹ and each entails some dynamic cost to the extent that information is not freely available for use as an input. That is, the quid pro quo of the patent system requires publicizing the protected information rather than keeping it secret. Private ordering may result in less information ultimately in the public domain.²⁸⁰ The costs of implementing and administering the mechanisms

2001) (highlighting the dangers of many broad patents to innovation as companies may encounter difficulties when attempting to invent around existing patents).

274. See Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 *SCIENCE* 698, 698 (1998) (discussing the dangers when “multiple owners each have a right to exclude others from a scarce resource and no one has an effective privilege of use”).

275. See Shapiro, *supra* note 273, at 120 (discussing the creation of a “dense web of overlapping intellectual property rights that a company must hack its way through in order to actually commercialize new technology”).

276. See Bar-Gill & Parchomovsky, *supra* note 39, at 409–10.

277. *Id.* at 404 (“[P]atent law traditionally did not afford protection to mere ideas . . .”).

278. See *id.* at 408–12 (arguing that protection of mere ideas will result in a reduction in idea development because idea conceivers will have too much bargaining power).

279. A separate objection to the increased use of private ordering is that it effectively allows protection beyond the scope of congressionally authorized patent and copyright systems. But the law has long enabled protection beyond the scope of congressional legislation. See generally *Kewanee Oil Co. v. Bicron Corp.*, 416 U.S. 470 (1974) (discussing the states’ power to enact intellectual property laws and regulations so long as they are not in conflict with the operation of laws passed by Congress).

280. See Fromer, *supra* note 178, at 581 (arguing that rules on how courts interpret patents incentivize patentees to write unhelpful descriptions which maximize the scope of the patent at the expense of effectively conveying technical information). But see Christopher A. Cotropia & Mark A. Lemley, *Copying in Patent Law*, 87 *N.C. L. REV.* 1421, 1440–58 (2009) (arguing that this is unlikely to be a significant concern because independent invention rather than copying is the primary driver of patent infringement litigation).

described above may in some circumstances be higher—at least as to the particular parties involved, if not to the public more broadly—than the costs of complying with positive law intellectual property systems. There is reason to believe, however, that the dynamic social welfare costs of the non-property mechanisms described in Part II will be lower than those of traditional intellectual property regimes. That is because exclusive rights regimes like the patent system assign a right to the invention that operates to preclude independent invention. It is a right as against the world. The protection conferred by the other mechanisms, by contrast, operates solely in the context of a commercial relationship.

B. *Dynamic Interactions*

Further complicating the social welfare analysis is the fact that the phenomena and solutions described in Part II are likely to interact in complex ways. To the extent, for example, that the availability of patent protection is curtailed, this may lead inventors to favor less self-disclosing forms of knowledge codification.²⁸¹ Contrariwise, strengthening the alternatives available to inventors may detract from the attractiveness of the patent system.

That basic dynamic varies with the nature of the information that the parties seek to exchange. The choice between information-flow design and patent protection, for example, depends heavily on how easy or hard it is to engage in information-flow design. To the extent that the degree of disclosure on the face of a particular product is less endogenous—less easily controlled, that is, by the information holder—then inventors' incentive likely is to design for less disclosure wherever feasible and then rely on the patent system when they are faced with no other choice.²⁸²

But there is another aspect to information-flow design to consider—homogeneity. As described above, the heterogeneity of information means that information flow can be self-consciously manipulated even when the underlying asset is self-disclosing. This is one of the central insights of the pharmaceutical example—despite the highly self-disclosing nature of the molecule, the parties generated information *about* the molecule that enabled them to engage in staged negotiations without full disclosure. Finally, consider that intellectual property can be layered into this scheme as well. The molecule in the pharmaceutical example was protected by patent, though the patents appeared to play little part in the information exchange.²⁸³ Of

281. Cf. J. Jonas Anderson, *Secret Inventions*, 26 BERKELEY TECH. L.J. 917, 960–69 (2011) (developing a framework for evaluating the choice between trade secrecy and patents); Lemley, *supra* note 111, at 340–41 (describing the influence of legal rules on the choice between patent and trade secret protection).

282. Available empirical evidence suggests that this is in fact the case. See Cohen et al., *supra* note 32, at 13–14; Levin et al., *supra* note 32, at 805.

283. See *supra* notes 171–73 and accompanying text.

course, the patent in that example likely served other purposes,²⁸⁴ but it is not hard to imagine a scenario where patent protection operates as an overlay on a system of partial information disclosure or other private ordering. In that circumstance, all of the social costs of both intellectual property and the non-property mechanisms may be incurred without a similar doubling up of social benefits.

Part of the difficulty in sorting out these effects is that, as the examples described above demonstrate, it is unclear whether intellectual property protection and non-property-based mechanisms are acting as complements, substitutes, or duplicates. In some circumstances, the various mechanisms work in concert to produce exchange.²⁸⁵ Examples of this dynamic include the use of tacit information combined with contracting for deeper teaching and exchange in platform technology deals, or the complementarity of contracts and norms in pharmaceutical development. Sometimes they may act as economic substitutes, as when a highly self-disclosing product for which other information-flow design is unavailable forces a choice between patents and secrecy. But sometimes these mechanisms may simply be layered on one another with little additional social benefit. That latter circumstance is of particular concern with respect to proposals to introduce intellectual property into areas where it does not currently apply on the ground that doing so will increase the efficiency of transactions in that area. If the parties operating in the relevant field of innovation have already developed mechanisms for exchange, and they continue to utilize those mechanisms in addition to securing intellectual property protection, then welfare loss is highly likely.

C. *The Need for Empirical Research*

As the analysis above demonstrates, the conditions under which one or another mechanism for overcoming the disclosure paradox is optimal are likely to vary significantly with the specific circumstances of the information exchange. The complexity of the social welfare analysis should make clear the necessity of further empirical research into the mechanisms that parties use to accomplish transactions in information. In this Article, I have outlined a framework for thinking through the various mechanisms that parties might use to facilitate the exchange of valuable information and have populated that framework with examples to demonstrate that these mechanisms actually are utilized in at least some cases. But in order to evaluate which mechanisms might be more favorable than others in particular circumstances—and in

284. See *supra* notes 253, 255.

285. See, e.g., Jonathan M. Barnett, *The Illusion of the Commons*, 25 BERKELEY TECH. L. J. 1751, 1795–96 (2010) (explaining that historic guilds were successful in generating low transaction cost exchanges of information by denying intellectual property rights at the member level, while continuing to facilitate innovation because of the exclusive intellectual property rights received at the group level).

order to evaluate potential policy interventions—more data is needed about the way that information exchange works and the prevalence and frequency with which information holders make use of the various alternatives available to them.²⁸⁶

Existing empirical work provides some clues. In industry surveys, several economists have concluded that patents play a lesser role in appropriating the gains from research and development than do first-mover advantages, ownership of complementary assets, and other such mechanisms.²⁸⁷ The more recent Berkeley Patent Study points in a slightly different direction.²⁸⁸ In that survey of entrepreneurs, the authors found that although patents provide mixed to weak incentives to engage in innovation, they often help start-ups to secure financing.²⁸⁹ Importantly, however, they find that patents link ideas to capital not by overcoming the disclosure paradox, but by providing potential funders with an appropriable asset in industries like biotech where patents are particularly important²⁹⁰ or by providing signals about the quality of the company's management or technology portfolio.²⁹¹ The Berkeley study nevertheless finds that the importance of patents to attracting startup capital varies by industry,²⁹² and that in at least some industries "patenting may not be a necessary condition for access to entrepreneurial capital."²⁹³ Ronald Mann, in a qualitative study of the software industry, finds that patent protection is usually not important in early-stage financing decisions, but takes on greater importance in later-stage companies.²⁹⁴ This suggests that the risk of appropriation in early-stage software deals is sufficiently small that the disclosure paradox can be overcome without intellectual property.

As Lemley points out, it often is difficult to find data on the role that intellectual property plays in the processes of technology transfer and

286. *Accord* Lemley, *supra* note 4, at 748 (acknowledging "licensing rationale for patent law" but concluding that "whether it is true is ultimately an empirical question").

287. *See generally* COHEN ET AL., *supra* note 32; Levin et al., *supra* note 32.

288. Stuart J.H. Graham et al., *High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Study*, 24 BERKELEY TECH. L.J. 1255 (2009).

289. *See id.* at 1303–08 ("[I]nvestors of many types value patents as an input into their investment decision.").

290. *See id.* at 1305 ("A reason why patents are so important in the biotechnology industry in particular is that, when one makes a biotech investment, fundamentally one is making an IP investment.").

291. *See id.* at 1306–07 (reporting that a survey of investors showed that some investors felt patents are a sign of managerial sophistication, while other investors suggested that patents signal quality). This finding is consistent with Clarisa Long's hypothesis that various signaling effects of patents offer another static social welfare benefit. *See generally* Clarisa Long, *Patent Signals*, 69 U. CHI. L. REV. 625 (2002) (discussing the valuable role patents can play as signaling mechanisms and in reducing information costs in capital markets).

292. Graham et al., *supra* note 288, at 1308–09.

293. *Id.* at 1305.

294. Ronald J. Mann, *Do Patents Facilitate Financing in the Software Industry?*, 83 TEXAS L. REV. 961, 981–85 (2005).

licensing.²⁹⁵ Licensing agreements usually are confidential and, as the discussion above demonstrates, much of information exchange takes place outside of the context of formal contract or legal proceedings.²⁹⁶ Future research to determine how the complex set of mechanisms and factors described in this Article interact with one another and, therefore, where policy interventions to promote markets for the exchange of information might be fruitful, must necessarily be qualitative in nature. This Article therefore provides a useful framework for case studies and qualitative interview-based work that will follow.²⁹⁷

Conclusion

Robust markets for the exchange of information are a critical driver of innovation and economic growth. For ideas to benefit society, they must be developed and commercialized. And in order for development and commercialization to take place, ideas must be linked with sources of capital and skills. In this Article, I have demonstrated that intellectual property is not necessary to forge those links. Instead, the complex nature of information goods gives rise to a host of strategies that, used alone or in combination, enable the exchange of commercially significant information. Given the potentially high costs of intellectual property, this complexity counsels against reflexive strengthening of existing intellectual property regimes to facilitate commercialization. Instead, policy interventions that seek to promote transactions in information must be made with a more complete understanding of both the social welfare trade-offs involved in different strategies and the specific business and legal environments in which information transactions take place. Reaching that understanding is fundamentally an empirical endeavor that I reserve for future work.

295. Lemley, *supra* note 4, at 748.

296. *Id.*

297. Cf. Michael J. Madison et al., *Constructing Commons in the Cultural Environment*, 95 CORNELL L. REV. 657 (2010) (offering a framework for qualitative research into cultural commons).

Solving the Patent Settlement Puzzle

Einer Elhauge* & Alex Krueger**

Courts and commentators are sharply divided about how to assess “reverse payment” patent settlements under antitrust law. The essential problem is that a PTO-issued patent provides only a probabilistic indication that courts would hold that the patent is actually valid and infringed, and parties have incentives to structure reverse payment settlements to exclude entry for longer than this patent probability would merit. Some favor comparing the settlement exclusion period to the expected litigation exclusion period, but this requires difficult case-by-case assessments of the probabilities of patent victory. Others instead favor a formal “scope of the patent” test that allows such settlements for nonsham patents if the settlement does not delay entry beyond the patent term, preclude noninfringing products, or delay nonsettling entrants. However, the formal scope of the patent test excludes entry for longer than merited by the patent strength, and it provides no solution when there is either a significant dispute about infringement or a bottleneck issue delaying other entrants.

This Article provides a way out of this dilemma. It proves that when the reverse payment amount exceeds the patent holder’s anticipated litigation costs, then under standard conditions the settlement will, according to the patent holder’s own probability estimate, exclude entry for longer than both the expected litigation exclusion period and the optimal patent exclusion period, and thus will both harm consumer welfare and undermine optimal innovation incentives. Further, whenever a reverse payment is necessary for settlement, it will also have those same anticompetitive effects according to the entrant’s probability estimate. This proof thus provides an easily administrable way to determine when a reverse payment settlement is necessarily anticompetitive, without requiring any probabilistic inquiry into the patent merits. We also show that, contrary to conventional wisdom, patent settlements without any reverse payment usually (but not always) exceed both the expected litigation exclusion period and the optimal patent exclusion period, and we suggest a procedural solution to resolve such cases.

* Petrie Professor of Law, Harvard Law School. Professor Elhauge has been an expert witness on both the plaintiff and defense side in antitrust challenges to reverse payment patent settlements.

** Executive Director, Legal Economics. Alex Krueger has consulted for the plaintiff side in an antitrust challenge to reverse payment patent settlements. For their helpful comments, we would like to thank Richard Brunell, Terry Fisher, Scott Hemphill, Al Klevorick, David Opperbeck, and the participants in the Harvard Health Law Policy workshop.

Introduction.....	284
I. The Two Relevant Benchmarks: <i>Ex Post</i> Consumer Welfare and the Optimal Patent Reward.....	293
II. Reverse Payments That Exceed the Patent Holder’s Anticipated Litigation Costs.....	297
A. The Proof.....	297
1. <i>Strong Patent</i>	299
2. <i>Weak Patent</i>	301
3. <i>Implications</i>	303
B. Presumption and Limited Grounds for Rebuttal.....	304
1. <i>Establishing Presumption by Comparing Reverse Payment to Anticipated Litigation Costs</i>	305
2. <i>Rebuttal by Showing At-Risk Entrant Is Sufficiently Judgment Proof</i>	307
3. <i>Rebuttal by Proving Other Procompetitive Justifications</i>	309
4. <i>No Rebuttal by Showing Lack of Market Power</i>	310
5. <i>No Rebuttal by Showing Risk Aversion</i>	311
III. Settlements Without Reverse Payments That Exceed Litigation Costs.....	312
A. Proof That Even Settlements with Zero Reverse Payment Are Usually Anticompetitive.....	313
1. <i>Strong Patent</i>	314
2. <i>Weak Patent</i>	317
3. <i>Summary</i>	319
B. Grounds for Rebuttal and Possible Procedural Solution.....	323
IV. Relationship to Prior Scholarship.....	325
Conclusion.....	328
Appendix.....	330

Introduction

Reverse payment patent settlements have led to widespread legal controversy. In such settlements, the patent holder agrees to make a payment to an allegedly infringing potential entrant (called a “reverse” payment because traditionally settlement-payment flow was from alleged infringer to patent holder) and the potential entrant agrees to stay out of the market until a later date.¹ Such settlements have anticompetitive potential because they can exclude entry for longer than the expected litigation exclusion period,

1. These are sometimes called “pay-for-delay” settlements, but we avoid that terminology because it presupposes that the settlement entry date does “delay” entry compared to the expected entry date, which is generally the disputed issue.

which would have reflected the often significant likelihood that the patent holder would have lost.² Indeed, unless constrained by the risk of antitrust liability, settling parties would (no matter how weak the patent) always have incentives to set the settlement entry date at the end of the patent term because that maximizes joint profits (by precluding competition for as long as possible), and they can use the reverse payment to split those joint profits in a way that leaves both better off. However, if antitrust liability could be designed to prevent settlements that exclude entry for more than the expected litigation exclusion period, then reverse payment settlements could theoretically avoid litigation costs without causing any anticompetitive effect.

Such reverse payment settlements have been a huge issue in the multitrillion dollar pharmaceutical industry. But the issue is even bigger than that because reverse payment settlements can occur in *any* market where the patent holder would have greater market power if the entrant were excluded.³

The federal courts of appeals are in utter conflict on when reverse payment settlements violate antitrust law. The Sixth Circuit has held that reverse payment settlements are *per se* illegal.⁴ This is the *categorical illegality* position. Two Eleventh Circuit cases rejected this position, holding that reverse payment settlements violate antitrust law only if the settlement exceeds the exclusionary “scope of the patent,” which means the “settlement cannot be more anticompetitive than litigation” and thus “underscores the need to evaluate the strength of the patent.”⁵ The test in these Eleventh

2. The literature often instead compares the settlement entry date to the “expected entry date” from litigation, but the latter term is imprecise in cases where, absent settlement, the entrant would have entered at risk and thus could have been excluded well after entry if it had later lost the patent litigation. So we use the more precise term “expected litigation exclusion period.” For example, suppose the entrant would enter at risk, the remaining patent term is 10 years, the patent litigation will last two years, and the entrant has a 20% chance of losing and being excluded. The expected entry date is immediate and thus would be exceeded by any settlement that excludes entry for any period at all. But this does not accurately determine whether the settlement harms consumer welfare because the expected litigation exclusion period is 20% of 8 years or 1.6 years. Thus, a settlement that excludes entry for less than 1.6 years does not harm consumer welfare.

3. Delaying entry through a reverse patent settlement is profit maximizing whenever entry reduces the joint profits of the patent holder and entrant, which is true whenever the patent holder has market power that the entrant would constrain to some degree. Further, we show below that if the reverse payment amount exceeds the patent holder’s anticipated litigation costs, then the patent holder must have believed it had market power that the settling entrant would uniquely constrain.

4. *In re Cardizem CD Antitrust Litig.*, 332 F.3d 896, 907–08 (6th Cir. 2003). In this case, Professor Elhaug filed an expert declaration for a generic defendant in which he opposed application of the *per se* rule.

5. *Schering-Plough Corp. v. FTC*, 402 F.3d 1056, 1065–66, 1075–76 (11th Cir. 2005); *id.* at 1071 (finding the evidence un rebutted that the settlement “entry date reasonably reflected the strength of [the patent holder’s] case”); *Valley Drug Co. v. Geneva Pharm., Inc.*, 344 F.3d 1294, 1312 (11th Cir. 2003) (requiring inquiry into whether the settlement terms exceeded patent protections, “considered in light of the likelihood of [the patent holder] obtaining such protections”).

Circuit cases thus turns on a case-by-case assessment of the objective probability that the patent holder would have won, which the court stressed should be determined at the time of settlement rather than by some later discrete outcome.⁶ Call this the *objective probabilistic* scope of the patent test.

The Second Circuit rejected an approach that required case-by-case assessments of patent probabilities as inadministrable.⁷ Instead, it concluded that, unless the patent was a sham or procured by fraud, reverse patent settlements were illegal only if the settlement exceeded the *formal* scope of the patent by delaying entry after the patent expires, precluding noninfringing products, or delaying the entry of nonsettling potential entrants.⁸ This *formal* scope of the patent test was then adopted by the Federal Circuit and another panel of the Eleventh Circuit.⁹

Finally, the Third Circuit has just rejected the formal scope of the patent test, adopting a presumption that reverse payment settlements are illegal unless the defendant shows that “the payment (1) was for a purpose other than delayed entry or (2) offers some pro-competitive benefit.”¹⁰ The Third Circuit held that this presumption could not be rebutted by proof about the merits of the patent suit because the reverse payment itself indicated that the purpose was to delay entry.¹¹ Instead, the Third Circuit indicated that the first rebuttal required proving that the patent holder received sufficient separate consideration to negate the existence of reverse payment and that the

6. *Valley Drug*, 344 F.3d at 1306–07 (holding that such a settlement could thus be proper even if the patent were later held invalid).

7. *In re Tamoxifen Citrate Antitrust Litig.*, 466 F.3d 187, 203–04 (2d Cir. 2006).

8. *Id.* at 212–13; *id.* at 213–16 (holding that the settlement did not exceed the scope of the patent because it did not preclude “non-infringing products,” did not delay other potential entrants, and allowed entry before the patent term expired); *Ark. Carpenters Health & Welfare Fund v. Bayer AG*, 604 F.3d 98, 106 (2d Cir. 2010) (stressing that a “settlement agreement did not exceed the scope of the patent where (1) there was no restriction on marketing non-infringing products; (2) a generic version of the branded drug would necessarily infringe the branded firm’s patent; and (3) the agreement did not bar other generic manufacturers from challenging the patent”).

9. *In re Ciprofloxacin Hydrochloride Antitrust Litig.*, 544 F.3d 1323, 1336–37 (Fed. Cir. 2008); *id.* at 1335 (indicating that if a reverse payment settlement created a bottleneck effect “delaying the entry of other generic manufacturers” or covered “non-infringing” products, then it would clearly lie “outside the exclusion zone of the patent”); *FTC v. Watson Pharm., Inc.*, 677 F.3d 1298, 1312 (11th Cir. 2012). The Eleventh Circuit panel in *Watson* claimed its conclusion was consistent with prior Eleventh Circuit panels, but it never addressed the language in *Valley Drug* saying that the exclusionary scope of a patent turned on “the likelihood of [the patent holder] obtaining [patent] protections.” *Valley Drug*, 344 F.3d at 1312. The *Watson* panel’s effort to reconcile *Schering-Plough* relied on the dubious assertion that *Schering-Plough*’s references to evaluating the “strength of the patent” merely meant the temporal length of the patent. *Watson*, 677 F.3d at 1311 n.8. But the *Watson* panel was reasonably concerned that the FTC’s proposed standard in that case (whether patent victory was unlikely) failed to reliably identify whether the settlement was anticompetitive and required difficult inquiries into the probability that the patent holder would have won. *Id.* at 1312–15.

10. *In re K-Dur Antitrust Litig.*, 686 F.3d 197, 218 (3d Cir. 2012).

11. *Id.*

second condition required proving some unrelated procompetitive benefit.¹² Call this the *presumptive condemnation* approach.

The antitrust enforcement agencies have advocated a similar presumption, but have suggested a broader range of rebuttal. The Department of Justice (DOJ) Antitrust Division has concluded that reverse payment settlements should be presumed unlawful, allowing defendants to rebut that presumption by showing either that (1) the reverse payment amount was “not greatly in excess of avoided litigation costs” or (2) the settlement exclusion period did not exceed the expected litigation exclusion period, given the settlors’ contemporaneous estimates of the likelihood that the patent holder would have won the patent litigation.¹³ The Federal Trade Commission (FTC) has advocated the Third Circuit approach of presuming that reverse payment settlements are illegal unless defendants demonstrate offsetting procompetitive effects.¹⁴ The FTC (like the Third Circuit) has also concluded that this presumption cannot be rebutted by proof of the actual *objective* likelihood that the patent holder would have prevailed.¹⁵ However, the FTC has suggested that maybe this presumption could be rebutted by evidence of the *perceived* probability at the time of settlement.¹⁶ Thus, at least for large reverse payment settlements, the Antitrust Division approach would require (and the FTC approach might permit) case-by-case assessments of the perceived probability of patent victory. Call this the *perceived probabilistic* scope of the patent test.

This split in authority does not simply reflect stubbornness or ideological conflict. There seems to be a real dilemma. The underlying problem is that the mere issuance of a patent by the Patent and Trademark Office (PTO) does not mean a court will hold that the patent is actually valid, let alone hold that another firm’s product infringes it. Indeed, even though patent holders get a presumption of patent validity, they lose 48%–73% of patent cases.¹⁷ On average, then, even without any at-risk entry during

12. *Id.*

13. Brief for the United States in Response to the Court’s Invitation at 10, 22, 28–32, *Ark. Carpenters Health & Welfare Fund*, 604 F.3d 98 (Nos. 05-2851-cv(L), 05-2852-cv (CON), 05-2863-cv (CON)), 2009 WL 8385027, at *10, *22, *28–32. The Antitrust Division used to favor a case-by-case inquiry into whether the patent holder actually would likely have won, but now rejects that sort of objective probabilistic approach in favor of the perceived probabilistic approach described in the text above. *Id.* at 24–27, 26 n.9.

14. *Schering-Plough Corp.*, 136 F.T.C. 956, 987–91, 1000–03 (2003), *vacated*, 402 F.3d 1056 (11th Cir. 2005).

15. *Id.* at 992–98.

16. *Id.*

17. See FED. TRADE COMM’N, *GENERIC DRUG ENTRY PRIOR TO PATENT EXPIRATION: AN FTC STUDY* vi (2002) (“Generic applicants have prevailed in 73 percent of the cases in which a court has resolved the patent dispute.”); ADAM GREENE & D. DEWEY STEADMAN, *RBC CAPITAL MKTS., PHARMACEUTICALS: ANALYZING LITIGATION SUCCESS RATES 1* (2010) (noting that patent holders lose 48% of the cases with generic entrants); Paul M. Janicke & LiLan Ren, *Who Wins Patent*

litigation, the expected litigation exclusion period across all cases is only 27%–52% of the patent term that remains after litigation. And this is the average; weaker patents would lose even more often and thus result in even shorter expected litigation exclusion periods, especially because they might provoke at-risk entry during litigation. Any settlement exclusion period that exceeds the expected litigation exclusion period has anticompetitive effects on consumer welfare because it increases the amount of time consumers must pay supracompetitive prices. Further, if we assume (for the purpose of antitrust analysis) that patent law has been optimally designed, then the odds of patent victory in litigation reflect the extent to which the patent holder should be rewarded with supracompetitive profits. Therefore, a settlement also exceeds the optimal patent exclusion period, and thus undermines optimal innovation incentives, if it excludes rivals for a percentage of the remaining patent period that exceeds the percentage chance of a patent victory.

The *objective probabilistic* scope of the patent test provides a straightforward solution: compare the settlement exclusion period to the expected litigation exclusion period by directly adjudicating the likelihood of patent victory. But that approach requires the very sort of inquiry into the patent merits that settlement is supposed to avoid, thus defeating the point of settlement. Moreover, once the court does investigate the patent merits, it will conclude that the patent holder should either have won or lost, and it may have difficulty calculating the perhaps imponderable probability that another court would have been (in its judgment) wrong.¹⁸ Further, this approach subjects settling parties who honestly believe that their settlement excludes entry for less than the expected litigation exclusion period to the threat of trebled antitrust damages if a court later decides the probability of patent victory was different, which may be affected by hindsight bias if other cases later adjudicate the same patent.

The *perceived probabilistic* scope of the patent test nicely avoids these problems when those perceived probabilities can reliably be ascertained. This may be possible when the parties carelessly record their probability judgments in contemporaneous documents. But if courts regularly depended on the parties' contemporaneous documents, then settling parties would likely stop documenting the true probabilities, and instead document inflated

Infringement Cases?, 34 AIPLA Q.J. 1, 20 (2006) (providing data that demonstrates that patent holders lose approximately 70% of the time).

18. This difficulty might be particularly acute because the Federal Circuit has exclusive jurisdiction over patent cases, while antitrust cases can go to any appellate panel. Thus, the appellate panel hearing the antitrust case might worry it lacks the expertise to predict how the Federal Circuit would decide any patent issues. See *FTC v. Watson Pharm., Inc.*, 677 F.3d 1298, 1314–15 (11th Cir. 2012) (expressing concern that “[t]his Court and the other non-specialized circuit courts have no expertise or experience in the area” of patent law and thus “are ill-equipped to make a judgment about the merits of a patent infringement claim”).

probabilities in order to protect their profitable settlements. Settling parties might also or instead simply offer self-serving testimony about those perceived probabilities. To avoid those problems, courts could critically examine such self-serving assessments, but to do so they would have to rely on an objective probability that would bring us back to the *objective probabilistic* scope of the patent test and all its problems.

One can thus understand the attraction of categorical approaches, but both categorical possibilities have serious problems. The problem with categorical *illegality* is that sometimes a positive reverse payment could be consistent with a socially desirable settlement.¹⁹ The problem with categorical *legality* for reverse payment settlements within the *formal* scope of a nonsham patent is that, if it were the accepted rule, settlements would always exclude entry until the patent expiration date, no matter how weak the patent was, because that would maximize firm profits. Indeed, the Second Circuit has indicated considerable ambivalence about its formal scope of the patent test, acknowledging that it produces the perverse result that the weaker a patent, the more such a rule would produce settlements that benefit the patent holder more than it deserves,²⁰ and in its most recent decision, the Second Circuit suggested that the policy problems were great enough that en banc review was merited to reconsider this rule.²¹ Moreover, the formal scope of the patent test by its own terms provides no guidance about how to assess settlements either when there is a serious dispute about whether the entrant product infringes the patent or when the settlement does create bottleneck effects that delay other entrants. Therefore, the one seeming virtue of the formal scope of the patent test, its apparent certainty, is illusory in most cases, where either infringement or bottleneck issues are seriously disputed.

Further, if the formal scope of the patent test prevails, its anticompetitive potential can be multiplied. In some cases, the parties to a patent dispute could *each* have some nonsham patent that applies to the relevant market. The formal scope of the patent test means such parties would maximize joint profits with settlements that declare the validity of whichever patent ends last, even if the other patent is actually more likely to be valid. They can then split those maximized joint profits with reverse payments to make both settling parties better off, while harming consumer welfare and providing rewards that bear no relation to any innovation. Firms would have incentives to further exacerbate the problem by creating a stream of weak (but nonsham) patents precisely for the purpose of enabling these

19. However, our proof below shows this possibility exists only when the reverse payment amount is lower than the patent holder's anticipated litigation costs, absent a judgment-proof entrant or a separate procompetitive justification.

20. *In re Tamoxifen Citrate Antitrust Litig.*, 466 F.3d 187, 211 (2d Cir. 2006).

21. *Ark. Carpenters Health & Welfare Fund v. Bayer AG*, 604 F.3d 98, 108–10 (2d Cir. 2010).

last-to-expire settlements that preclude competition as long as possible. Such a stream could even allow horizontal competitors to create a chain of reverse payment settlements that span multiple patent periods, trading the monopoly power back and forth between each other and splitting the profits with their counterpart throughout.

Clearly then, adopting the formal scope of the patent test can have disastrous consequences. But how can courts prevent anticompetitive settlements without either using a categorical condemnation approach that deters some socially beneficial settlements, engaging in a direct inquiry into the probability of patent victory that may be imponderable and effectively precludes real settlement of the patent issues, or relying on perceptions of patent strength that the settling parties will predictably exaggerate?

In this Article, we present a proof that solves this puzzle for the lion's share of cases. We begin in Part I by separating out two relevant benchmarks: (1) the expected litigation exclusion period and (2) the optimal patent exclusion period. References to the "probabilistic scope" of a patent could mean either one of these benchmarks, which have generally been conflated in prior work and cases, but in fact the benchmarks differ from each other. The expected litigation exclusion period is the expected time that the rival would be excluded with litigation, accounting for the reality of litigation delays and the fact that at-risk entry during litigation might occur or be deterred. Any settlement that excludes the entrant for more than the expected litigation exclusion period thus harms *ex post* consumer welfare because it subjects consumers to supracompetitive prices for longer than litigation would have. The optimal patent exclusion period, in contrast, equals the probability of patent-holder victory times the remaining patent term (ignoring litigation delays and the issue of at-risk entry), and thus is the exclusion period that the patent holder deserves on the merits according to substantive patent law. Assuming that substantive patent law is optimal, any settlement that excludes the entrant for more than the optimal patent exclusion period undermines optimal innovation incentives. Both benchmarks are thus relevant to policy, but the two can differ from each other. A strong patent deters at-risk entry with certainty during litigation, even though there is a probability of patent loss. Therefore, for strong patents, the expected litigation exclusion period always exceeds the optimal patent exclusion period. In contrast, a weak patent produces at-risk entry with certainty during litigation, even though there is a probability of patent victory. Therefore, for weak patents, the optimal patent exclusion period always exceeds the expected litigation exclusion period.

We then prove in Part II that whenever a reverse payment exceeds the patent holder's anticipated litigation costs, the settlement exclusion period will always exceed *both* the expected litigation exclusion period *and* the optimal patent exclusion period, according to the patent holder's *own* probability estimate. Further, whenever such a reverse payment is necessary for settlement, the settlement will also have both those anticompetitive

effects according to the probability estimates of *both* the patent holder and the entrant. Nor is there any reason to tolerate any reverse payment that is not necessary for settlement because without it the settlement would have provided an earlier entry date less harmful to consumer welfare, while still exceeding the optimal patent exclusion period according to the patent holder's own estimate. This proof thus provides an easily administrable way to determine when a reverse payment settlement is anticompetitive, without requiring any probabilistic inquiry into the patent merits. Unlike prior analyses, this proof does not depend at all on actually knowing what the patent holder or entrant perceive the patent strength to be, and it applies even if patent holders and entrants disagree about the patent strength or the future profitability of the patented product.

Although we formally illustrate our proof using a mathematical model below, the basic logic is as follows. If the reverse payment amount exceeds the patent holder's anticipated litigation costs, then we show that the settlement exclusion period necessarily exceeds the expected litigation exclusion period and the optimal patent exclusion period according to the patent holder's own estimate of the patent strength because otherwise the patent holder would be better off litigating. If the entrant's estimate of patent strength is below the patent holder's, then this settlement exclusion period must also exceed the entrant's estimate of the expected litigation exclusion period and optimal patent exclusion period. If the entrant's estimate of patent strength exceeds the patent holder's, then no reverse payment is necessary for settlement because without any reverse payment the parties could have agreed to a settlement exclusion period that is greater than the litigation exclusion period that the patent holder expects but less than what the entrant expects, which would make both better off than they have would been if they litigated.

Our proof assumes that at-risk entrants are not judgment proof and that the reverse payment does not have some other procompetitive justification. Courts therefore should presumptively condemn reverse payments that exceed the patent holder's anticipated litigation costs, but allow defendants to rebut that presumption by proving either: (a) the entrant would have entered at risk and is judgment proof to a sufficient extent to indicate the settlement exclusion period could or would be within the optimal patent exclusion period or (b) that some other procompetitive justification exists that offsets the anticompetitive effects. Absent one of those rebuttals, the proof holds. One important implication is that, contrary to the recommendations by the DOJ (and perhaps the FTC), defendants should *not* be able to rebut this presumption by arguing that the settlement exclusion period did not exceed the expected litigation exclusion period or the optimal patent exclusion period because our proof precludes that possibility for reverse payments that exceed the patent holder's anticipated litigation costs. Nor should the defendants be able to rebut the presumption by arguing that the patent holder lacks market power because our proof also shows that the patent holder

would never make a reverse payment of that size unless it had the requisite market power.

In Part III, we address patent settlements that set entry dates without using reverse payments that exceed the patent holder's anticipated litigation costs. Many, including the FTC and DOJ, have assumed that settlements with no reverse payments will likely set exclusion periods that equal the expected litigation exclusion period.²² However, we prove that this conventional wisdom is untrue. Although patent settlements without any reverse payment will not *necessarily* exceed the expected litigation exclusion period and optimal patent exclusion period, it turns out that they *usually* will exceed both benchmarks. The magnitude of anticompetitive harm is certainly much smaller without the reverse payments, but that does not alter the fact that those harms are undesirable.

One approach to deal with this problem would be to extend presumptive condemnation to settlements that exclude entry without any reverse payment, but to allow parties to rebut this presumption by showing that their settlement exclusion period did not exceed the expected litigation exclusion period or the optimal patent exclusion period. Sometimes this inquiry can be limited with a market power screen or by bounding the possible probabilities that could satisfy the relevant benchmarks. However, in other cases, this approach would require courts to directly adjudicate the patent strength, which is what courts are generally trying to avoid. If direct inquiry into probabilistic patent strength is too unreliable, then the best substantive solution would be categorical condemnation because the proof shows that most such settlements are anticompetitive. However, the better solution in such cases may be procedural rather than substantive. Because the underlying problem that allows anticompetitive settlements is that patent law does not ordinarily give buyers standing to challenge dubious patents, a possible procedural solution would provide that when such settlements are reached buyers should have standing to challenge the patent's validity.

Finally, in Part IV, we relate our analysis to prior scholarship. Although some prior commentators have conjectured that reverse payments that exceed litigation costs are usually anticompetitive, this conjecture has not previously been proven. Our proof establishes the validity of this conjecture under two benchmarks that prior work had generally conflated. Our analysis also proves that other commentators are mistaken in instead claiming that we can tell when settlements are anticompetitive by determining whether the odds of patent victory exceed some general threshold or by comparing the reverse payment to entrant profits. Further, our proof allows us to more accurately determine the conditions under which reverse payments should be

22. Schering-Plough Corp., 136 F.T.C. 956, 987 (2003), *vacated*, 402 F.3d 1056 (11th Cir. 2005); Brief for the United States, *supra* note 13, at 21–22.

presumptively condemned and the proper grounds for rebuttal. We show that the right benchmark is not all litigation costs, as prior proponents of this conjecture have assumed, but only the patent holder's future anticipated litigation costs at the time of settlement. More important, we disprove the claim by prior proponents of this conjecture that courts should allow rebuttal based on alleged direct proof regarding the likelihood of patent victory, the expected litigation exclusion period, risk aversion, or varying party estimates of patent strength. We also prove that courts need to allow a limited rebuttal for judgment-proof entrants that prior proponents of this conjecture have missed. Finally, we disprove the conventional wisdom in prior scholarship that settlements without reverse payments generally do not cause anticompetitive effects.

I. The Two Relevant Benchmarks: *Ex Post* Consumer Welfare and the Optimal Patent Reward

To determine whether a given patent settlement is anticompetitive, one must focus on two benchmarks: (1) but-for *ex post* consumer welfare and (2) optimal patent rewards for *ex ante* innovation. But-for *ex post* consumer welfare reflects the level of expected consumer welfare that would have resulted had the relevant patent disputes been litigated rather than settled. It is called "*ex post*" consumer welfare because it is calculated assuming that the innovation has already occurred. Because the patent holder can charge a significantly higher price while the potential entrant is excluded from the market, a settlement reduces *ex post* consumer welfare below but-for levels if the settlement excludes the entrant from the market for a larger portion of the patent's remaining life than one would have expected to result from litigation. Thus, any settlement exclusion period that exceeds the expected litigation exclusion period necessarily harms *ex post* consumer welfare.

However, not all things that increase *ex post* consumer welfare above but-for levels are desirable or increase *overall* consumer welfare. For example, refusing to enforce any patent (no matter how valid) would increase *ex post* consumer welfare above but-for levels. But that is only because the *ex post* perspective assumes the innovation has already occurred, when in reality patent protection is often necessary to encourage the innovation *ex ante*.²³ If designed optimally, the patent system will maximize overall consumer welfare by giving patent holders the optimal fraction of *ex post*

23. See John H. Barton, *Patents and Antitrust: A Rethinking in Light of Patent Breadth and Sequential Innovation*, 65 ANTITRUST L.J. 449, 450 (1997) ("The patent-antitrust analysis has always had to take into account and balance benefit to consumers by maintaining the competitive structure of existing markets against benefit to consumers by permitting the intellectual property rights system to provide an incentive for research toward new and improved products.").

total surplus created by their innovations.²⁴ Reducing the patent exclusion period below the optimal level will thus result in an inefficiently low amount of innovation.

Exceeding the optimal patent exclusion period is likewise inefficient for several reasons. First, the economic literature shows that patent profits that exceed the optimal level result in excessive investments in innovation that reduce social welfare compared to the optimal investments in innovation.²⁵ Second, excessive patent protection can produce a net reduction in innovation by precluding subsequent innovations by others.²⁶

Third, settlements that overreward the patent holder with a longer exclusion period than it deserves reduce the *net* reward for true innovation by increasing the reward *more* for less-deserving patents than for more-deserving patents. As the proof below shows and the Second Circuit has already pointed out,²⁷ settlements that exclude entry increase patent-holder profits more for weaker patents than for stronger patents. For example, the holder of a weak patent that is only 5% likely to be deemed a valid innovation could use such a settlement to secure exclusion throughout the entire patent term, even though its patent is 95% likely to be deemed a non-innovation, while the holder of an ironclad patent that is 100% likely to be deemed a true innovation could not increase its exclusion period through settlement because it would already expect 100% exclusion from litigation. Thus, settlements with an excessive exclusion period reduce the net reward for investing in a true innovation that leads to a stronger patent rather than in a pseudo-innovation that leads to a weaker patent. When a firm faces a choice between investing in true innovation or pseudo-innovation, this artificially reduced net reward for true innovation will distort its choice, and can reduce the rate of true innovation because it is generally harder, more costly, or less certain than pseudo-innovation.

24. See SUZANNE SCOTCHMER, INNOVATION AND INCENTIVES 100–03 (2004); Partha Dasgupta & Joseph Stiglitz, *Uncertainty, Industrial Structure, and the Speed of R&D*, 11 BELL J. ECON. 1, 18 (1980); Pankaj Tandon, *Rivalry and the Excessive Allocation of Resources to Research*, 14 BELL J. ECON. 152, 152, 156–57 (1983). Such a system will also maximize overall *total* welfare because competing innovators will keep spending on *ex ante* investments until their investment costs equal their expected *ex post* profits, so that the profits to patent holders wash out *ex ante*.

25. See *supra* note 24.

26. For a theoretical model proving that this is possible, see generally Michele Boldrin & David K. Levine, *A Model of Discovery*, 99 AM. ECON. REV. 337 (2009). For empirical work showing that expanding patent protections have had net negative effects on patent filings and suppressed later innovations, see generally Josh Lerner, *The Empirical Impact of Intellectual Property Rights on Innovation: Puzzles and Clues*, 99 AM. ECON. REV. 343 (2009); Fiona Murray et al., *Of Mice and Academics: Examining the Effect of Openness on Innovation* (Nat'l Bureau of Econ. Research, Working Paper No. 14819, 2009); Heidi L. Williams, *Intellectual Property Rights and Innovation: Evidence from the Human Genome* (Nat'l Bureau of Econ. Research, Working Paper No. 16213, 2010).

27. *In re Tamoxifen Citrate Antitrust Litig.*, 466 F.3d 187, 211 (2d Cir. 2006).

For example suppose that a true innovation will produce a gross patent reward of \$1 billion, but that the *net* reward for this true innovation is only \$400 million because the firm can instead get \$600 million in the same market by creating a pseudo-innovation that it can convert into a long exclusion period using a reverse payment settlement. Suppose further that the true innovation requires a \$500 million investment, but the pseudo-innovation requires no investment. Then the true innovation will be deterred because the excessive reward for the pseudo-innovation reduces the net reward for true innovation below the investment required for it. Therefore, settlements that overreward patent holders with longer exclusion periods than they deserve can actually *decrease* incentives to invest in true innovation.

More generally, by reducing the net reward for investing in stronger patents rather than weaker patents, settlements that provide excessive exclusion periods distort investment choices away from the stronger patents that are more likely to reflect real innovation. In all three ways summarized above, settlements that exceed the optimal patent exclusion period will undermine optimal innovation incentives.

For the purpose of antitrust analysis of these settlements, it is best to assume that substantive patent law is optimal. Although scholars sometimes argue that current patent law upholds too many patents,²⁸ or too few,²⁹ some balance must be struck. Even if one believes that current patent law does not strike the correct balance, the correct solution is to reform patent law, not to allow courts in antitrust cases to second-guess patent law doctrine and try to offset it imperfectly for the limited set of cases that produce patent settlements that raise antitrust issues. This second-guessing approach would not work both because it would require litigating the optimality of the patent system in *every* antitrust case that involved patent rights (not just reverse payment settlement cases), and because it would alter the innovation reward in the odd subset of cases that lead to such antitrust suits, which would distort firm incentives in choosing among possible innovations. Therefore, antitrust analysis of patent settlements should assume the optimality of patent law.

Given that Congress and the courts have crafted the substantive doctrines that determine the probability that a patent would be found valid and infringed, the amount of exclusion that the patent holder deserves on the merits is equal to the probability that the patent would be found valid and infringed times the remaining patent term.³⁰ To formalize this, call the

28. See, e.g., Ian Ayres & Gideon Parchomovsky, *Tradable Patent Rights*, 60 STAN. L. REV. 863, 864 (2007) (arguing that issuance of too many patents chills innovation).

29. See, e.g., Joshua L. Sohn, *Can't the PTO Get a Little Respect?*, 26 BERKLEY TECH. L.J. 1603, 1605 (2011) (arguing that federal courts do not give PTO decisions enough deference during invalidity contests).

30. For the same reasons, we think antitrust law should assume the optimality of the Hatch-Waxman Act, which gives pharmaceutical patent holders the additional exclusion right of an automatic 30-month stay on generic entry, which helps incentivize patent holders to incur the costs

probability that the patent will be found valid and infringed θ , and normalize the remaining patent term so that it spans from 0 to 1. For example, if 100 months remained on the patent term, then 100 months would be 1.0 on the normalized scale, 50 months would be 0.5, 10 months would be 0.1, and so forth. According to patent law, the patent holder deserves exclusivity for θ of the remaining patent period because θ percent of the time it deserves exclusivity for the entire period and $1 - \theta$ percent of the time it deserves no exclusivity. This means that if a settlement exclusion period T (again on the normalized 0 to 1 timeline) is greater than θ , then T exceeds the optimal patent exclusion period, and thus gives the patent holder more exclusivity and patent reward than it deserves. For example, if the remaining patent term is 100 months, and the probability of patent victory is 0.5, then the settlement exclusion period exceeds the optimal patent exclusion period only if $T > 0.5$; in other words, if the settlement excludes entry for more than 50 months. This measure entitles the patent holder to all the expected profits it would get if patent litigation were instant and costless, and thus enables patent holders to reap any legitimate settlement benefits that come from avoiding the delay and cost of litigation.

Whether a settlement is anticompetitive or procompetitive therefore depends both on whether: (a) the settlement harms or benefits *ex post* consumer welfare, which turns on whether the settlement exclusion period is greater or less than the expected litigation exclusion period; and (b) the patent holder receives more or less than the optimal patent reward, which turns on whether the settlement exclusion period is greater or less than the optimal patent exclusion period. The net effect could be murky if these tests pushed in opposite directions because that would require us to weigh the *ex post* effect on consumer welfare against the *ex ante* effect on innovation (which also affects consumer welfare). We avoid this difficulty by proving that both tests point in the same direction for settlements with reverse payments that exceed the patent holder's anticipated litigation costs.

The reason that these tests could in theory point in opposite directions is that the optimal patent exclusion period might be less or more than the expected litigation exclusion period. If the patent is strong enough to deter at-risk entry during litigation, then the optimal patent exclusion period would be smaller than the expected litigation exclusion period. The reason is that entry would be deterred during litigation with 100% probability, even though such entry would be legal and outside the scope of the patent with a

of new drug applications that can secure FDA approval. 21 U.S.C. § 355(j)(5)(D)(i)-(ii) (2006). Thus, in a Hatch-Waxman case, we would treat the residual patent period as starting once that 30-month stay expires because monopoly profits before the stay expires are part of the special intellectual property reward for investing in new drug applications, which are valuable even if the patent proves invalid. Although the Hatch-Waxman Act provides this limited exclusion right, it nowhere approves anticompetitive settlements that extend beyond that 30-month exclusion right.

probability of $1 - \theta$. If the patent is too weak to deter at-risk entry during litigation, then the optimal patent exclusion period would exceed the expected litigation exclusion period. The reason is that entry would occur during litigation with 100% probability, even though such entry would be illegal and within the scope of the patent with a probability of θ . In either case, the theoretical concern is that if the settlement exclusion period was between the optimal patent exclusion period and the expected litigation exclusion period, then the two tests would produce conflicting conclusions.

However, we prove that when a settlement has a reverse payment that exceeds the patent holder's anticipated litigation costs, the settlement exclusion period will exceed *both* (1) the expected litigation exclusion period and thus harm *ex post* consumer welfare *and* (2) the optimal patent exclusion period and thus exceed the optimal patent reward for innovation *ex ante*. Such a settlement is thus unambiguously anticompetitive.

II. Reverse Payments That Exceed the Patent Holder's Anticipated Litigation Costs

A. *The Proof*

To begin, we must define some variables. Call θ_P and θ_E the respective estimates by the patent holder and entrant of the probability that the patent will be found valid and infringed. Call P_N the supracompetitive profits that the patent holder would earn with *no* entry by the entrant for the remainder of the patent term. This would equal monopoly profits in the case where the patent holder is a monopolist, but could reflect a lesser degree of market power if other rivals exist that constrain the patent holder from charging the full monopoly price but do not fully constrain the patent holder to price at cost. Call P_Y the more competitive profits that the patent holder would earn over that period if the entrant enters as soon as it can. We normalize the remaining patent term to extend from time 0 (when the entrant is first capable of entering the market) to time 1 (the patent expiration date), with no discount rate and the assumption that each time slice reflects an equal share of the total profits that could be earned during that period.³¹ Call E the profits the entrant would earn if it were in the market for the remainder of the patent term (from time 0 to time 1), and call C_P and C_E the expected cost of litigation for the patent holder and entrant, respectively.

31. Altering the model to include discount rates, make profitability differ over time, or both would not change any of the conclusions in the proof but would significantly complicate the mathematical formulas. In fact, adding either of these complications would only strengthen our proof. Adding discount rates would reduce the net present value of the patent holder's anticipated litigation costs but would not reduce the net present value of any reverse payment made at time 0. Discounting future profit streams would only increase the extent to which an entrant would be willing to delay entry in exchange for an upfront settlement payment and reduce the extent to which a patent holder is willing to speed up entry.

The proof below does not require one to know P_N , P_Y , or E . Instead, the proof holds so long as the patent holder has enough market power that the joint profits of the patent holder and entrant are lower with entry by the potential entrant than without it, which is what standard economic models and common observation predict. For example, standard economic models indicate that monopoly profits exceed duopoly profits, and empirical evidence indicates that each entrant added to a market generally reduces profit margins until one gets to the point where market power is fully constrained.³² Moreover, one need not even assume this because in the Appendix we prove in the alternative that if joint profits were *not lower* with entry, then a reverse payment would be completely unnecessary for settlement. For now, let's focus on the more realistic case where $P_N > P_Y + E$.

Absent settlement, the entrant must decide whether to enter before resolution of the patent litigation or instead wait and enter only if it wins that patent litigation. Entry before resolution of the patent litigation is often called "at-risk" entry because the entrant risks having to pay infringement damages to the patent holder if it loses the litigation. Call L the expected duration of the patent litigation, again on the normalized 0 to 1 timeline.³³ The entrant will enter at risk if its expected profits during at-risk entry, LE , exceed its expected infringement liability, which is equal to θ_E (its expected probability of losing) times the patent holder's lost profits during at-risk entry $L(P_N - P_Y)$.³⁴ This means the entrant will enter at risk only if $\theta_E < E/(P_N - P_Y)$. As shorthand, define $\theta^* = E/(P_N - P_Y)$, and call a patent "strong" if it deters at-risk entry ($\theta_E > \theta^*$) and "weak" if it does not ($\theta_E < \theta^*$).³⁵

We model settlements that do two things: set a settlement exclusion period (of T , on the normalized 0 to 1 time scale) and give the entrant a reverse payment (in amount R). Thus, the entrant's settlement payoff is

32. See, e.g., Timothy F. Bresnahan & Peter C. Reiss, *Entry and Competition in Concentrated Markets*, 99 J. POL. ECON. 977, 984 (1991); Richard G. Frank & David S. Salkever, *Generic Entry and the Pricing of Pharmaceuticals*, 6 J. ECON. & MGMT. STRATEGY 75, 84 (1997); David Reiffen & Michael R. Ward, *Generic Drug Industry Dynamics*, 87 REV. ECON. & STAT. 37, 43 (2005).

33. For example, if the remaining patent term is 100 months, and the parties expect the patent litigation to last 10 months, then $L = 0.1$. We assume both parties share the same expected litigation duration L because it makes the mathematical model easier to understand but does not change any of the relevant conclusions. If we instead assumed that the entrant was relatively pessimistic about litigation length (so that $L_E > L_P$), that would only widen the range of possible settlement entry dates that (without any reverse payment) can provide settlement payoffs to both the entrant and patent holder that exceed their litigation payoffs. If we instead assumed the entrant was relatively optimistic about litigation length (so that $L_E < L_P$), that would only increase the extent to which a settlement exclusion period that exceeds the patent holder's estimate of the expected litigation exclusion period will exceed the entrant's estimate of that expected litigation exclusion period.

34. This formula assumes that the entrant has sufficient assets to pay damages, i.e., that it is not judgment proof. We discuss below the case of a judgment-proof entrant. See *infra* section II(B)(2).

35. Our model assumes firms are risk neutral, but as we show below, our conclusions do not depend on this assumption.

$(1 - T)E + R$ because the entrant earns nothing during T , when the settlement excludes it from the market, earns profits at a rate of E during the remaining patent term (i.e., during $1 - T$), and gets the reverse payment R . Conversely, the patent holder's settlement payoff is $TP_N + (1 - T)P_Y - R$ because it earns supracompetitive profits at a rate of P_N during T , when the settlement excludes the entrant, earns more competitive profits at a rate of P_Y for the remaining patent term $(1 - T)$, and pays R to the entrant.

The parties' joint payoff from settlement is thus $TP_N + (1 - T)P_Y - R + (1 - T)E + R$, which simplifies to $P_Y + E + T(P_N - P_Y - E)$. Because the patent holder's profits without entry exceed joint profits with entry, $P_N - P_Y - E$ is positive. The parties' joint payoff is thus clearly maximized by choosing the maximum T of 1, that is, by setting the settlement exclusion period equal to the entire remaining patent term. At this $T = 1$, the joint settlement payoff is P_N , which in the case where the patent holder was a monopolist would mean monopoly profits throughout the patent period. This is the settlement we can expect if the *formal* scope of the patent test were adopted and the settling parties were free to choose any settlement exclusion period as long as it did not exceed the patent expiration date because such a settlement maximizes their joint profits. Because θ_P and θ_E are both less than 1, a settlement exclusion period of $T = 1$ means that the settlement exclusion period necessarily exceeds both the optimal patent exclusion period and the expected litigation exclusion period.

However, if the formal scope of the patent test were *not* adopted, we might hope that the threat of antitrust liability would cause the parties to choose a settlement exclusion period of $T < 1$. Even then, however, neither party would ever enter into a patent settlement that leaves it worse off than it would be if it litigated. We prove next that the unwillingness of either party to approve a settlement that leaves it worse off suffices to assure that a settlement must be anticompetitive if the reverse payment exceeds the patent holder's anticipated litigation costs.

1. Strong Patent.—With a strong patent, the entrant would not enter at risk during the litigation. Thus, expected entry would be delayed by at least the length of litigation L , and also delayed with probability θ for the remainder of the patent term, $1 - L$. Accordingly, the expected litigation exclusion period is $L + \theta(1 - L)$, which can be rearranged as $\theta + (1 - \theta)L$.

The patent holder's expected litigation payoff is $LP_N + (1 - L)[\theta_P P_N + (1 - \theta_P)P_Y] - C_P$. The first term reflects the fact that the patent holder earns profits at a rate of P_N during the patent litigation, no matter how that litigation turns out. The next two terms reflect the fact that, after the patent litigation ends, it earns profits at a rate of P_N if it wins that litigation and P_Y if it loses. The last term reflects its anticipated litigation costs. Given a reverse payment that exceeds the patent holder's anticipated litigation costs by some additional amount A , the patent holder's settlement payoff is $TP_N + (1 - T)P_Y - A - C_P$. Thus, the patent holder will accept the settlement only if $TP_N +$

$(1 - T)P_Y - A - C_P > LP_N + (1 - L)[\theta_P P_N + (1 - \theta_P)P_Y] - C_P$. Rearranging, this is true only when $T > \theta_P + (1 - \theta_P)L + A/(P_N - P_Y)$.

Therefore, the minimum settlement exclusion period that the patent holder will demand is $\theta_P + (1 - \theta_P)L + A/(P_N - P_Y)$. According to the patent holder's own probability estimate, the optimal patent exclusion period is θ_P and the expected litigation exclusion period is $\theta_P + (1 - \theta_P)L$. Thus, the minimum settlement exclusion period will, by its own estimate, always exceed the optimal patent exclusion period by $(1 - \theta_P)L + A/(P_N - P_Y)$ and the expected litigation exclusion period by $A/(P_N - P_Y)$. These terms are all positive because by definition $A > 0$, $P_N > P_Y$, $L > 0$, and $\theta_P \leq 1$. Moreover, the more the reverse payment exceeds the patent holder's anticipated litigation costs, the more the minimum settlement exclusion period will exceed both benchmarks.

In short, according to the patent holder's own probability estimate, a settlement with a reverse payment that exceeds its anticipated litigation costs will always exclude the entrant for greater than the expected litigation exclusion period and the optimal patent exclusion period, even though the patent is strong enough to deter at-risk entry. This is true no matter what the entrant estimates the patent strength to be.

If the patent holder and the entrant disagree about the patent strength θ , there are two possibilities. One possibility is that $\theta_P > \theta_E$, meaning that the patent holder's estimate of patent strength exceeds the entrant's, so that we can say the entrant is relatively optimistic. If so, then all the above propositions will also be true according to the entrant's probability estimate. Indeed, according to the entrant's lower probability estimate, the settlement exclusion period will even more greatly exceed the expected litigation exclusion period and the optimal patent exclusion period.

The other possibility is that $\theta_E > \theta_P$, meaning that the entrant's estimate of patent strength exceeds the patent holder's, so that we can say the entrant is relatively pessimistic. If so, then the parties will always be able to reach a settlement without any reverse payment. Without any reverse payment, $A = -C_P$, so the above analysis shows that the patent holder will agree to such a settlement if $T > \theta_P + (1 - \theta_P)L - C_P/(P_N - P_Y)$, which can be rearranged as $L + \theta_P(1 - L) - C_P/(P_N - P_Y)$. The entrant will agree as long as its settlement payoff exceeds its expected litigation payoff. Without a reverse payment, the entrant's settlement payoff is $(1 - T)E$. The entrant's expected litigation payoff, given litigation delays and no at-risk entry, is $(1 - L)(1 - \theta_E)E - C_E$ because the entrant earns nothing during the litigation period, earns profits at a rate of E after the litigation period if it wins, and must pay litigation costs of C_E . Thus, the entrant will agree to such a settlement if $(1 - T)E > (1 - L)(1 - \theta_E)E - C_E$. Rearranging, this is true if $T < L + \theta_E(1 - L) + C_E/E$. Thus, a settlement without any reverse payment will be possible for a strong patent as long as the maximum exclusion period T that the entrant would agree to, $L + \theta_E(1 - L) + C_E/E$, is greater than the minimum T that the patent holder would demand, $L + \theta_P(1 - L) - C_P/(P_N - P_Y)$. This can be rearranged as $(\theta_E -$

$\theta_P)(1 - L) + C_E/E + C_P/(P_N - P_Y) > 0$. This inequality is always satisfied because $\theta_E > \theta_P$, given that the entrant is relatively pessimistic, and all the other terms are positive.

Therefore, when the entrant is relatively pessimistic, a reverse payment is never necessary to reach settlement, even if the patent is strong. Further, because adding a reverse payment can only increase the entrant's willingness to agree to a larger settlement exclusion period, the settlement they would have reached without the reverse payment would provide a shorter exclusion period less harmful to consumer welfare, while still exceeding the optimal patent exclusion period according to the patent holder's own estimate.

2. *Weak Patent.*—With a weak patent, the entrant would enter at risk immediately because the entrant thinks that its expected profits during entry exceed its expected infringement liability. The patent holder's expected litigation payoff would thus be $P_N\theta_P + P_Y(1 - \theta_P) - C_P$. The first term reflects the fact that, if the patent holder wins the patent litigation, the patent holder receives supracompetitive profits P_N throughout the patent term because the patent victory means it will recover damages for any reduction in profits below that level during the litigation and it can exclude its rival after the litigation ends.³⁶ The second term reflects the fact that, if the patent holder loses the patent litigation, it will receive the more competitive profits P_Y throughout the patent term because the lost litigation means it gets no damages for the entry during litigation and cannot exclude the entrant after the litigation ends. The last term reflects its anticipated litigation costs. Given a reverse payment that exceeds the patent holder's expected litigation costs by A , its settlement payoff is $P_NT + P_Y(1 - T) - A - C_P$. Thus, the patent holder will agree to a settlement if $P_NT + P_Y(1 - T) - A - C_P > P_N\theta_P + P_Y(1 - \theta_P) - C_P$, which simplifies to $T > \theta_P + A/(P_N - P_Y)$.

Therefore, the minimum settlement exclusion period that the patent holder will demand is $\theta_P + A/(P_N - P_Y)$. Accordingly, the shortest possible settlement exclusion period will exceed the optimal patent exclusion period, θ_P , by $A/(P_N - P_Y)$. Given at-risk entry, the entrant would have entered at time 0, but would remain in the market after the patent litigation ends only if

36. The formula in the above text assumes the entrant has sufficient assets to pay any patent damages. If the entrant does not, then it is judgment proof to some extent, which does provide a possible ground for rebuttal that we discuss below in section II(B)(2). The formula in the text also assumes that damages are not trebled for willful infringement. Because we are talking here about a weak patent, where by definition the odds are relatively low that a court would sustain the patent claims, it is very unlikely willful infringement would ever be found, especially because willful infringement is found in only 2.1% of all patent disputes. Kimberly A. Moore, *Empirical Statistics on Willful Patent Infringement*, 14 FED. CIR. B.J. 227, 234 (2005). In any event, the prospect that damages might be trebled would either: (1) raise damages high enough to deter entry, in which case the strong patent proof would apply; or (2) raise the patent-holder returns from litigation if the patent remained too weak to deter entry, which would make the patent holder demand an even larger settlement exclusion period, worsening all the effects predicted by the model.

it won. Thus, the expected litigation exclusion period is $\theta_P(1 - L)$, or $\theta_P - \theta_P L$. The minimum settlement exclusion period of $\theta_P + A/(P_N - P_Y)$ will thus exceed the expected litigation exclusion period by $\theta_P L + A/(P_N - P_Y)$.

In short, according to the patent holder's own probability estimate, a settlement with a reverse payment that exceeds its anticipated litigation costs will always exclude entry for longer than both the expected litigation exclusion period and the optimal patent exclusion period. Again, the more the reverse payment exceeds the patent holder's anticipated litigation costs, the more the minimum settlement exclusion period will exceed both benchmarks.

If the entrant is relatively optimistic, then we can also say that the patent holder's probability estimate θ_P exceeds the entrant's probability estimate θ_E , and therefore all the above propositions will also be true according to the entrant's probability estimate. Indeed, according to the entrant's lower probability estimate, the settlement exclusion period will even more greatly exceed the expected litigation exclusion period and the optimal patent exclusion period.

If the entrant is relatively pessimistic, then the parties will be able to reach a settlement without any reverse payment. Without any reverse payment, $A = -C_P$, and thus the above analysis shows that the patent holder will agree if $T > \theta_P - C_P/(P_N - P_Y)$. The entrant will agree as long as its settlement payoff exceeds its litigation payoff. Without any reverse payment, the entrant's settlement payoff is $(1 - T)E$. The entrant's litigation payoff with at-risk entry is $L[E - \theta_E(P_N - P_Y)] + (1 - L)(1 - \theta_E)E - C_E$ because during the litigation period it earns profits at a rate of E but must pay infringement damages at a rate of $P_N - P_Y$ if it loses; after the litigation period, it earns profits at a rate of E if it wins and nothing if it loses, and it must pay litigation costs of C_E either way. Therefore, with a weak patent, the entrant will agree to a settlement without any reverse payment as long as $(1 - T)E > L[E - \theta_E(P_N - P_Y)] + (1 - L)(1 - \theta_E)E - C_E$. Rearranging, this is true if $T < \theta_E + (1/E)[\theta_E L(P_N - P_Y - E) + C_E]$. Thus, a settlement without any reverse payment will be possible for any weak patent as long as the maximum exclusion period T that the entrant would agree to, $\theta_E + (1/E)[\theta_E L(P_N - P_Y - E) + C_E]$, is greater than the minimum T that the patent holder would demand, $\theta_P - C_P/(P_N - P_Y)$, which can be rearranged as $\theta_E - \theta_P + (1/E)[\theta_E L(P_N - P_Y - E) + C_E] + C_P/(P_N - P_Y) > 0$. This inequality is always true because $\theta_E > \theta_P$, given that the entrant is relatively pessimistic, and the other terms are all positive, given that the patent holder's profits without entry exceed joint profits with entry.

Therefore, when the entrant is relatively pessimistic, a reverse payment is never necessary to reach settlement. Further, because increasing the reverse payment amount beyond the patent holder's anticipated litigation costs can only increase the entrant's willingness to agree to a longer settlement exclusion period, the settlement they would have reached without a reverse payment above this level would provide a shorter exclusion period

less harmful to consumer welfare, while still exceeding the optimal patent exclusion period according to the patent holder's own estimate.

3. *Implications.*—In sum, if the reverse payment amount exceeds the patent holder's anticipated litigation costs, the following propositions hold true, whether the patent is weak or strong. *First*, the settlement exclusion period must exceed the expected litigation exclusion period and the optimal patent exclusion period according to the patent holder's own estimate of the patent strength. This is true whether the entrant is relatively optimistic or pessimistic. Further, the higher the reverse payment, the worse these effects are.

Second, if the entrant is relatively optimistic, then both benchmarks are exceeded even further according to the entrant's own estimate of the patent strength. Such a settlement thus anticompetitively excludes the entrant for longer than the expected litigation exclusion period and the optimal patent exclusion period according to *both* the patent holder's *and* the entrant's estimates of the patent strength. This means that both the entrant and patent holder knew the settlement was anticompetitive.

Third, if the entrant is relatively pessimistic, the parties could always settle without any reverse payment at all. In this case, a reverse payment that exceeds the patent holder's anticipated litigation costs is not only *unnecessary* to reach settlement, but this also means that the settlement exclusion period *must* exceed the expected litigation exclusion period and the optimal patent exclusion period, according to the patent holder's *own* probability estimate. There is thus no reason to tolerate a reverse payment of this size because without it the alternative settlement the parties could have reached would have provided a shorter exclusion period less harmful to consumer welfare, while still exceeding the optimal patent exclusion period according to the patent holder's own estimate. Because the patent holder's own estimate is the only estimate that can affect *its* incentives to invest in innovation, it is the key estimate to consider in determining whether the settlement exceeds the optimal patent exclusion period.

Defenders of reverse payments often stress that they may sometimes be necessary to reach settlement.³⁷ But the above analysis proves that a reverse payment that exceeds the patent holder's anticipated litigation costs is never necessary to secure a *desirable* settlement. Instead, a reverse payment of this size can be necessary to reach settlement only when both the patent holder and entrant know the settlement is anticompetitive. Therefore, courts can safely condemn settlements with reverse payments of this size because doing so can deter only anticompetitive settlements.

37. See, e.g., Robert D. Willig & John P. Bigelow, *Antitrust Policy Toward Agreements That Settle Patent Litigation*, 49 ANTITRUST BULL. 655, 659–62, 667–77 (2004).

To put it another way, when a reverse payment exceeds the patent holder's anticipated litigation costs, a court can be confident that the settlement exclusion period will exceed the optimal patent reward, while anticompetitively reducing consumer welfare as compared either to litigation or to an alternative settlement without a reverse payment of that size. This conclusion does not rely on any particular level of patent strength θ or any assumption that the parties agreed on that level. Nor does it require knowledge of the parties' varying estimates of patent strength or even knowing which side's estimate is greater. It does not even require us to assume that the parties picked the settlement that maximized profits or to make any particular assumption about the extent to which the parties considered the risk of antitrust liability. It simply requires us to assume that neither party to the patent dispute would agree to a settlement that made it worse off.

The proof above nowhere needed to assume the existence of anything like the Hatch-Waxman Act's 180-day generic exclusivity period, during which only the first-filing generic entrant is permitted to enter.³⁸ That generic exclusivity period is often cited as the main problem with reverse payment settlements because it means that a settlement between the patent holder and first-filing generic that delays the entry of that generic will also create a bottleneck effect that delays the entry of other nonsettling generic entrants.³⁹ When settlements have this bottleneck effect, that certainly exacerbates their anticompetitive consequences. Because our proof does not rely on this bottleneck effect, it not only provides a conservative estimate of the anticompetitive consequences of reverse payment settlements under the Hatch-Waxman Act, but also shows that the problem with reverse payment settlements extends well beyond the Hatch-Waxman Act and the pharmaceutical industry regulated by it.

B. Presumption and Limited Grounds for Rebuttal

The above proof assumes that at-risk entrants are not judgment proof and that the reverse payment does not have some other procompetitive justification. The proof thus suggests that courts should presumptively condemn settlements when the reverse payment exceeds the patent holder's litigation costs, unless the defendants can rebut this presumption by showing either: (1) that the entrant would have entered at risk and is judgment proof to a sufficient effect to change the results or (2) that some other procompetitive justification exists and offsets the anticompetitive effect. Absent either of those rebuttals, the proof shows that a settlement with a

38. 21 U.S.C. § 355(j) (2006).

39. See, e.g., Michael A. Carrier, *Unsettling Drug Patent Settlements: A Framework for Presumptive Illegality*, 108 MICH. L. REV. 37, 71 (2009).

reverse payment of this size always has anticompetitive effects. In particular, absent such rebuttals, a reverse payment of this size precludes the possibilities: (1) that the settlement exclusion period is actually shorter than the expected litigation exclusion period or the optimal patent exclusion period and (2) that the patent holder lacks market power. Defendants thus should not be permitted to rebut the presumption by trying to prove that either of those possibilities is true.

1. Establishing Presumption by Comparing Reverse Payment to Anticipated Litigation Costs.—To apply the presumption indicated by our proof, a court need only determine whether the reverse payment amount exceeds the patent holder's anticipated litigation costs. The amount of the reverse payment is easy to ascertain if the settlement specifies just a monetary payment to the entrant. Sometimes, however, a payment to the entrant consists of consideration other than money, such as a business license, in which case the reverse payment amount equals the expected value (at the time of settlement) of that consideration. Other times there is also some return consideration, in which case the reverse payment amount is the difference between the expected value of the consideration flowing to and from the entrant, leaving aside the value of setting the entry date and avoiding litigation costs.

One must also estimate anticipated litigation costs. For the purpose of applying this proof, only the *forward-looking* anticipated litigation costs are relevant; past litigation expenses are sunk costs and thus should not affect the patent holder's willingness to settle. This means that the patent holder's anticipated litigation costs will be relatively small in cases where the parties settle after discovery or trial, which are by far the most expensive aspects of litigation.⁴⁰ However, when parties settle before discovery or trial, those potentially expensive litigation costs have not yet occurred. Thankfully, there are three easily administrable ways that a court can determine whether the reverse payment amount exceeded the patent holder's anticipated litigation costs. We present them in order from easiest to hardest.

First, the reverse payment amount may sometimes exceed the patent holder's own estimate of litigation costs in its documents. This is a sufficient

40. See AIPLA, REPORT OF THE ECONOMIC SURVEY 2011, at 35 (2011) [hereinafter 2011 AIPLA REPORT] (stating that, as of 2011, median litigation costs for a patent infringement suit with more than \$25 million at stake were \$3 million through the end of discovery and \$5 million in total); Meredith Addy, *Appellate Strategy Before the U.S. Court of Appeals for the Federal Circuit*, in PATENT LITIGATION, NEGOTIATION, AND SETTLEMENT: LEADING LAWYERS ON STRATEGIES FOR EFFECTIVELY RESOLVING PATENT DISPUTES 7, 8 (2006), available at <http://www.brinkshofer.com/files/201.pdf> ("Generally, once a patent case has gone through a district court trial, it has already cost, on average, \$3 to \$5 million, or more. Comparatively, the cost of appeal is far less: perhaps a few hundred thousand dollars for an easy case, a few million for a complicated one, but almost always exponentially less than the initial litigation.").

but not necessary condition for finding that the reverse payment exceeded its anticipated litigation costs because, if this presumption were adopted, patent holders would predictably start to inflate their recorded estimates of litigation costs in order to evade antitrust liability. Thus, courts should move onto the next method if this first method does not indicate that the reverse payment exceeded the patent holder's anticipated litigation costs.

Second, a court could compare the reverse payment amount to the upper bound of litigation costs from similar cases. The largest publicly documented amount spent on patent litigation that we could find was \$32 million, spent by Apple in a suit against Google's Motorola Mobility unit.⁴¹ Empirical literature confirms that \$32 million is an upper bound. Surveys of intellectual property lawyers indicate that the 75th percentile for patent litigation costs through trial, for cases with more than \$25 million in controversy, was around \$7.5 million in 2011.⁴² This 75th percentile is \$10 million for cases in New York,⁴³ but even that figure is only one-third of the \$32 million upper bound. A court could, therefore, be confident that any reverse payment settlement in excess of the \$32 million upper bound exceeds the patent holder's anticipated litigation costs.

The reverse payments in past cases have often far exceeded \$32 million, including \$66.4 million in *Tamoxifen*,⁴⁴ \$151.5 million in *Valley Drug*,⁴⁵ \$264–\$382.5 million in *Watson*,⁴⁶ and \$398.1 million in *Arkansas Carpenters*.⁴⁷ These past cases show that in many situations, including those arising in the past key appellate cases, applying this test should not require significant fact-finding. Further, in most cases the relevant upper bound will

41. See Susan Decker, *Apple Patent Battles Create Lawyer Boon at \$1,200 an Hour*, BLOOMBERG (Aug. 23, 2012, 3:00 PM), <http://www.bloomberg.com/news/2012-08-23/apple-patent-battles-create-lawyer-boon-at-1-200-an-hour.html>.

42. 2011 AIPLA REPORT, *supra* note 40, at 35–36, I-154.

43. *Id.* at I-154. The reports from previous years have similar figures, with the highest 75th percentile reported in any year being \$11.5 million for cases in the Los Angeles region in 2009. See AIPLA, REPORT OF THE ECONOMIC SURVEY 2009, at I-129 (2009) [hereinafter 2009 AIPLA REPORT].

44. *In re Tamoxifen Citrate Antitrust Litig.*, 466 F.3d 187, 190, 194 (2d Cir. 2006) (noting the \$66.4 million total given to two generics).

45. The patent holder agreed to pay one generic \$6 million every three months and the other generic \$4.5 million per month for the period from March 31, 1998 to March 1, 2000. *Valley Drug Co. v. Geneva Pharm., Inc.*, 344 F.3d 1294, 1300–01 (11th Cir. 2003). Because of an FTC investigation, the parties terminated this settlement agreement early on August 13, 1999, so \$39 million of this was not actually paid out, but that was not part of the original settlement agreement and thus not relevant to the inference at issue here. *Id.*

46. The patent holder agreed to pay \$60 million to one generic and \$19–\$30 million annually to another generic for the 10.75 years from January 2006 to September 2015. *FTC v. Watson Pharm., Inc.*, 677 F.3d 1298, 1304–05 (11th Cir. 2012). The patent holder also agreed to pay one generic \$12 million for backup manufacturing assistance, but if we assume that this \$12 million constituted fair consideration for that assistance, then it should be excluded from the calculation of the reverse payment amount. *Id.*

47. *Ark. Carpenters Health & Welfare v. Bayer AG*, 604 F.3d 98, 102 n.8 (2d Cir. 2010).

be much smaller than \$32 million. Reverse payment settlements have most often arisen in the pharmaceutical industry, where patent litigation usually involves one or a few patents and the largest publicly documented amount spent on litigation that we could find was \$15 million, for a drug that had over \$1 billion in annual sales.⁴⁸ In contrast, the \$32 million litigation cost mentioned above involved many patents relevant to Apple's iPhone, on which it earned annual revenue of \$47 billion, which likely explains Apple's willingness to spend so much on that patent litigation.⁴⁹ Thus, \$15 million is likely a safe upper bound for pharmaceutical cases.

If neither of the above tests is dispositive, then the parties could call patent lawyers as experts to estimate the patent holder's anticipated litigation costs. For several reasons, this method of objectively measuring the patent holder's anticipated litigation costs is significantly more administrable than trying to objectively measure the patent strength, as courts would have to endeavor to do under an objective probabilistic scope of the patent test. First, because law firms that try large patent cases almost exclusively bill by the hour, one need only know the average amount of time that was expected to be necessary for patent litigation in order to estimate the patent holder's anticipated litigation costs; one need not estimate the probability that the patent holder would win or lose, which is conceptually far more difficult, if not imponderable. Second, hindsight bias is not a concern because most firms' hourly billing structures mean that patent litigation costs do not depend on whether the patent holder wins or loses. Third, firms that honestly are trying to keep their settlement payments below the patent holder's anticipated litigation costs could easily insulate themselves from being second-guessed by a court by soliciting arm's-length estimates of their litigation costs from law firms prior to the settlement.

2. Rebuttal by Showing At-Risk Entrant Is Sufficiently Judgment Proof.—Whether an entrant is judgment proof can affect whether or not the patent holder expects it to enter at risk, although the effects are mixed. On the one hand, if the entrant is judgment proof, then it pays only a fraction of damages if it loses, which makes it more likely to enter at risk. On the other hand, being judgment proof also means that if it loses, the entrant will go bankrupt and the managers may lose their jobs, which will make the corporation more likely to behave in a risk-averse fashion due to managerial risk aversion.

If the patent holder concludes that the net effect is that the entrant will not enter at risk, then our proof for strong patents continues to apply without any modification. The reason is that if the patent holder does not expect at-

48. See Richard D. Margiano, *Cost and Duration of Patent Litigation*, MANAGING INTEL. PROP., Feb. 2009, at 150, 150.

49. See Decker, *supra* note 41.

risk entry, then whether the entrant is judgment proof is irrelevant to assessing the patent holder's litigation payoff because it does not expect to sue for damages anyway. However, if the patent holder concludes that the net effect is that the entrant will enter at risk, then it becomes relevant that our model for weak patents assumes that the entrant is not judgment proof, that is, that it has sufficient assets to fully pay any patent damages. If we instead assume that an at-risk entrant would be judgment proof, then a patent holder could suffer an uncompensated loss of patent profits from such at-risk entry. This would reduce the expected litigation payoff to the patent holder, and thus would make it willing to accept a settlement with a smaller exclusion period than our proof predicted with at-risk entry.

This effect could mean that, even with a reverse payment that exceeds litigation costs, the patent holder might accept a settlement exclusion period that is less than the optimal patent exclusion period. To see why, call J the share of damages (between 0 and 1) that a judgment-proof entrant will be unable to pay. To simplify, assume here that both the patent holder and entrant perceive the same patent strength of θ . Then, assuming at-risk entry, we must subtract $JL\theta(P_N - P_Y)$ from the previously predicted litigation payoff because the patent holder no longer expects to collect that share of its lost-profits damages during litigation if it wins. Thus, its litigation payoff is now $P_N\theta + P_Y(1 - \theta) - C_P - JL\theta(P_N - P_Y)$. If this litigation payoff is exceeded by its settlement payoff, $P_N T + P_Y(1 - T) - A - C_P$, the patent holder will agree to settlement. This simplifies to saying the patent holder will agree to a settlement exclusion period of $T > \theta + A/(P_N - P_Y) - JL\theta$. The minimum settlement exclusion period of T that the patent holder would accept could thus be less than the optimal patent exclusion period if $JL\theta > A/(P_N - P_Y)$, which we can rearrange as when $JL\theta(P_N - P_Y) > A$. That is, the minimum settlement exclusion period might be less than the optimal patent exclusion period if the expected amount of uncollectable lost profits exceeds the difference between the reverse payment and the patent holder's anticipated litigation costs.

However, this is just the minimum settlement exclusion period. The actual settlement exclusion period could also be larger. The at-risk entrant's litigation payoff would be $L[E - \theta(P_N - P_Y)] + (1 - L)(1 - \theta)E - C_E$ plus the $JL\theta(P_N - P_Y)$ in damages it expects to avoid because it is judgment proof. If this is exceeded by its settlement payoff, $(1 - T)E + R$, it will accept settlement. This can be rearranged to conclude that the at-risk entrant will accept settlement if $T < \theta + (1/E)[R + C_E + L\theta(P_N - P_Y - E) - JL\theta(P_N - P_Y)]$. The maximum settlement exclusion period that the at-risk entrant would accept can thus exceed the optimal patent exclusion period when the expected amount of uncollectable lost profits is less than the sum of the reverse payment, avoided entrant litigation costs, and joint profits from excluding at-risk entry for a valid patent. Because the settlement could be reached anywhere between the minimum T the patent holder would accept and the maximum T the entrant would accept, one cannot know in such a

case whether the settlement exclusion period will or will not exceed the optimal patent exclusion period. On the other hand, where the defendants can show that the expected amount of uncollectable lost profits does exceed the sum of the reverse payment, avoided entrant litigation costs, and joint profits from excluding at-risk entry for a valid patent, then we do know that the settlement exclusion period must have been less than the optimal patent exclusion period.

However, even in such a case, the settlement exclusion period will still exceed the expected litigation exclusion period if the reverse payment exceeds litigation costs. As shown above, the expected litigation exclusion period without litigation is $\theta(1 - L)$. The settlement exclusion period will always exceed this if $\theta + A/(P_N - P_Y) - JL\theta > \theta(1 - L)$, which rearranges to $(1 - J)L\theta + A/(P_N - P_Y) > 0$. This is always true because $J \leq 1$, A is positive, and $P_N > P_Y + E$. The extent to which the minimum settlement exclusion period will exceed the expected litigation exclusion period will thus increase the higher the reverse payment and the higher the length of litigation, odds of patent victory, or share of damages the judgment-proof entrant will pay.

In sum, even if the defendants can show that at-risk entry would have occurred by a judgment-proof entrant, that showing will mean only that the settlement exclusion period *could* be less than the optimal patent exclusion period *only* if they also show that the expected amount of uncollectable lost profits exceeds the difference between the reverse payment and the patent holder's anticipated litigation costs. It will mean the settlement exclusion period necessarily *will* be less than the optimal patent exclusion period *only* if the defendants can also show that the expected amount of uncollectable lost profits exceeds the sum of the reverse payment, avoided entrant litigation costs, and joint profits from excluding at-risk entry for a valid patent. Further, even in such a case, the settlement will clearly increase the exclusion period relative to litigation, thus creating at most a murky tradeoff between harming *ex post* consumer welfare through that increased exclusion of entry and benefiting *ex ante* consumer welfare by increasing the patent reward to a level still within the optimal patent exclusion period.

3. *Rebuttal by Proving Other Procompetitive Justifications.*—Leaving aside cases of judgment-proof entrants, the proof above shows that when a settlement does nothing else other than set an entry date and provide reverse payments that exceed the patent holder's anticipated litigation costs, then the settlement cannot be justified as necessary to reach a settlement that: (a) shortens the expected exclusion period (which would increase *ex post* consumer welfare); or (b) increases the patent reward to a level still within the optimal patent exclusion period (which would increase *ex ante* consumer welfare). The reason is that our proof precludes those procompetitive justifications.

However, in some cases, settlements might have unique features that create *other* procompetitive justifications that can offset any anticompetitive

effects.⁵⁰ For example, one of us was a defense expert in the *In re Cardizem* case and found that the settlement there had the unique feature that it allowed the generic to bring a reformulation of its generic drug onto the market more quickly than otherwise possible. The entry in that case was governed by the Hatch-Waxman Act, which allows a patent holder to automatically delay entry by a reformulated generic by an additional 30 months.⁵¹ The settlement prevented this additional delay by providing that the reformulated generic would be treated like the original generic. The reverse payment was then used to fund the reformulation, which the patent holder ultimately conceded was outside the patent. It thus resulted in earlier generic entry. Further, in that case, the evidence indicated that the generic was judgment proof to a significant extent. It was thus a particularly strong case to rebut a presumption that reverse payment settlements have anticompetitive effects.⁵²

4. *No Rebuttal by Showing Lack of Market Power.*—The model above assumed only that the joint profits of the patent holder and entrant would be higher without entry than with it. This condition will hold as long as the patent holder has any degree of market power that the entrant would constrain, even if it falls well short of monopoly power.

Further, the fact that the reverse payment exceeds litigation costs itself proves that the patent holder has market power that the settling entrant would constrain. If the patent holder lacked this market power, then by definition its business profits would be identical no matter when the entrant entered the market because other firms would constrain the patent holder to price at cost regardless of when this entrant entered. If so, the patent holder would earn the same business profits whether it won the patent litigation, lost the patent litigation, or settled and excluded the entrant for some period. The only effect of settlement would thus be that the patent holder would save its anticipated litigation costs and incur the cost of making the reverse payment.

50. The Third Circuit has recognized the need to allow this sort of rebuttal for other procompetitive justifications. *In re K-Dur Antitrust Litig.*, 686 F.3d 197, 218 (3d Cir. 2012).

51. 21 U.S.C. § 355(j)(5)(D)(i)–(ii) (2006).

52. Aiding this sort of rebuttal evidence was other evidence in that case that minimized the possible anticompetitive effects that needed to be rebutted. The *Cardizem* settlement differed from the sort we model in this Article because it did not end the patent litigation and set a fixed settlement entry date. Rather, it was an interim settlement that required the parties to continue the patent litigation, precluded entry only during the litigation and only if the litigation did not last too long, and allowed the generic to keep the reverse payment only if it won the litigation. Further, in that case anticompetitive effects were undermined by strong evidence that: (1) the entrant would not have entered at risk anyway, so that such a purely interim settlement did not preclude any entry by the settling generic; and (2) no other generic entry was delayed because (a) under the rules that then prevailed, the settling generic had to win the patent litigation to preclude other generics and (b) no other generic received FDA approval in time to enter any earlier anyway. Given this evidence, the FTC concluded that the settlement had not actually delayed any generic entry. FED. TRADE COMM’N, DOCKET NO. 9293, ANALYSIS TO AID PUBLIC COMMENT 4 (2001), available at <http://www.ftc.gov/os/2001/04/hoechstanalysis.pdf>.

Therefore, if the reverse payment amount exceeds the litigation costs, the settlement would always make the patent holder worse off if it lacked market power. Accordingly, the patent holder's willingness to make a reverse payment that exceeds its anticipated litigation costs necessarily means that it believes it has market power.

Given this, if the reverse payment exceeds litigation costs, courts should not allow defendants to rebut the presumption by arguing that the patent holder lacked market power. Instead, a reverse payment of that size itself proves market power, and obviates any need to establish market definition or other methods of showing market power.⁵³

This same analysis also rebuts the claim that anticompetitive effects could be eliminated because nonsettling entrants can still challenge the patent.⁵⁴ Even though that possibility generally exists, our analysis proves that the patent holder would never make a reverse payment of this size if nonsettling entrants could—through entry or patent litigation—create the same constraint on its market power. The patent holder would make a reverse payment that exceeds its anticipated litigation costs only if excluding the settling entrant confers an enhanced market power on the patent holder that it otherwise would not enjoy.

5. *No Rebuttal by Showing Risk Aversion.*—Our model assumes firms are risk neutral. This assumption generally holds, but even if it did not, it would not alter our conclusions.

Entrants are typically public corporations with a market capitalization that exceeds the potential patent damages from the case at hand. In those circumstances, the entrant's managers and shareholders have incentives to behave in a risk-neutral manner that maximizes expected profits. In other situations, entrants might be risk averse, in which case they might not enter at

53. See *FTC v. Ind. Fed'n of Dentists*, 476 U.S. 447, 460–61 (1986) (“[P]roof of actual detrimental effects, such as a reduction of output, can obviate the need for an inquiry into market power, which is but a ‘surrogate for detrimental effects.’”); *Toys “R” Us, Inc. v. FTC*, 221 F.3d 928, 937 (7th Cir. 2000) (Wood, J.) (“[T]he share a firm has in a properly defined relevant market is only a way of estimating market power, which is the ultimate consideration. The Supreme Court has made it clear that there are two ways of proving market power. One is through direct evidence of anticompetitive effects.” (citations omitted)); *Allen-Myland, Inc. v. Int’l Bus. Machs. Corp.*, 33 F.3d 194, 209 (3d Cir. 1994) (Easterbrook, J.) (“Market share is just a way of estimating market power, which is the ultimate consideration. When there are better ways to estimate market power, the court should use them.” (quoting *Ball Mem’l Hosp., Inc. v. Mut. Hosp. Ins., Inc.*, 784 F.2d 1325, 1336 (7th Cir. 1986))); *United States v. Baker Hughes Inc.*, 908 F.2d 981, 992 (D.C. Cir. 1990) (Thomas, J., joined by Ruth Bader Ginsburg, J.) (same); see also IIB PHILLIP E. AREEDA, HERBERT HOVENKAMP & JOHN L. SOLOW, *ANTITRUST LAW: AN ANALYSIS OF ANTITRUST PRINCIPLES AND THEIR APPLICATION* 108 (3d ed. 2007) (“[D]irect indicators of market power . . . can be independent of market definition and are sometimes superior to it. . . . [M]arket definition may not be necessary to prove market power.”).

54. See *FTC v. Watson Pharm., Inc.*, 677 F.3d 1298, 1315 (11th Cir. 2012) (making such a claim).

risk even though a risk-neutral entrant would. But this would merely expand the set of cases for which the proof for strong patents applies, which proved anticompetitive effects. Thus, whether or not risk aversion would deter at-risk entry, a reverse payment that exceeds anticipated litigation costs would be anticompetitive.

Risk aversion is unlikely to be a serious issue for the patent holder. Because the patent holder does not face the risk of patent damages, aversion to loss is not relevant to it. Although individuals might sometimes prefer to avoid variation in profits by accepting certain profits with lower expected value, this is unlikely to be relevant for a publicly held corporation, which generally has incentives to maximize expected profits on behalf of a diversified set of shareholders. Managers who do not maximize expected profits increase the risk that their conduct will be punished by product markets, capital markets, labor markets, takeover threats, shareholder voting, and lower valuation of their stock options. Further, because the issue for patent holders is merely variation in the degree of profits, decisions to litigate are unlikely to create a risk that the corporation will go out of existence that could override those ordinary managerial incentives.

In any event, to the extent that risk aversion could cause managers of the patent holder to enter into settlements that fail to maximize its expected corporate profits, that effect reflects an undesirable agency cost that can only be exacerbated by reverse payments that make such settlements more likely. Facilitating managerial risk aversion that reduces expected firm profits is certainly not a procompetitive efficiency that could justify the anticompetitive effects of a reverse payment settlement. Indeed, considering possible managerial risk aversion would only strengthen the case for invalidating reverse payments that exceed the patent holder's litigation costs. The reason is that, to the extent that managerial risk aversion might make a difference, it means that if managers were allowed to make reverse payments that exceed the patent holder's anticipated litigation costs, the managers might do so in order to get a more certain exclusion period even though the settlement lowers expected corporate profits. Allowing managers to make reverse payments of this size would accordingly lower the expected corporate profits on the underlying innovation, reducing incentives to innovate below optimal levels. Curbing this possible distortion of innovation incentives from managerial risk aversion would thus provide another benefit from condemning reverse payments that exceed the patent holder's anticipated litigation costs.

III. Settlements Without Reverse Payments That Exceed Litigation Costs

If the reverse payment does not exceed the patent holder's anticipated litigation costs, then we can no longer be sure that the settlement exclusion period will *necessarily* exceed the expected litigation exclusion period and the optimal patent exclusion period. But we prove below that such a

settlement *usually* will have these anticompetitive effects. We do so by modeling the simple case of a settlement that sets an exclusion period but has no reverse payment. The FTC, DOJ, and many prominent antitrust and patent scholars have assumed that such a settlement will likely produce a settlement exclusion period that equals the expected litigation exclusion period.⁵⁵ We prove that this widespread assumption is incorrect; instead entry-excluding settlements with no reverse payment are usually anticompetitive. This necessarily means that entry-excluding settlements are also usually anticompetitive if they have a positive reverse payment that is lower than anticipated patent holder litigation costs because increasing the reverse payment from zero to any positive amount can only increase the settlement exclusion period that the patent holder would demand and that the entrant would accept.

Because this subset of settlements is not susceptible to proofs showing that they are necessarily anticompetitive, it may make sense to allow rebuttal through direct inquiry into the expected litigation exclusion period or the optimal patent exclusion period. Although that inquiry is difficult, it can be bounded in various ways that we describe below. To the extent those bounds do not apply and a court concludes that such a direct inquiry is too unreliable, the best substantive solution would be to preclude rebuttal because most such settlements are anticompetitive. However, the better method for resolving these cases might be procedural rather than substantive. The underlying problem that makes it possible for patent holders and entrants to collude in settlements that benefit themselves at the cost of buyers is the ordinary legal rule that patent law does not give buyers standing to challenge dubious patents. Thus, a possible procedural solution to address such settlements would be to give buyers standing to challenge the patent's validity.

A. Proof That Even Settlements with Zero Reverse Payment Are Usually Anticompetitive

Because we are just trying to get a rough sense of likelihood, rather than prove necessary effects, we adopt the simplifying assumption that the entrant and patent holder perceive the same patent strength of θ and same anticipated litigation cost C . Although their perceptions could vary, that possibility could increase or decrease the likelihood of anticompetitive settlements and thus it has no clear effect on the overall likelihood. Otherwise, we use the same model as in Part II. Although we illustrate this proof formally below, the intuition is easy to grasp. Any delay in entry increases the patent holder's profits by more than it decreases the entrant's profits (i.e., $P_N - P_Y > E$). Therefore, the patent holder will be less willing to accept a shorter exclusion period in order to avoid litigation costs than the entrant is willing to accept a

55. See *supra* note 22; see *infra* Part IV.

longer exclusion period to avoid litigation costs. This will push the range of possible settlement exclusions higher.

In a Hatch-Waxman case, there is the additional factor that settlement guarantees the first-filing generic a 180-day period of generic exclusivity after the settlement exclusion period ends, whereas with litigation the odds are θ that the first-filing generic will lose the patent litigation and never get any generic exclusivity period. This factor will make the first-filing generic even more willing to accept a longer exclusion period to get a settlement, especially because a generic's profits during any period of generic exclusivity far exceed its profits during any period when it has to compete with other generics. In other words, this factor means that in a Hatch-Waxman case, T_{\max} would be much greater than calculated in the proof that follows, and thus the odds would be even higher that a settlement without any reverse payment will exclude entry for longer than the expected litigation exclusion period and the optimal patent exclusion period, and the magnitude of the additional exclusion can be much greater. Because the proof that follows does not rely on this additional factor, it conservatively understates the likelihood and magnitude of anticompetitive effects in a Hatch-Waxman case and also shows that the problem exists even outside the Hatch-Waxman context.

1. *Strong Patent.*—Take first the case of a strong patent. As shown above, the patent holder's expected litigation payoff is $LP_N + (1 - L)[\theta P_N + (1 - \theta)P_Y] - C$. Without a reverse payment, its settlement payoff is $TP_N + (1 - T)P_Y$. It will accept a settlement only if the latter is greater than the former, which one can rearrange to show that the minimum settlement exclusion period it will accept is $T_{\min} = \theta + L(1 - \theta) - C/(P_N - P_Y)$. Therefore, T_{\min} will exceed the optimal patent exclusion period whenever $L(1 - \theta) > C/(P_N - P_Y)$, which can be rearranged as $(1 - \theta)L(P_N - P_Y) > C$. In words, the minimum settlement exclusion period will exceed the optimal patent exclusion period whenever the patent holder's anticipated litigation costs are less than the additional supracompetitive profits it expects to make because the strong patent deters entry during litigation even when in fact the patent holder would have lost. Thus, even without any reverse payment, there are some cases where the minimum settlement exclusion period will necessarily exceed the optimal patent exclusion period but other cases when it will not.

With a strong patent, the expected litigation exclusion period is $L + \theta(1 - L)$, which is the same as $\theta + L(1 - \theta)$. Therefore, T_{\min} is always lower than the expected litigation exclusion period by $C/(P_N - P_Y)$.

However, T_{\min} just tells us the bottom edge of the bargaining range. To get the full bargaining range, one needs to also know the maximum settlement exclusion period that the entrant would accept. The entrant's expected litigation payoff is $(1 - L)(1 - \theta)E - C$. Without a reverse payment, its settlement payoff is $(1 - T)E$. It will accept a settlement only if the latter is greater, which one can rearrange to show that the maximum settlement

exclusion period it will accept is $T_{\max} = \theta + L(1 - \theta) + C/E$. This maximum thus always exceeds the optimal patent exclusion period by $L(1 - \theta) + C/E$. It also always exceeds the expected litigation exclusion period by C/E . Because bargaining can produce a settlement anywhere between T_{\min} and T_{\max} , the above analysis proves that, for strong patents, settlements without any reverse payment can produce a settlement exclusion period that exceeds both the optimal patent exclusion period and the expected litigation exclusion period.

Given that any settlement between T_{\min} and T_{\max} is possible, it makes some sense to assume that all such settlements are equally likely. Under this assumption, the middle of this settlement range equals the average expected settlement exclusion period, T_{avg} . Given the above, $T_{\text{avg}} = \theta + L(1 - \theta) + C/(2E) - C/(2(P_N - P_Y))$. T_{avg} will thus exceed the expected litigation exclusion period, $\theta + L(1 - \theta)$, whenever $C/(2E) > C/(2(P_N - P_Y))$, which is true if $P_N - P_Y - E > 0$, which is always true because the patent holder's profits without entry exceed joint profits with entry. T_{avg} will exceed the optimal patent exclusion period by this amount plus $L(1 - \theta)$.

Therefore, even with zero reverse payment and a strong patent, the middle of the settlement range always exceeds both the expected litigation exclusion period and the optimal patent exclusion period. This rebuts the prevailing view that settlements without reverse payments will do neither. To the contrary, the above proof establishes that, if we assume all settlements in the bargaining range are equally likely, settlements without reverse payments are usually anticompetitive.

To get some sense of just how likely these anticompetitive effects are in an average case, we need to estimate the average for some of these parameters. Given the data summarized above, \$10 million appears to be a good high-end estimate of average litigation costs, so we will use that as our average estimate of C . The lion's share of reverse payment settlements have occurred in pharmaceutical markets, where on average, the residual patent term is 90.2 months and monthly pre-entry sales by the patent holder are \$72.6 million.⁵⁶ This means average total sales for a patent holder during the

56. This data is drawn from Professor Scott Hemphill's invaluable survey of 143 patent settlements from 1984 to 2008. See generally C. Scott Hemphill, *An Aggregate Approach to Antitrust: Using New Data and Rulemaking to Preserve Drug Competition*, 109 COLUM. L. REV. 629 (2009); C. Scott Hemphill, *Drug Patent Settlements Between Rivals: A Survey* (March 12, 2007) (unpublished manuscript), available at <http://academiccommons.columbia.edu/catalog/ac%3A129331>. We use the date of the settlement agreement, rather than the expiration of the 30-month Hatch-Waxman stay, as the best indicator of when generic entry was first possible because sometimes the Hatch-Waxman stay gets extended for other reasons, like pediatric exclusivity. If one instead uses the expiration of the 30-month stay, the residual patent period would be 93.2 months. Because our focus is on the prospective issue of how likely it is that settlements that exclude entry would be anticompetitive if no reverse payment were allowed, we combine results from settlements that did and did not have a reverse payment. If one instead wanted to ask about the likelihood that past settlements without any reverse payment were anticompetitive, then the

remaining patent term would be \$6.548 billion. The average length of patent litigation after the end of the automatic Hatch-Waxman stay is about 18 months.⁵⁷ We can thus estimate that on average $L = 18/90.2 = 0.1996$.

On average, a single generic entrant prices at 70%–88% of the pre-entry price charged by the patent holder.⁵⁸ To get a single average, we average these figures to estimate the generic price is 79% of the patent holder's pre-entry price. (With multiple generic entrants, the drop in price is much higher, so the results with the assumption here of only one generic entrant are quite conservative.) Empirical studies show that incumbent drug prices remain fairly constant in response to entry.⁵⁹ Costs are around 20% of the patent holder's pre-entry price,⁶⁰ which suggests that on average $P_N = 80\%$ of \$6.548 billion = \$5.238 billion. Generic producers get 40%–50% of the market.⁶¹ To get a single average, we average these to estimate a 45% generic market share. Because the empirical evidence indicates that generic entry does not alter total market volume,⁶² this means that on average $P_Y = \$2.881$ billion and $E = \$1.732$ billion.⁶³

Given these numbers, the average threshold probability to have a strong patent $\theta^* = E/(P_N - P_Y) = 47.2\%$. Therefore, in the average pharmaceutical case, the patent holder's odds of winning the patent would have to exceed 47.2% to be a strong patent that deters at-risk entry.

For such a strong patent, if we plug the above values into the equation for T_{\min} , we get that $T_{\min} = 0.1954 + 0.8004\theta$. Thus, in the average case, the

residual patent period for only those settlements would be 75.4 months and the average monthly sales figure would be \$42.4 million. This would not alter our qualitative conclusions. See *infra* section III(A)(3) (showing that cutting the residual patent period and annual profit level in half would actually make it somewhat *more* likely that settlements without reverse payments are anticompetitive).

57. See GREENE & STEADMAN, *supra* note 17, app. C.

58. Frank & Salkever, *supra* note 32, at 84 fig.3 (reporting 70%); Reiffen & Ward, *supra* note 32, at 43–44 (reporting 88%).

59. In fact, incumbents increase their drug prices slightly in response to generic entry, but because the price increase is only 0.7% with one generic entrant, we treat it as unchanged. Frank & Salkever, *supra* note 32, at 87. Apparently, the incumbent makes more money by keeping its price high and selling only to price-insensitive customers than the incumbent would make if it lowered its price to compete with the generic for price-sensitive customers.

60. Reiffen & Ward, *supra* note 32, at 43.

61. Frank & Salkever, *supra* note 32, at 89.

62. See Gautier Duflos & Frank R. Lichtenberg, *Does Competition Stimulate Drug Utilization? The Impact of Changes in Market Structure on US Drug Prices, Marketing and Utilization*, 32 INT'L REV. L. & ECON. 95, 106–07 (2012) (concluding that net volume is unchanged by entry into drug markets because entry leads to a decline in both prices and marketing expenditures, which “approximately offset” each other's effects on output).

63. Because the patent holder profit per sale is unchanged, $P_Y = 55\%$ of P_N . The generic who is a single entrant has a price that is 79% of the patent holder's with the same marginal cost of 20%, and thus earns 59% of the patent holder gross sales for 45% of volume, which means average monthly profits of 59% of 45% of \$72.46 million = \$19.2 million. Thus, if it could obtain those profits for the entire residual patent period, it would get \$1.732 billion.

minimum settlement exclusion period exceeds the optimal patent exclusion period whenever $0.1954 + 0.8004\theta > \theta$, which is true whenever $\theta < 97.9\%$. Therefore, even without any reverse payment, the minimum settlement exclusion period exceeds the optimal patent exclusion period in the average case unless the patent holder is all but assured of winning the patent litigation.

We can also ascertain the portion of the settlement range that exceeds the optimal patent exclusion period in the average case. Given the above, this portion is 100% for patent strengths between 47.2% and 97.9%. For extremely strong patents with strengths from 97.9% to 100%, the portion of the settlement range above $\theta = (T_{\max} - \theta)/(T_{\max} - T_{\min})$. Plugging the values into the equation for T_{\max} , we get that $T_{\max} = 0.2054 + 0.8004\theta$. Inserting that into the prior equation, we find that (given average numbers) the portion of the settlement range that exceeds the optimal patent exclusion period is $20.54 - 19.96\theta$. Over this range of extremely strong patents, this portion drops from 99% to 58% as the patent strength goes from 97.9% to 100%.

For a strong patent, the expected litigation exclusion period is $L + \theta(1 - L) = 0.1996 + 0.8004\theta$. The portion of the settlement range above the expected litigation exclusion period in the average case is $(T_{\max} - 0.1996 - 0.8004\theta)/(T_{\max} - T_{\min}) = 58\%$ for all patent strength levels that constitute a strong patent.

In sum, even with zero reverse payment and a strong patent, the middle of the settlement range produces an exclusion period that always exceeds both the expected litigation exclusion period and the optimal patent exclusion period. Assuming parties are equally likely to reach any settlement in the bargaining range, this means that settlements without reverse payments are usually anticompetitive. Further, using actual average numbers for such settlements, the settlement exclusion period for a strong patent is 100% likely to exceed the optimal patent exclusion period unless the patent is extremely strong (more than 97.9% certain to win). Even for extremely strong patents whose patent strength ranges from 97.9% to 100%, the settlement is still 58% to 99% likely to exceed the optimal patent exclusion period. Further, the settlement exclusion period is 58% likely to exceed the expected litigation exclusion period for all levels of patent strength that qualify as a strong patent.

2. *Weak Patent.*—Next consider a weak patent. As shown above, the patent holder's expected litigation payoff is $P_N\theta + P_Y(1 - \theta) - C$. It will accept a settlement without a reverse payment if this is exceeded by its settlement payoff of $TP_N + (1 - T)P_Y$. Thus, $T_{\min} = \theta - C/(P_N - P_Y)$. Accordingly, a settlement exclusion period that is lower than the optimal patent exclusion period is always possible without a reverse payment. Indeed, T_{\min} is always lower than the optimal patent exclusion period by $C/(P_N - P_Y)$. The expected litigation exclusion period for a weak patent is $\theta(1 - L)$. Thus, T_{\min} will exceed the expected litigation exclusion period if

$\theta L > C/(P_N - P_Y)$, which can be rearranged as $\theta L(P_N - P_Y) > C$. Accordingly, there are some cases where the minimum settlement exclusion period will necessarily exceed the expected litigation exclusion period.

The entrant's expected litigation payoff given a weak patent is $L[E - \theta(P_N - P_Y)] + (1 - L)(1 - \theta)E - C$. It will accept a settlement without a reverse payment if this is exceeded by its settlement payoff of $(1 - T)E$. Thus, $T_{\max} = \theta(1 - L) + (1/E)[\theta L(P_N - P_Y) + C]$. This maximum thus always exceeds the expected litigation exclusion period by $(1/E)[\theta L(P_N - P_Y) + C]$. This maximum also exceeds the optimal patent exclusion period if $(1/E)[\theta L(P_N - P_Y) + C] > L\theta$. This can be rearranged as $\theta L(P_N - P_Y - E) + C/E > 0$. This is always true given that joint profits without entry exceed joint profits with entry and litigation costs are positive. Therefore, the maximum settlement exclusion period always exceeds both the optimal patent exclusion period and the expected litigation exclusion period. Accordingly, for weak patents as well as strong, settlements without any reverse payment can produce settlement entry dates that exceed both the optimal patent exclusion period and the expected litigation exclusion period.

Given the above, $T_{\text{avg}} = \theta + \theta L(P_N - P_Y - E)/2 + C/(2E) - C/(2(P_N - P_Y))$. The second term is positive because joint profits without entry exceed joint profits with entry, so T_{avg} will always exceed the optimal patent exclusion period, θ , if $C/(2E) > C/(2(P_N - P_Y))$. This is always true given that $P_N - P_Y - E > 0$. Thus, T_{avg} will always exceed the optimal patent exclusion period. The expected litigation exclusion period for a weak patent is $\theta(1 - L)$, thus T_{avg} will exceed this expected exclusion period by this amount plus θL .

Therefore, even with zero reverse payment and a weak patent, the middle of the settlement range always exceeds both the expected litigation exclusion period and the optimal patent exclusion period. If we assume all settlements in the bargaining range are equally likely, settlements without reverse payments are usually anticompetitive for weak patents as well as strong.

Using the numbers above, we can get a sense of just how likely these anticompetitive effects are in the average case. Given those numbers, a weak patent exists if $\theta < 47.2\%$. For such a weak patent, if we plug the above values into the equation for T_{\min} , we get that $T_{\min} = \theta - 0.004243$. Plugging in the values into the equation for T_{\max} , we get that $T_{\max} = 0.005774 + 1.072\theta$. The portion of the settlement range above $\theta = (T_{\max} - \theta)/(T_{\max} - T_{\min}) = (0.005774 + 0.072\theta)/(0.010017 + 0.072\theta)$. This ranges from 58% to 90% for weak patents. Thus, if we assume parties are equally likely to reach any settlement in the bargaining range, a settlement with no reverse payment is 58% to 90% likely to exceed the optimal patent exclusion period even for a weak patent that cannot deter at-risk entry.

The expected litigation exclusion period for a weak patent is $\theta(1 - L) = 0.8004\theta$. The minimum settlement exclusion period $T_{\min} = \theta - 0.00424$. Thus, the minimum settlement exclusion period exceeds the expected litigation exclusion period if $\theta - 0.00424 > 0.8004\theta$, which is true if

$\theta > 2.1\%$. Accordingly, even without any reverse payment, in the average case the minimum settlement exclusion period exceeds the expected litigation exclusion period other than for a very weak patent that is less than 2.1% likely to win the patent litigation.

We can also ascertain the portion of the settlement range that exceeds the expected litigation exclusion period. Given the above, this portion is 100% for patent strengths between 2.1% and 47.2%. For patent strengths less than 2.1%, the portion of the settlement range that exceeds the expected litigation exclusion period is $(T_{\max} - 0.8004\theta)/(T_{\max} - T_{\min}) = (0.005774 + 0.2716\theta)/(0.010017 + 0.072\theta)$. Over this range of extremely weak patents, this portion ranges from 58% to 99%.

Accordingly, for a weak patent, like a strong one, a settlement with no reverse payment still results in a settlement range whose midpoint always exceeds both the expected litigation exclusion period and the optimal patent exclusion period. Assuming parties are equally likely to reach any settlement in the bargaining range, this means that settlements without reverse payments are usually anticompetitive. Further, using actual average numbers for such settlements, a settlement is 100% likely to exceed the expected litigation exclusion period unless the patent is extremely weak (less than 2.1% likely to win). Even for extremely weak patents whose patent strength ranges from 0% to 2.1%, the settlement is still 58% to 99% likely to exceed the expected litigation exclusion period. Further, the settlement exclusion period is 58% to 90% likely to exceed the optimal patent exclusion period.

3. *Summary.*—We have thus proven that, even with a zero reverse payment, a settlement that excludes entry for some period will produce a settlement range whose midpoint exceeds both the optimal patent exclusion period and the expected litigation exclusion period at *any* level of patent strength and for *any* level of market profits, residual patent period, litigation length, and litigation costs. Because we have no particular reason to assume that some settlements in the possible range are any more likely than others, it makes some sense to assume all of them are equally likely. If so, then we can say that all settlements that exclude entry for some period with no reverse payment are usually anticompetitive, regardless of the market particulars.

If we do use typical numbers for such settlements, we can go further and estimate the likelihood that they are anticompetitive. The graph below combines the above analysis to depict the portion of the bargaining range that exceeds the optimal patent exclusion period (vertical axis) at each level of possible patent strength (horizontal axis).

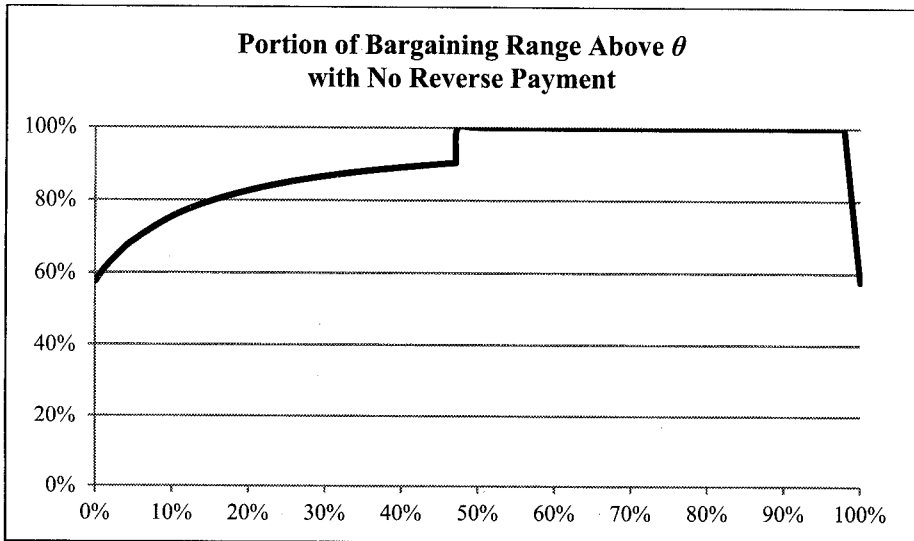


Figure 1.

As Figure 1 shows, this proportion exceeds 50% for all levels of patent strength. It ranges from 58% to 90% for weak patents (for which patent victory is less than 47.2% likely), but rises to 100% for strong patents (above this 47.2% threshold), unless the patent is extremely strong, in which case the proportion declines from 99% to 58% as the patent strength goes from 97.9% to 100%. It is thus at least 58% likely that the optimal patent exclusion period is exceeded at all patent strength levels, and usually the likelihood is much higher than that. Further, if the evidence shows that the patent was strong enough to deter at-risk entry, then the settlement will certainly exceed the optimal patent exclusion period unless patent victory was a slam dunk.

The next graph below combines the above analysis to depict the proportion of the bargaining range that exceeds the expected litigation exclusion period at each level of possible patent strength.

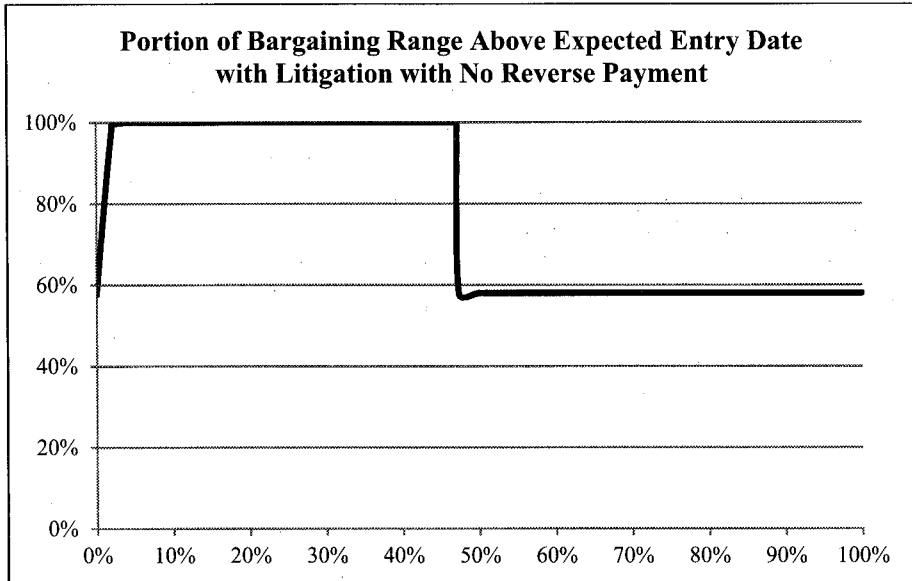


Figure 2.

This portion exceeds 50% for all levels of patent strength. It ranges from 58% to 99% for extremely weak patents (which are less than 2.1% likely to prevail), is 100% for all other weak patents (those 2.1% to 47.2% likely to win), and then drops back to 58% for strong patents. Thus, it is at least 58% likely that the expected litigation exclusion period is exceeded at all patent strength levels. Further, if the evidence indicates that the patent was weak enough that the entrant would have entered at risk, then the settlement will certainly exceed the expected litigation exclusion period unless a patent loss was virtually assured.

Further, the portion of possible settlement exclusion periods that will exceed *both* standards can be depicted by the following graph, which puts together the graphs above. The bottom edge of the range of likelihood is 58% with the likelihood increasing to 90% for some patent strength levels. In short, even without any reverse payment, the bulk of possible settlement exclusion periods will exceed both the expected litigation exclusion period and the optimal patent exclusion period at every possible level of patent strength.

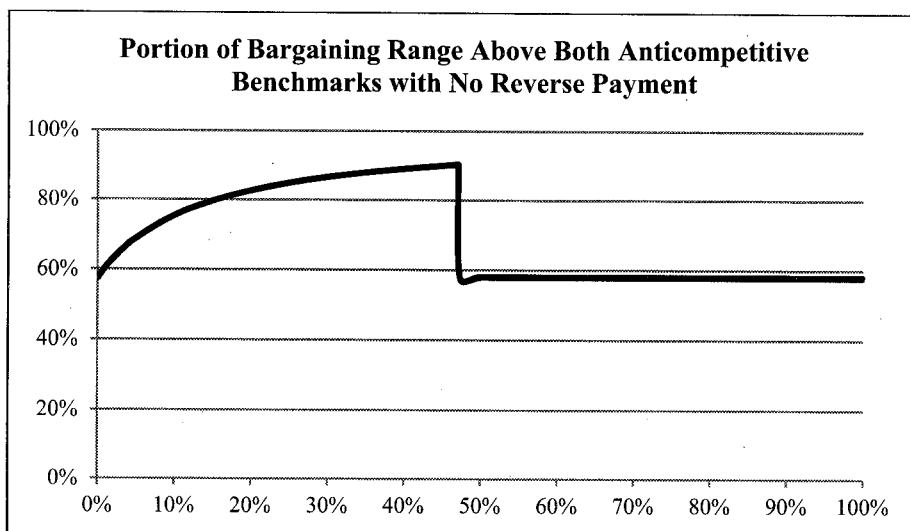


Figure 3.

The above analysis proved that these portions will exceed 50% regardless of the particular assumptions we make about market factors. To illustrate, suppose we cut in half our estimates of both the patent holder profits per month without entry and the residual patent period, so that the total patent holder profits without entry at stake are only one-fourth of what we estimated. Using the same analysis as above, we get the following graph for the portion of the bargaining range that is above both anticompetitive benchmarks for a settlement with zero reverse payment.

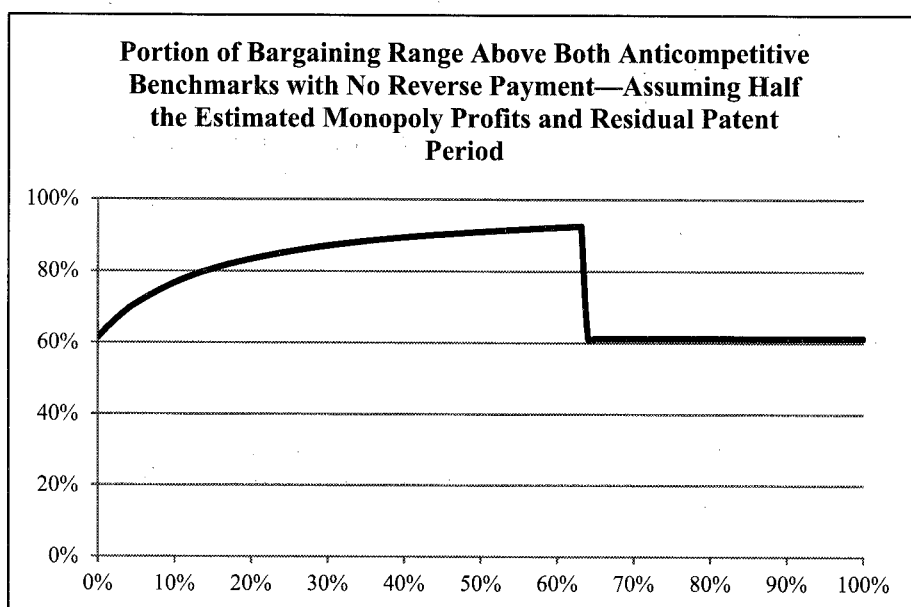


Figure 4.

With lower monthly profits and a smaller residual patent period, the threshold for a strong patent is higher (now 63.2%), but the portion of settlements that are above both anticompetitive benchmarks remains above 50% at each patent strength level. Indeed, the bottom of the range of likelihood is now higher, at 61%, as is the top of the range, now at 93%.

B. Grounds for Rebuttal and Possible Procedural Solution

The above analysis demonstrates that, even without any reverse payment, a settlement that excludes entry for some period will likely exceed both the expected litigation exclusion period and the optimal patent exclusion period. Because adding any reverse payment will only increase the settlement exclusion period, settlements with a reverse payment that is smaller than the patent holder's anticipated litigation costs are even more likely to cause these anticompetitive effects.

However, we cannot exclude the possibility that some such settlements might not be anticompetitive. This proof thus suggests that courts should in these cases also adopt a presumption of illegality, but allow it be rebutted by direct proof that the settlement exclusion period was shorter than the expected litigation exclusion period or the optimal patent exclusion period. The problem is that such a rebuttal would require the sort of direct case-by-case inquiry into probabilistic patent strength that many courts seek to avoid in antitrust cases.

One way to narrow the inquiry would be to add a market-power screen. Here such a screen makes sense because one cannot exclude the possibility that no market power exists if the reverse payment has not exceeded the patent holder's anticipated litigation costs. Further, there can be no harm to consumer welfare without market power, which is by definition the power to raise prices above competitive levels. However, this screen will not help in the typical patent settlement case where such market power can be proven.

Nonetheless, the above analysis can help bound the probabilistic analysis in a way that makes it more tractable. For example, suppose a court concludes that the relevant standard is whether the settlement exceeds the optimal patent exclusion period and that there is good evidence that at-risk entry would not have occurred. Then, if the case at hand matched average numbers for things like market profits and residual patent period, we know that the relevant standard must have been violated unless it was an extremely strong patent whose patent strength exceeded 97.9%. Even if courts would have difficulty assessing the precise probability of patent victory, it may be easier for courts to decide whether that lower bound seems likely to be exceeded. In an actual case, expert witnesses would simply plug in the case-specific values for market profits, residual patent term, expected litigation length, and anticipated cost to reach the appropriate lower bound for that case. Those values are easier to ascertain than the probability of patent victory.

Similarly, suppose a court concludes that the relevant standard is whether the settlement exceeds the expected litigation exclusion period and that there is good evidence that at-risk entry would have occurred. Then it can conclude that the relevant standard must have been violated unless it was an extremely weak patent, with the upper bound being 2.1% with typical numbers or by using another upper bound based on case-specific numbers.

However, a court will not be able to set upper and lower bounds that guarantee that both standards are violated in the same case. Thus, if it wants to allow rebuttal under both standards at once, then it cannot avoid a direct inquiry into probabilistic patent strength. Given the difficulty with this sort of inquiry, this might be unattractive but there may be no better alternative.

If courts do not think they can reliably assess probabilistic patent strength, one solution would be not to allow rebuttal at all. This would reach the wrong result in some cases, but by hypothesis, the problem is that courts cannot distinguish those cases. Therefore, their substantive choices are to either condemn all such settlements or allow them all. Given our proof that most settlements without reverse payments are anticompetitive, allowing all such settlements would produce worse results than condemning them all. To be sure, the magnitude of anticompetitive harm is much smaller without a reverse payment, but that does not make such harms desirable, and antitrust law generally has no exception for small anticompetitive harms.

Given the problems with these possible substantive responses, the better solution in such cases might be procedural. The underlying problem that allows anticompetitive patent settlements is that patent law ordinarily does not allow buyers to sue to prevent the anticompetitive exclusion of rivals through invalid patents. If patent law did allow such buyer standing, then patent holders and entrants could not collude in settlements that bar patent scrutiny of dubious patents at the expense of buyers. Those buyers would have a strong interest in challenging dubious patents.

There may well be good reasons to change this patent rule against buyer standing generally. But at a minimum, one could lift this bar on buyer standing in patent cases when the only rival or rivals that could have challenged the patent have settled in a way that prevents them from entering immediately. This sort of procedural remedy would sharply lessen the incentive for a patent settlement that excludes rivals because it could not preclude a buyer class action seeking to invalidate the patent.⁶⁴ When the patent is not dubious, then plaintiffs' attorneys would have little incentive to lose money by funding a class action to challenge the patent. But when the patent is dubious, they would have incentives to bring such a buyer class action, and courts could directly address the issue of whether the patent is

64. Because the suggestion is to allow only buyer patent actions to invalidate the patent prospectively, one need not worry that a risk of paying damages in such a buyer class action would deter the patent holder from ever entering into a settlement with the rival.

valid, rather than adjudicate the difficult issue of the probability with which the court thinks another court would have held the patent valid.

IV. Relationship to Prior Scholarship

Some leading antitrust and patent scholars have previously conjectured that reverse payments that exceed litigation costs should be deemed presumptively anticompetitive.⁶⁵ However, this conjecture has not previously been proven.⁶⁶ By providing this proof, we not only validate this conjecture, under two benchmarks that previously were usually conflated, but we also are able to more accurately specify the conditions under which this presumption holds and what sort of rebuttals should be permitted.

To begin with, while these scholars word this conjecture as applying when the reverse payment exceeds all litigation costs, our proof shows that the payment need only exceed the patent holder's anticipated future litigation costs. More importantly, our analysis proves that the appropriate grounds for rebuttal are very different from those suggested by prior proponents of this conjecture. None of them provide for rebuttals based on judgment-proof entrants or procompetitive justifications, which we show above are necessary. Further, they would all allow rebuttal based on grounds that our proof precludes.

Professors Hovenkamp, Janis, and Lemley would allow rebuttal by proof that (a) there was some "legitimate" likelihood of patent victory and (b) the settlement entry date was in the "range" of possible expected litigation exclusion periods.⁶⁷ We reject this possible rebuttal because our proof shows that the fact that a reverse payment exceeds the patent holder's anticipated litigation costs itself precludes the possibility that the settlement exclusion period is shorter than the expected litigation exclusion period. Nor, even if their proposed rebuttal could be established, would we find it sufficient because: (1) the fact that the odds of patent victory are "legitimate" does not mean that the settlement exclusion period did not exceed those odds and thus overreward innovation; and (2) the fact that the settlement exclusion period is within the "range" of the expected litigation exclusion period does not mean it did not exceed the actual expected litigation exclusion period and thus harm consumer welfare.

Professor Shapiro would allow this conjecture to be rebutted by proof of varying party estimates or risk aversion, and Professor Carrier would

65. See Michael A. Carrier, *Unsettling Drug Patent Settlements: A Framework for Presumptive Illegality*, 108 MICH. L. REV. 37, 75–76 (2009); Herbert Hovenkamp, Mark Janis & Mark A. Lemley, *Anticompetitive Settlement of Intellectual Property Disputes*, 87 MINN. L. REV. 1719, 1720, 1759 (2003); Carl Shapiro, *Antitrust Limits to Patent Settlements*, 34 RAND J. ECON. 391, 408 (2003).

66. Professor Shapiro does present proofs on other issues but not on this conjecture.

67. Hovenkamp, Janis & Lemley, *supra* note 65, at 1734–35.

similarly allow rebuttal based on informational asymmetries.⁶⁸ We would not do so because our proof indicates anticompetitive effects despite varying party estimates of patent strength, and we show that risk aversion would not alter our conclusions. Moreover, Shapiro's ultimate test is that patent settlements should be illegal if the settlement exclusion period exceeds the expected litigation exclusion period.⁶⁹ He thus uses only one of the two benchmarks we use and applies it as a case-by-case test rather than (as we do) as a policy benchmark by which to assess the desirability of a more administrable test. One problem with Shapiro's approach is that a settlement exclusion period could fail his benchmark and still be within the optimal patent exclusion period and thus help provide the patent holder with the appropriate reward for innovation.⁷⁰ Such a settlement might thus benefit *ex ante* consumer welfare more than it harms *ex post* consumer welfare, so that his test would not allay the concern of some courts about denying patent holders the full patent reward they deserve. The other problem is that his test requires a case-by-case inquiry into the probability of patent victory,⁷¹ which involves the sort of inquiry into the patent merits that the patent settlement was trying to avoid and the sort of probabilistic patent assessment that many courts have been reluctant to undertake in antitrust cases.

Finally, while we prove that even settlements without any reverse payment are generally anticompetitive, these prior scholars assumed that a settlement with no reverse payment will produce a settlement exclusion period that equals the expected litigation exclusion period. Hovenkamp, Janis, and Lemley thus favor a presumption of legality for such settlements with the only rebuttal being proof that the patent was a sham.⁷² Shapiro concludes that settlements without reverse payments should be *per se* legal.⁷³ Carrier also seems to advocate *per se* legality if there is no reverse payment or if the reverse payment is less than litigation costs.⁷⁴ Because we have

68. Shapiro, *supra* note 65, at 408; Carrier, *supra* note 65, at 77. Professor Carrier would also allow rebuttal if a cash-strapped generic needs cash quickly. *Id.* To the extent he means to rely on varying risk aversion, we would not allow rebuttal. To the extent he means that the generic might be judgment-proof, we agree with that possible ground for rebuttal, as limited by the conditions we prove are necessary to establish it.

69. Shapiro, *supra* note 65, at 396, 407–08.

70. Shapiro mistakenly conflates the expected litigation exclusion period with the optimal patent exclusion period, *id.* at 396, but as we show above, the former can be less than the latter when at-risk entry would have occurred without settlement.

71. *Id.* at 397.

72. Hovenkamp, Janis & Lemley, *supra* note 65, at 1762–63. They do suggest another possible rebuttal consisting of evidence that a reverse payment was actually made, but such evidence would mean that this presumption does not apply in the first place.

73. Carl Shapiro, *Antitrust Analysis of Patent Settlements Between Rivals*, ANTITRUST, Summer 2003, at 70, 72.

74. Carrier, *supra* note 65, at 76–77.

disproved their underlying assumption, we show that presumptive or conclusive legality is inappropriate even without any reverse payment.

Other professors have not focused on the relationship between reverse payments and litigation costs at all. Professors Daniel Crane and Thomas Cotter have instead focused on the absolute odds of patent victory. Crane argues that one should allow reverse payment settlements when the *ex ante* probability of patent victory is high but not when it is low.⁷⁵ But even if the probability of patent victory is high, a settlement exclusion period that is higher than merited by that probability would still be undesirable, and even if the probability of patent victory is low, a settlement exclusion period that is lower than merited by that probability would still be desirable. Thus, his test does not correspond to social desirability of the patent settlement and also requires the sort of case-by-case inquiry into the probability of patent victory that patent settlements and many courts in antitrust cases seek to avoid. Moreover, our proof shows that such a probabilistic inquiry into the patent merits is unnecessary when the reverse payment exceeds the patent holder's anticipated litigation costs.

Cotter shows that it can be rational for a patent holder to offer the entrant a reverse payment even if the odds of patent victory are high, and concludes from this that although reverse payments should be presumptively unlawful, this presumption should be rebuttable by proving the odds of patent victory are high, with "high" meaning at least 50% and certainly being provable by showing 75% odds.⁷⁶ However, the fact that a patent holder finds it rational to make a reverse payment tells us nothing about whether the settlement is desirable, especially because a settlement that excludes entry funds that reverse payment with other people's money—namely the money of buyers. Moreover, whether the probability of patent victory exceeds 50% or 75% also tells us nothing about the settlement's desirability. Even if the probability were 75%, a settlement that excludes entry for more than 75% of the residual patent period would still be anticompetitive, and even if the probability were 10%, a settlement entry date that covers less than 10% of the residual patent period would still be procompetitive. Further, his test also requires a difficult case-by-case inquiry into the probability of patent victory. Our proof shows that whether a reverse payment exceeds litigation costs provides a more reliable indicator of social desirability, without requiring any such case-by-case inquiry into what the probability of patent victory might be.

75. Daniel A. Crane, *Exit Payments in Settlement of Patent Infringement Lawsuits: Antitrust Rules and Economic Implications*, 54 FLA. L. REV. 747, 779–96 (2002).

76. Thomas F. Cotter, *Refining the "Presumptive Illegality" Approach to Settlements of Patent Disputes Involving Reverse Payments: A Commentary on Hovenkamp, Janis & Lemley*, 87 MINN. L. REV. 1789, 1807, 1812 & n.92 (2003).

Professor Blair argues that one should simply apply a rule of reason to reverse patent settlements to determine if their net effects are procompetitive.⁷⁷ But his rule would presume legality, which our proof shows is unwarranted, especially if the reverse payment amount exceeds anticipated litigation costs. Nor does he provide clear guidance as to how courts could conduct the suggested rule-of-reason analysis. Further, he suggests that one should not infer likely illegality unless the reverse payment is close to the amount of entrant profits from entry. We prove that the key comparison is instead to the patent holder's anticipated litigation costs.

Professors Willig and Bigelow argue that reverse payments may sometimes be necessary for desirable patent settlements, and they conclude from this that settlements with reverse payments thus should not be presumptively unlawful.⁷⁸ However, when one examines the details, one sees that their argument applies for a desirable settlement only when the reverse payment amount is "less than the incumbent's litigation costs."⁷⁹ Our proof shows this possibility goes away when the reverse payment amount exceeds those litigation costs, which fully justifies the presumption.

Further, Willig and Bigelow consider only whether desirable settlements are possible, not whether they are likely. We prove that even without *any* reverse payment, the lion's share of settlement exclusion periods that the parties could reach would be anticompetitive. This justifies presumptive condemnation even without a reverse payment, and thus even more strongly justifies it with a positive reverse payment amount, which only increases the share of possible settlements that are anticompetitive. Finally, in the end Willig and Bigelow simply argue that courts should sustain patent settlements if the settlement entry date is earlier than the expected entry date.⁸⁰ Their test thus, like Shapiro's ultimate test, both: (1) ignores the potential disjunction between the expected litigation exclusion period and the optimal patent exclusion period; and (2) requires the sort of case-by-case inquiry into the probability of patent victory that patent settlements and many courts seek to avoid.

Conclusion

In assessing whether patent settlements are anticompetitive, it is relevant to use two benchmarks that are often conflated: (1) whether the

77. Roger D. Blair & Thomas F. Cotter, *Are Settlements of Patent Disputes Illegal Per Se?*, 47 ANTITRUST BULL. 491, 533–34 (2002) (reporting the views of just Professor Blair).

78. See Willig & Bigelow, *supra* note 37, at 659–62, 667–77.

79. *Id.* at 671. Much of their analysis actually addresses a different question: whether a reverse payment might be necessary for patent settlement *without* showing that settlement would actually be desirable. Our proof shows that although this is true, a reverse payment that exceeds litigation costs is necessary only for undesirable settlements.

80. *Id.* at 662, 677.

settlement harms *ex post* consumer welfare by excluding the entrant for longer than expected from litigation; and (2) whether the settlement harms *ex ante* welfare by exceeding the optimal patent exclusion period and thus the optimal reward for innovation. However, courts have been reluctant to apply such benchmarks in a case-by-case way because it would require the sort of inquiry into the patent merits that settlement aims to avoid, with the addition of a probabilistic twist that is conceptually difficult for courts to resolve.

Our proof avoids this administrative difficulty by proving that, under ordinary conditions, a patent settlement with a reverse payment that exceeds the patent holder's anticipated litigation costs is *always* anticompetitive under both benchmarks. We prove that this is true even if the patent holder and alleged infringer differ in their estimates of patent victory. We also show that this claim should not be defeated by claims that market power was lacking, that the parties were risk averse, or that the particular settlement exclusion period did not violate the two benchmarks. On the other hand, we show that rebuttal is appropriate when the entrant would have entered at risk and is judgment proof to a sufficient extent. We also show that rebuttal is appropriate when there are other procompetitive justifications.

Finally, we show that, contrary to conventional wisdom, patent settlements that exclude entry without any reverse payment are also usually anticompetitive. However, such settlements are not always anticompetitive, so a broader array of rebuttal would be advisable. To the extent that those rebuttals require a probabilistic inquiry into the patent merits that is too difficult for the courts, then the best solution may be the procedural one of giving buyers standing to challenge the patent's validity.

Appendix

Proof That Reverse Payments Cannot Be Necessary for Settlement If Joint Profits with Entry Exceed the Patent Holder's Profits Without Entry

Weak Patent

$$T_{\max} = \theta_E (1 - L) + [\theta_E L(P_N - P_Y) + C_E + R]/E$$

$$T_{\min} = \theta_P - (C_P + R)/(P_N - P_Y)$$

The parties can settle only if $T_{\max} > T_{\min}$

$$\theta_E(1 - L) + [\theta_E L(P_N - P_Y) + C_E + R]/E > \theta_P - (C_P + R)/(P_N - P_Y)$$

Thus, if R increases by δ from 0 or any positive number, the left side (T_{\max}) will increase by δ/E and the right side (T_{\min}) will increase by $\delta/(P_N - P_Y)$.

Therefore, if $P_N - P_Y < E$ (just $P_N < P_Y + E$ rearranged) then $\delta/E < \delta/(P_N - P_Y)$, meaning that increasing a settlement payment by δ can only make it less likely that $T_{\max} > T_{\min}$. A corollary is that if $P_N - P_Y < E$ but the parties nevertheless settled, the parties must have necessarily been able to settle without any reverse payment.

Strong Patent

$$T_{\max} = \theta_E + L(1 - \theta_E) + (C_E + R)/E$$

$$T_{\min} = \theta_P + L(1 - \theta_P) - (C_P + R)/(P_N - P_Y)$$

Increasing R by δ from 0 or any positive number can only reduce $T_{\max} - T_{\min}$ if $P_N - P_Y < E$ because then T_{\max} would increase by only δ/E and T_{\min} would increase by the greater $\delta/(P_N - P_Y)$. Therefore, if $P_N - P_Y < E$ but the parties nevertheless settled, the parties must have necessarily been able to settle without any reverse payment.

Book Reviews

Henry Friendly: The Judge, the Man, the Book

HENRY FRIENDLY: GREATEST JUDGE OF HIS ERA. By David M. Dorsen.
Cambridge, Massachusetts: Harvard University Press, 2012. 512 pages.
\$35.00.

Reviewed by Mary Coombs*

For those of us who neither live in the rarefied world of the famous nor are aficionados of self-published memoirs by the earnest and obscure, reading a biography of someone we knew personally is a rare event. David Dorsen's biography of Judge Friendly—in addition to being a surprisingly engrossing read for anyone¹—was, for someone like me, both confirmatory and revealing.

The reason to remember Henry Friendly and write—or read—his biography is Friendly the Judge.² The subtitle calls him the “greatest judge of his era.” With this assessment (if not with all his holdings), I can heartily agree.

While one often associates Friendly with a mastery of the law, he also had a concern for getting the facts right, which was somewhat unusual for an appellate judge.³ He would pore through the record where the lawyers didn't cite to what seemed important to him. I believe that this focus on facts sometimes bridged his concern for reaching an outcome that seemed compatible with justice to the parties and his desire not to distort the law for future cases.⁴

* I would like to thank Warren Stern, my co-clerk, and my research assistant, Andrea Solano. All remaining mistakes and misjudgments are my own.

1. To misquote ALICE IN WONDERLAND, it is a book with conversations but, unfortunately, no pictures. LEWIS CARROLL, ALICE'S ADVENTURE IN WONDERLAND I (Richard Kelly ed., Broadview P. 2d ed. 2011). To be honest, the audience is likely limited to lawyers, which is still enough for respectable sales. (In 2010 there were an estimated 728,200 lawyers in the United States. *Occupational Outlook Handbook*, BUREAU OF LABOR STATISTICS (Apr. 26, 2012), <http://bls.gov/ooh/legal/lawyers.htm>.)

2. DAVID M. DORSEN, HENRY FRIENDLY: GREATEST JUDGE OF HIS ERA (2012).

3. This focus is perhaps less surprising when one considers his background as a lawyer for whom much of his practice was before administrative agencies and who was steeped in the forms of common law adjudication.

4. One example is Judge Friendly's finding parallels between the theology of Paul Tillich and the claims of Mr. Jakobson, a rather odd conscientious objector (CO), to find that Mr. Jakobson met the statutory standard for CO status. *Id.* at 245–47; *United States v. Jakobson*, 325 F.2d 409, 415–16 (2d Cir. 1963), *aff'd sub nom. United States v. Seeger*, 380 U.S. 163 (1965).

Dorsen provides an example of Friendly's fact consciousness in his discussion of the *Biaggi* case.⁵ Friendly wrote an opinion that released the transcripts of a grand jury investigation only after he knew what the grand jury transcripts revealed; namely, confirming his suspicion that Biaggi was trying to manipulate the courts with his motion to release in part the transcripts of the grand jury that was investigating him.⁶

During my term, we had a case where a would-be author sought the release of the Bureau of Alcohol, Tobacco, and Firearms "raid manual" under the Freedom of Information Act.⁷ Judge Friendly's concurring opinion found that the manual was protected by one of the exceptions in the statute and thus said that the plaintiff had no standing to question the constitutionality of the procedures set out therein.⁸ Before he wrote that opinion, however, he had instructed me to review the manual and inform him if it did seem to authorize any unconstitutional actions by agents. I believe it mattered to him that, in my estimation, the manual did not.⁹

His working process not only produced masterful opinions with great rapidity, but it also was as good an intellectual training ground as any clerk could receive.¹⁰ Immediately after the day's oral arguments, we were called *seriatim* to discuss the cases for which we were responsible—a discussion that began with him asking us what we thought.¹¹ If one could give an account of how the case should be decided that met with his approbation (if not his concurrence), one felt an extraordinary sense of achievement (or at least relief for not having stumbled).¹² And, as Dorsen notes, that discussion was immediately followed by the judge dictating his voting memo to his secretary.¹³ These were inevitably the first memoranda distributed to the other judges and, one assumes, they guided the way the case would be analyzed.¹⁴ Many judges rightly assumed that they should intellectually

5. DORSEN, *supra* note 2, at 222–26; *In re Biaggi*, 478 F.2d 489, 494 (2d Cir. 1973) (Friendly, J., supplemental opinion) ("It [the majority opinion] rests on the exercise of a sound discretion under the special circumstances of this case.").

6. DORSEN, *supra* note 2, at 222–26.

7. *Caplan v. Bureau of Alcohol, Tobacco & Firearms*, 587 F.2d 544, 548–49 (2d Cir. 1978).

8. *Id.*

9. I do not know if I was assigned this case in part because I had been a mentee of Yale Kamisar. For a description of Yale Kamisar and his work, see *Yale Kamisar*, U. MICH. L. SCH., http://web.law.umich.edu/_facultybiopage/facultybiopagenew.asp?ID=201.

10. In addition to the interactions with the judge, one learned by watching the production of great legal analysis and by hearing his responses to bad legal work by lawyers and, less frequently, other judges. DORSEN, *supra* note 2, at 87–88.

11. A similar description of the judge's working methods can be found in Lawrence B. Pedowitz, *Judge Friendly: A Clerk's Perspective*, 1978 ANN. SURV. AM. L. xl, xli.

12. As Dorsen notes, Friendly could be quite cutting about poor performances by lawyers, other judges, or clerks. DORSEN, *supra* note 2, at 87–88, 95–97. As I have told colleagues, he did not suffer fools gladly, and from his intellectual perch, there appeared to be many fools.

13. *Id.* at 91.

14. *Id.* at 90–91.

dominate their clerks; I think Friendly made a similar assumption about most other judges.

The description of Friendly as “Greatest Judge of His Era,” however, rests not merely on the judge’s working method or on his focus with facts, but also on his contribution to jurisprudence.¹⁵ The term brings to mind the iconic great judge of legal theory, Ronald Dworkin’s Hercules, who can (as all judges ideally should) find the single best solution to hard cases—one that is consistent both with a defensible interpretation of existing law and with a coherent understanding of deep principle.¹⁶ It also echoes Duncan Kennedy’s counter image of the judge as half-consciously following his ideological predispositions in interpreting law in hard cases.¹⁷

Based on Dorsen’s book and my impression, Friendly fits neither model.¹⁸ About as well as any real judge, he sought (most of the time) to get “the law” right, consistent with both his sense of justice to the parties and a set of predilections that did not fit neatly into any simple liberalism or conservatism. His substantive political views were sometimes aligned with conservatism, particularly in his critical stance toward Warren Court constitutional criminal procedure law, which impeded law enforcement even where there was no plausible risk of convicting the innocent and no fundamental right, in his view, at stake.¹⁹ But he also tended to favor

15. That contribution is circumscribed by the facts that he served on a lower federal court and that I examine a period decades after he was active. His impact was larger than that position would suggest. Since 2000, Supreme Court Justices have cited to Friendly’s judicial or other writings by name nineteen times (in Lexis Nexis, within the “Federal Court Cases, Combined” database, search the following: COURT(supreme) and “Judge Friendly” or “Friendly, J.” or “Henry J. Friendly” or “Henry Friendly”). This is not simply an artifact of John Roberts being his former clerk; he has also been cited by Ginsburg, Stevens, Souter, Scalia, Kennedy, and Sotomayor. *E.g.*, *Tellabs, Inc. v. Makor Issues & Rights, Ltd.*, 551 U.S. 308, 320 (2007) (Ginsburg, J.); *Stoneridge Inv. Partners, LLC v. Scientific-Atlanta, Inc.*, 552 U.S. 148, 177–79 (2008) (Stevens, J.); *Sosa v. Alvarez-Machain*, 542 U.S. 692, 712 (2004) (Souter, J.); *Wal-Mart Stores v. Samara Bros.*, 529 U.S. 205, 210 (2000) (Scalia, J.); *Brown v. Plata*, 131 S. Ct. 1910, 1946 (2011) (Kennedy, J.); *Blueford v. Arkansas*, 132 S. Ct. 2044, 2057 (2012) (Sotomayor, J.).

16. *See*, RONALD DWORGIN, *TAKING RIGHTS SERIOUSLY* 116–18 (1978) (theorizing that an ideal judge, such as the metaphorical Judge Hercules, would have complete knowledge of the law and sufficient time to decide all cases; in such circumstances, Judge Hercules could create the perfect rule for a particular case that justifies the law as a whole).

17. Duncan Kennedy, *Strategizing Strategic Behavior in Legal Interpretation*, 1996 UTAH L. REV. 785, 792–93.

18. Interestingly, Friendly himself does not discuss Kennedy in his writings but does briefly mention Dworkin. After noting the jurisprudential debate “generated by” Ronald Dworkin in *Hard Cases*, 88 HARV. L. REV. 1057 (1975), which discussed whether a judge may use policy or only principle in deciding cases where law seems indeterminate, Friendly concludes that “it is not clear to me how far apart, in any practically significant sense, the disputants really are.” Henry J. Friendly, *The Courts and Social Policy: Substance and Procedure*, 33 U. MIAMI L. REV. 21, 24 n.14 (1978).

19. DORSEN, *supra* note 2, at 188, 214–15; *see also* Henry J. Friendly, *Is Innocence Irrelevant? Collateral Attack on Criminal Judgments*, 38 U. CHI. L. REV. 142, 142 (1970) (arguing that “with a few important exceptions, convictions should be subject to collateral attack only when the prisoner supplements his constitutional plea with a colorable claim of innocence”).

prosecutors or unsophisticated investors in cases involving the regulation of business.²⁰ He also placed more consistent emphasis than either Dworkin or Kennedy does on the role of the judge in (a) creating and maintaining a coherent and predictable body of law,²¹ which he did not see as embodying a substantive preference for conservative and liberal policy choices,²² and (b) leaving space for institutions to make policy choices (sometimes the government, sometimes private institutions being protected from government).²³ Finally, he was sometimes (though not always, as Dorsen notes)²⁴ more modest than Hercules. On occasion, after seeking to turn his first analysis (usually from his voting memorandum) into an opinion, he would be brought up short by the existing legal materials and conclude, “It won’t write.”²⁵ The winning party might not change, but the argument would be revised to be consistent with his best reading of the law he was interpreting. Perhaps somewhere in the world there is or will be a Hercules who can always find a “right” opinion on every topic consistent with her philosophical principles. In the meantime, we are unlikely to find a better judge than one like Friendly, who so often got it right, who wrote so fluently²⁶ and so well, and who recognized when it “wouldn’t write.”

Nonetheless, a biography and a memory must also consider Friendly the Man, particularly as it may help illuminate Friendly the Judge. The book does so, based on interviews with surviving family, a wide range of other judges, and famous folk who could shed light on various aspects of Friendly’s life and character. Dorsen also interviewed every clerk Friendly had had. While each of us interacted with him intensely for only a year (and some of us almost not at all beyond that), that year was indeed intense. As the book demonstrates, there were common elements, but our experiences—

20. DORSEN, *supra* note 2, at 249–53.

21. He similarly criticized administrative agencies, largely in terms appropriate to courts as well, for doing a poor job of “[providing] standards and reasoned analysis” for their conclusions. *Id.* at 295.

22. I think he would reject Kennedy’s view that a preference for rules over standards is linked to a substantive “conservative” position. See Duncan Kennedy, *Form and Substance in Private Law Adjudication*, 89 HARV. L. REV. 1685, 1753 (1976) (connecting the conservative attack on judicial activism to a preference for judicial rulemaking and application of rules to judicial creation and enforcement of standards).

23. For more on Friendly’s jurisprudence, see generally Michael Boudin, *Judge Henry Friendly and the Craft of Judging*, 159 U. PA. L. REV. 1 (2010) (surveying Judge Friendly’s judicial decision-making process and noting Judge Friendly’s understanding of the importance of predictability and the maintenance of stable rules).

24. At one point, Dorsen chastises Friendly for his “creative, if not cavalier, treatment of precedent.” DORSEN, *supra* note 2, at 179.

25. See *id.* at 90–91, 150–51 (collecting examples of Friendly stating that he would change his ruling if one could find authority for the contrary position or expressing discomfort with a result he believed he could not avoid based on the law as it stood). Though he once said that “he could distinguish just about every decision,” he sometimes felt more constrained by the body of statutory and decisional law. *Id.* at 89.

26. One stylistic quirk: he had a habit of the “not quite double negative” (like “not unreasonable”). A Westlaw search found twenty-eight Friendly opinions using that locution.

and our assessments of those experiences—varied. In thinking back and in light of the book, I think those differences are in part a result of changes in Judge Friendly over time, in part a result of differences among us, and in part a reflection of the meshing—or not—of our personalities with his.

By the time I clerked for the judge in 1978–1979, he was on the downward arc of a judicial career that lasted from 1959 through 1986. It was clear that the opportunity for a Supreme Court appointment had passed. As Dorsen notes, Friendly was at times “dispirited,” if not necessarily clinically depressed.²⁷ He had difficult relationships with two of his three children.²⁸

It may be that his depression was worsening—certainly we rarely saw him cheerful. His eyesight seemed to have gotten worse with time,²⁹ a disability especially salient for someone whose professional life was so bound up with reading and writing.³⁰ He seemed to flourish largely in the company of his wife Sophie, who had the warmth and natural social skills he lacked.³¹ One feels acutely what a blow it must have been when she predeceased him.³²

I was very much an atypical choice for a Friendly clerk. I was one of only two women clerks (two years after Ruth Wedgewood) and, as a Michigan graduate, one of only four clerks not from Harvard (more than half), other Ivy League schools, or Chicago.³³ I was also, unusually I believe, a second-life law student; I turned 33 during my clerkship year. Together these may have made for a poor fit for the judge’s style in interacting with his clerks, apart from the more intellectual aspects of the court’s work.

Friendly was a man of his time, formed in an era before feminism. He lived in a fairly sheltered world, growing up comfortably middle-class in a small city and living for much of his adult life in a luxurious apartment on Park Avenue in Manhattan.³⁴ Neither his mother nor his wife worked outside the home. In his world, he succeeded by merit and may have been less sensitive to how merit alone would not suffice for all.³⁵ The judge read

27. *Id.* at 53.

28. *Id.* at 52–59.

29. *Id.* at 341.

30. He also relied extensively on his prodigious memory. This usually served him well, though clerks could be frustrated by his referring to some prior case that he thought relevant in a way of little use to a clerk with less than a year’s tenure and before Westlaw and Lexis, such as “the case with the lawyer from X firm.”

31. DORSEN, *supra* note 2, at 36–37, 55–56.

32. *Id.* at 339–40.

33. The other three “outreach” clerks are William Lake from Stanford, Martin Glenn from Rutgers, and William Bryson from the University of Texas. *Id.* at 361–66.

34. *Id.* at 6–10, 37.

35. Dorsen’s book gives little sense of Judge Friendly’s interactions with or understanding of the lives of racial minorities. His relationship with Judaism and WASP Anti-Semitism was complex, as shown by his somewhat inconsistent responses to Harvard President Lowell’s proposal for a quota for Jewish students. *Id.* at 18. My sense is that the judge was also largely insensitive to

widely: law, history, and legal philosophy, but not apparently social sciences or current affairs.³⁶ His history reading, given his interests and the forms of history most common during his formative years, would have been intellectual and political, not social history.³⁷ Only one person on the long list of his regular correspondents was a woman.³⁸ Dorsen notes that when Friendly and others formed Cleary Gottlieb, two of its newly-hired nine associates were women.³⁹ This was unusual at the time and place.⁴⁰ We do not know how Friendly interacted with these women lawyers. I was unsurprised by the anecdote Dorsen recounts of Friendly's shock that Ruth Bader Ginsburg responded negatively when he pulled out a chair for her at a luncheon.⁴¹ I suspect the shock was genuine surprise and dismay. My memory is that the clerks' dinners during his lifetime were held at the Century Association, his club. I doubt that the judge even really noticed that the Association had no women members.⁴²

Similarly, I resented—more than many clerks—the “menial” tasks that were expected of us, such as ensuring that the bench was prepared precisely to his requirements for each sitting⁴³ and that he always had a working pen.⁴⁴ When buzzed in, you would enter, take the pen held out in his nonwriting hand, replace the innards with those from a government issue pen and return it to him, all without exchanging a word or glance. What was for him, as Dorsen suggests, a manifestation of routine and hierarchy,⁴⁵ felt to me like patriarchy as well.⁴⁶

class—that there were (and are) people who grow up in economic and family circumstances that do not dare even to dream of Harvard, though their native intelligence might have permitted them to thrive there.

36. *Id.* at 10–14, 54–55.

37. His senior paper at Harvard explored the relations of Church and State in England under William the Conqueror. *Id.* at 16–17.

38. *Id.* at 101.

39. *Id.* at 60.

40. See VIRGINIA G. DRACHMAN, SISTERS IN LAW: WOMEN LAWYERS IN MODERN AMERICAN HISTORY 255 (1998) (stating that in 1939, only 14.2% of lawyers in New York were women); David M. Margolick, *Wall Street's Sexist Wall*, NAT'L L.J., Aug. 4, 1980, at 60 (stating that in the 1940s, very few lucky women found positions at New York's most prestigious law firms and almost all found positions in trusts and estates law).

41. DORSEN, *supra* note 2, at 118.

42. He was hardly alone. Women were not admitted until 1989 and then only after a very contentious battle. See Felicia R. Lee, *121 Years of Men Only Ends at Club*, N.Y. TIMES, July 28, 1989, at B1, available at <http://www.nytimes.com/1989/07/28/nyregion/121-years-of-men-only-ends-at-club.html?pagewanted=all&src=pm> (describing the end of the long battle to end the male-only policy of the New York Athletic Club and recounting the Century Association's admission of women the previous year).

43. DORSEN, *supra* note 2, at 108.

44. *Id.* at 93.

45. *Id.* at 108.

46. This may be my projection. Other clerks may not have resented this part of the job. In any event, they were unlikely to attribute it to patriarchy.

I do not mean to suggest that the clerkship year was some unrelieved Dickensian misery. His work style meant there was little relaxed interaction between judge and clerk. But clerks and secretaries could often relax and enjoy the chambers on the other side of the judge's closed door. Furthermore, while the judge did not show much of a warm sense of humor with his clerks, he did have wit and cleverness.

Dorsen mentions two opinions that were pivotal in my decision to seek and take a clerkship with Judge Friendly (though he wrote nothing quite so clever the term I was there): *Frigaliment Importing Co. v. B.N.S. International Sales Corp.*⁴⁷ and *Nolan v. Transocean Airlines*.⁴⁸ I, and at least one of Chief Judge Kaufman's clerks, saw a mix of wit and hostility in the way Judge Friendly, when he had arranged to ride home with Kaufman, would interrogate him during the ride on his views of recent advance sheets or slip opinions. As soon as the ride was arranged, Kaufman would reassign one of his clerks to prepare him for it. Dorsen's story of Kaufman's ignominious role in Friendly's Second Circuit nomination process⁴⁹ may go a long way in explaining this behavior, which we both thought showed more than just a desire to save money or make conversation on Friendly's part.⁵⁰

Friendly the Man—like Friendly the Judge—was a complicated individual. And it may be that his personal history was more impressive than even those aspects that made my clerkship a legacy. We always want our heroes without feet of clay. We want those of great accomplishment to be great as people. Life doesn't always cooperate. To the extent that the judicial legacy would have been less had my clerkship been more pleasant, that is a trade I would not—at least in retrospect—have thought worth making.

47. 190 F. Supp. 116, 117 (S.D.N.Y. 1960) (“The issue [in this case] is, what is chicken?”). *Frigaliment* was a contract case brought in diversity where the contract called for chickens and the plaintiff-buyer argued that this did not extend to stewing chickens. Friendly rightly used the standard tools of contract interpretation. As a cook and grocery shopper, I can say that stewing hens would be found in the “chicken” section of the meat and poultry case, but I would have been deeply unhappy if my husband had brought home a stewing hen when the grocery list included “chicken.”

48. 276 F.2d 280, 281 (2d Cir. 1960) (“Our principal task, in this diversity of citizenship case, is to determine what the New York courts would think the California courts would think on an issue about which neither has thought.”), *vacated and remanded*, 365 U.S. 293 (1961), *adhered to*, 290 F.2d 904 (2d Cir. 1961). Friendly concurred in my assessment (or vice-versa); he called this his “best opening paragraph.” DORSEN, *supra* note 2, at 315.

49. DORSEN, *supra* note 2, at 74–75.

50. *Id.* at 120–21.

On Becoming a Great Judge: The Life of Henry J. Friendly

HENRY FRIENDLY: GREATEST JUDGE OF HIS ERA. By David M. Dorsen.
Cambridge, Massachusetts: Harvard University Press, 2012. 512 pages.
\$35.00.

Reviewed by Frederick T. Davis*

In writing a biography of Henry Friendly, author David Dorsen has taken on an enormous challenge: the subtitle is “Greatest Judge of His Era”—a claim that few who knew Judge Friendly, or are familiar with his remarkable legal legacy, would dispute. Judge Friendly left an unparalleled body of written opinions from his twenty-five-year career on the bench and was a vigorous presence at the very highest level of his profession through prolific writings, energetic participation in groups such as the American Law Institute, and his many professional friends.¹ His opinions remain, even today, among the most cited in the federal jurisprudence;² for those who knew him, he was an incomparably towering influence. To summarize the life of this remarkable person, and to offer some explanation of how he developed his formidable skills and extraordinary impact, is no easy task. David Dorsen does a remarkable job. His biography is not only rewarding for those who knew Judge Friendly or are familiar with his work, but also provides a readable and accessible exploration of how one person arrived at such a remarkable level of excellence in his profession.

I was a law clerk for Judge Friendly during the 1972–1973 term of the United States Court of Appeals for the Second Circuit. As it was for every lawyer who had this extraordinary opportunity, the year was one of the most remarkable experiences of my professional life. Unusually for a judge who died more than twenty years ago, his law clerks still reunite every three years or so to share recollections about our year with the Judge and his impact on our own thoughts and careers. This is no group of underachievers—it includes a number of very prominent professors and judges, including the Chief Justice of the Supreme Court—yet the prevailing sentiment is universally one of awe, occasionally tinged with a sense of fear that Judge Friendly might somehow look over our shoulders and remind us of standards of excellence that all of us still strain to meet.

* Frederick T. Davis is a partner in the Paris office of Debevoise & Plimpton LLP and a member of the Paris and New York bars. He was a law clerk for Judge Friendly in 1972–1973.

1. DAVID M. DORSEN, HENRY FRIENDLY: GREATEST JUDGE OF HIS ERA 3, 131–33 (2012).

2. *Id.* at 353–55.

I approached the Dorsen biography with a particular question that has always fascinated me: how was it that the son of a small-town manufacturer in upstate New York became the titan of his profession?³ Is it possible to find an explanation, or even a description, of his path to brilliance? A few years before he died, Judge Friendly permitted me to spend several hours tape-recording his reminiscence from both before and during his judicial career. While those recordings were transcribed, I never succeeded in editing or publishing them, and thus was thrilled when David Dorsen took them over and skillfully used them in his biography.⁴ Complemented by the thorough research he has done and access to Judge Friendly's files, friends, and family, the biography offers some clues to Friendly's emergence as one of the principal legal voices of his generation.

The first clue may seem obvious: Henry Friendly was simply a brilliant intellect, endowed with extraordinary skills. David Dorsen describes, and all of Friendly's law clerks well remember, the Judge's ability to sit down at a table with a ballpoint pen and two pads—one for the text of his opinions, the other for the footnotes—and simply write them out in one draft, often in one sitting, citing precedent from memory and when necessary marching over to find the text of the decision he wanted to quote, from memory pulling exactly the right volume of the Federal Reporter from the shelf. This technical brilliance was not a late development. When he arrived at Harvard College in 1919 at age 16, he had a keen interest in mathematics and took the most advanced course in mathematics available to entering undergraduates.⁵ When the grades arrived, he had received the second-highest grade ever received by a student in the history of the course. To his chagrin, however, the holder of the highest grade—by a minuscule margin—was a classmate. That was enough for Henry Friendly: he abandoned any dreams of becoming a mathematician.⁶ I had heard this story before doing my oral history with the Judge, and after confirming its basic outlines I was about to move on when I casually asked who the other student had been. It turns out that the competitor had been Marshall Stone, son of future Chief Justice Harlan Fiske Stone, who went on to have a distinguished career as a Professor of Mathematics at Harvard, and is credited with discovering several noted theorems. To be even neck-and-neck with such a scholar would be beyond the competence of virtually any other student, but to Henry Friendly being anything other than the best was insufficient. He later majored in European history, and when the time came for him to defend his thesis in an oral exam, the number of professors and students who wanted to watch was so great that the event took place in the Sanders Theater at Harvard College.

3. *Id.* at 5–6, 8.

4. *Id.* at 371–72.

5. *Id.* at 14.

6. *See id.* (“He changed his mind [about taking additional math classes] when he compared his performance in one course [with his classmate] Stone’s.”).

Undoubtedly through his mother, Friendly early on developed a passion for learning and an intellectual curiosity of extraordinary scope. His mother was evidently a woman of intellect and energy.⁷ Nor was she lacking in ambition for her near-sighted and unathletic son: after he arrived at Harvard College, she wrote to Professor Felix Frankfurter, who was known to her through a family connection, and who quickly befriended this young prodigy and did his utmost to entice him into the study of law.⁸ The persuasion was not immediately successful: Friendly remained fascinated with (and deeply knowledgeable about) European history throughout his life, and upon graduation at the top of his class in 1923 was courted not only by Professor Frankfurter at the law school but by the leading professors in liberal arts to pursue a career in academics.⁹ After a year of studying abroad to consider his options, he entered the law school¹⁰—but only really made up his mind to commit to the practice of law after receiving his first round of grades. He went on to achieve an academic record at Harvard Law School that, according to many, ranks even today as the statistically highest performance of any student in the history of the School.¹¹

The key trait that emerges from the Dorsen biography is that once Friendly focused on the law, he made it the passion of his professional life with a sustained and unwavering focus. With energy, curiosity, voracious reading habits, and prodigious memory, he saw the law in all of its dimensions—not as a series of rules to be memorized, nor even as tools to achieve ends, but rather as a process that goes to the core of society and how it is supposed to work. To this passionate commitment he brought insights drawn from his remarkable knowledge of history, literature, and philosophy. A trivial anecdote brought home to me the breadth of his reading and the depth of his ability to recall: once when I was with him he noticed that I was carrying a book and, with characteristic inquisitiveness, asked me what it was. It turned out to be a long and quite dense history of Russia, which I was going to visit for the first time later that year. “Oh,” he said, “that seems familiar, I think I read it once.” But, he then went on, “I must have read a different book because the one I read was more than one volume.” I checked, and sure enough the book I was reading was a one-volume simplification of an exhaustive seven-volume history of Russia—which the Judge had not only read, but mastered: when he questioned me about my meager insights from the slimmed-down version, it was clear that his grasp of the subject many times exceeded mine, even though he had read the lengthy opus more than twenty years before.

7. *Id.* at 6–7.

8. *Id.* at 20–21.

9. *Id.* at 20.

10. *Id.*

11. *See id.* at 26 (outlining Friendly’s excellent academic performance at law school).

When he joined the bench in 1959, Friendly brought to the job prodigious academic skills, broad learning, and more than three decades of challenging practice—which included founding what is today one of New York’s major law firms, and serving as General Counsel for Pan American Airways at the apex of its success as the first truly international American airline.¹² But most importantly, he brought an uncanny ability not only to parse a legal issue, but to see it in its three-dimensional context, shorn of ideology or preconceived notions. Before joining the bench, for example, Friendly had had relatively little experience with criminal procedures—he had never been a prosecutor or a criminal defense lawyer.¹³ Yet to this day, his decisions in this area are beacons of thoughtfulness and common sense, as well as learning. Many thought of him as a pro-government “conservative,” in part based upon a superficial interpretation of one of his well-known articles entitled “Is Innocence Irrelevant?,” in which he questioned some aspects of federal review of state criminal convictions via habeas corpus.¹⁴ But in each criminal case before him, his interest was in understanding exactly what happened in the case in question, and whether the procedures met the standards of transparency, honesty, and excellence that society demands. During my clerkship year, he wrote opinions in at least two instances reversing convictions because he felt that the prosecutor or the trial judge had not acted appropriately—even though the innocence or guilt of the accused was not really in question.¹⁵ In each case, he delved into the facts in meticulous detail, and concluded that the process had not satisfied acceptable standards upon which he insisted.

Judge Friendly was an internationalist. His work with Pan Am and his law firm put him at the cutting edge of international business during and after World War II.¹⁶ He read widely in French, once publishing a review of a lengthy French-language legal treatise¹⁷ and, as a student, remarking to a startled professor that a text apparently written in early English was actually in Law French, which Friendly offered to translate.¹⁸ But his heart was in the common law, where his insights derived not only from American precedent but from his deep understanding of English precedent as well. In his

12. *Id.* at 60–61.

13. *Id.* at 81.

14. Henry Friendly, *Is Innocence Irrelevant? Collateral Attack on Criminal Judgments*, 38 U. CHI. L. REV. 142 (1970).

15. *See generally* *United States v. Fernandez*, 480 F.2d 726 (2d Cir. 1973) (reversing a robbery conviction on the grounds that the trial judge’s questioning and discernible distrust of the defense’s expert witness was both improper and prejudicial); *United States v. Estepa*, 471 F.2d 1132 (2d Cir. 1972) (reversing a conviction for the prosecutor’s improper use of hearsay before a grand jury).

16. DORSEN, *supra* note 1, at 61–68.

17. Henry J. Friendly, Book Review, 54 HARV. L. REV. 169 (1940) (reviewing JEAN VAN HOUTTE, *LA RESPONSABILITÉ CIVILE DANS LES TRANSPORTS AÉRIENS INTÉRIEURS ET INTERNATIONAUX* (1940)).

18. Michael Boudin, *Judge Henry Friendly and the Mirror of Constitutional Law*, 82 N.Y.U. L. REV. 975, 977 (2007).

legendary *Kinsman Transit* tort decision,¹⁹ where he explored and essentially recast the law of causation,²⁰ he delved into English precedent at some length and with noteworthy insight—even though the applicability of that law had not been argued by either party.²¹ While respectful of the separation of the powers and the legislative function, he earnestly believed that judges contributed to the making of the law, and did not just interpret it in the manner of his continental counterparts. When the Federal Rules of Evidence were discussed, and ultimately adopted, in the 1970s, they were the culmination of years of work;²² today they are a fundamental component of federal trial practice. But Judge Friendly was not a fan because he felt that codified rules could never match the nuances and contextual appropriateness of judge-made decisions, and would stultify the flexibility and evolution of the law of evidence. It did not appear to occur to him that many judges, lacking his erudition, memory, and objectivity—Judge Friendly read *Wigmore on Evidence*²³ so thoroughly that he virtually had it memorized—would be helped by having a handy, consistent code of common-sense rules.

What are we to make of this remarkable man, looking back more than 25 years after his death?

On the credenza behind the desk in his chambers, there was a black-and-white photograph of Justice Louis Brandeis, for whom Henry Friendly served as law clerk at the beginning of his legal career after graduating from law school in 1927.²⁴ On it the Justice had scrawled “To Henry Friendly, a born lawyer.” While prescient, these words may understate Judge Friendly’s achievement: he was “born” with prodigious skills, but he became a masterful lawyer and judge through hard work, passion, an open mind, a high degree of curiosity, and relentless focus—and, to my mind, with an unwavering, almost brutal insistence upon intellectual honesty. While we are unlikely to see his like again, David Dorsen’s biography reminds us of the standards of excellence on which Judge Friendly insisted and the importance they hold for his profession today.

19. *In re Kinsman Transit Co.*, 338 F.2d 708 (1964).

20. *Id.* at 719–26.

21. David M. Dorsen, *Judges Henry J. Friendly and Benjamin Cardozo: A Tale of Two Precedents*, 31 PACE L. REV. 599, 610 n.69 (2011).

22. Paul R. Rice & Neals-Erik William Delker, *Federal Rules of Evidence Advisory Committee: A Short History of Too Little Consequence*, 191 F.R.D. 678, 683 (2000).

23. JOHN HENRY WIGMORE, A TREATISE ON THE ANGLO-AMERICAN SYSTEM OF EVIDENCE IN TRIALS AT COMMON LAW (2d ed. 1923).

24. DORSEN, *supra* note 1, at 27.

Henry Friendly: As Brilliant as Expected but Less Predictable

HENRY FRIENDLY: GREATEST JUDGE OF HIS ERA. By David M. Dorsen.
Cambridge, Massachusetts: Harvard University Press, 2012. 512 pages.
\$35.00.

Reviewed by Peter Edelman *

We are fortunate that David Dorsen is Henry Friendly's biographer. Chronicling the life of a judge, in this case a longtime and hardly flamboyant private practitioner before he became a judge, could easily yield a product about which "wooden" would be a compliment. Dorsen is sophisticated, very smart and very wise, a fine lawyer, and a really good writer. He has produced a highly readable and truly interesting book. He combines astute analysis of the voluminous list of cases on which Judge Friendly sat with perceptive discussion of their political and social context and significance—both inside the court and in relation to the world outside. It is actually entertaining.

I was Henry Friendly's third law clerk from 1961–1962, more than fifty years ago. For a variety of reasons, I did not follow closely his judicial output or other writings after that time, so for me that portion of Dorsen's book (which is most of it) was largely new and in many respects fascinating. The first part of the book covers quite well his family life, with which I was already quite familiar, and his earlier career as a practicing lawyer. This is all interesting and important to know to get some idea of Friendly the man outside the court. But it is when Dorsen turns to the judicial substance that I find myself enthralled. It was a totally pleasant surprise to discover how engaging the descriptions and backstories of the cases were.

My experiences as Judge Friendly's law clerk bear out Dorsen's account, although almost entirely on the positive end of the continuum. He was rigorous and demanding but seldom if ever short with me in a hurtful way. I drafted only one opinion, a very short one toward the end of the year, and I distinctly remember feeling something close to ecstatic when he gave me the assignment. I had a similar rush earlier in the year when he allowed me to write a few paragraphs of the opinion in a fairly complicated case. And now and then he asked me to draft a textual footnote. Yes, those were special moments, too. Just out of law school, I thought this was how all judges functioned.

* Professor of Law, Georgetown University Law Center.

Friendly's routine was largely as Dorsen describes. He rarely took more than a day to write an opinion. At 5:30 or so, ready to embark for home, he would emerge with a finished draft opinion, at which point (after his secretary, Mrs. Flynn, typed it) it would be my job to cite-check it.¹ The draft often contained citations to old English cases. How did he find them, I wondered. I knew he didn't have the original cases in his office. Did he have a secret door or escape hatch in his office from which he could get to a library that had the cases? There seemed to be two answers. One, he knew some of the citations by heart, and two, he found the citations in later cases and cited the earlier originals. Either way, my job was to check his work. Either way, it was impressive.

Conversations in his office were brief. Whether I was asking or answering a question or putting forward an idea, the drill was largely the same. I would get half a sentence out and he would finish my sentence and then respond. Usually he had grasped instantly and correctly what I was trying to say. (Remember, he was brilliant, and that's an understatement.) I would be somewhat at my peril if I wanted to disagree or suggest that he had not understood what I was trying to say. Mostly he would say gruffly that he had heard me correctly the first time (although sometimes he hadn't) and occasionally I would get a second shot.

(A parenthetical note: I had another boss whose *modus operandi* in office conversations was exactly the same—Robert Kennedy. Kennedy and Friendly were poles apart in many ways, but conversations in their offices about work issues were identical. Just like Friendly, Kennedy would jump in and finish my question or suggestion and then reply. Getting a second bite at it was similarly iffy. Notably, Kennedy was brilliant, too, in an especially intuitive way.)

Dorsen notes that Friendly held grudges and cites one example that involved me.² Writing this review gives me a chance to clarify the facts about that. The story was about Judge Friendly's lobbying Robert Kennedy through me in 1966 to get him to ask President Johnson to nominate Judge Edward Weinfeld for a seat on the Second Circuit.³ The ultimate result was that the appointment went to then-district Judge Wilfred Feinberg (who has been an outstanding appellate judge for the better part of fifty years),⁴ and Friendly blamed Robert Kennedy and me for not advocating strongly enough for Judge Weinfeld.⁵

Friendly was wrong on multiple counts. He had to know full well that Senators can only suggest court of appeals nominations to the President, as opposed to the process used for district court appointments when Senators

1. DAVID M. DORSEN, HENRY FRIENDLY: GREATEST JUDGE OF HIS ERA 87, 94 (2012).

2. *Id.* at 119.

3. *Id.*

4. *Id.*

5. *Id.*

from the relevant state are members of the same party as the President.⁶ And he should have known that in any event Robert Kennedy and Lyndon Johnson did not like each other very much, to put it mildly. At Kennedy's direction, I told Friendly at least twice that Feinberg's brother was a major donor to Johnson and a personal friend as well, and that Weinfeld should get the labor leaders, David Dubinsky and Jacob Pitofsky, whom he knew well, to lobby the President on his behalf. Kennedy did try quite hard to sell the White House on Weinfeld, and I told Friendly that more than once. When I had first broached the matter to Kennedy, he said immediately that he felt badly about having elevated Judge Irving Kaufman instead of Weinfeld some years earlier when he was Attorney General, and he wanted to rectify the mistake. I had conveyed this to Friendly as well. Nonetheless, Friendly was furious at the Senator and me when Judge Feinberg was named to the seat.

I was frankly hurt by this. I did not go to the annual clerk dinners for a couple of years (I was pretty busy, too), but what Dorsen does not report (which may be my fault) is that my wife Marian and I invited the judge to the naming ceremony for our first child, Joshua, in 1969, and he came specially from New York to Washington to attend. So maybe the grudge was on my side rather than Friendly's, and quite possibly Friendly's attendance at the ceremony was his way of apologizing for his anger three years earlier. When I resumed going to the dinners during the Nixon years, the judge always called on me along with three or four others to talk about my latest activities. As Dorsen points out, Friendly strongly valued people who engaged in public service.

Another personal note, about Friendly's wife Sophie. As Dorsen points out, my then-wife, Arlyn, and I were invited to dinner fairly frequently throughout the year that I clerked.⁷ This was more about Arlyn than it was about me, because Arlyn was charming in the same kind of way that Sophie was, and both the judge and Sophie were captivated. Whatever the reason, we saw Sophie's charm and life spirit firsthand and also got to know the Friendly children, especially Joan and her husband, Frank Goodman. Sophie was an extra special person, and we saw the "other" Henry Friendly on those evenings (which, fortunately, seemed to have a positive effect on our office relationship as well).

I did not think of Friendly as conservative or liberal when I clerked for him. I see now that he was in general a moderate conservative in the vein of John Marshall Harlan. But I didn't have the perspective to understand that at the time. This was at least partly because the Harvard Law School of the day enshrined Felix Frankfurter as the model Justice and "neutral principles" as the reigning judicial philosophy, and discussions of legal issues at Harvard (at least as I recall) did not articulate issues in terms of conservative or liberal

6. DENIS STEVEN RUTKUS, CONG. RESEARCH SERV., RL34405, *ROLE OF HOME STATE SENATORS IN THE SELECTION OF LOWER FEDERAL COURT JUDGES* 1 (2008).

7. DORSEN, *supra* note 1, at 111.

values. We were taught that the Roosevelt Court had settled everything both substantively and in terms of the judicial role.

To me, reflexively, Judge Friendly was the perfect example of the Harvard idea of the law, although that idea was so ingrained in me that I did not even articulate to myself that there was a Harvard idea of the law. (Justice Brennan was a Harvard Law graduate, too, but he had not been on the *Law Review* and the “vibe” I felt at the school was that he was not of the same caliber as Justice Frankfurter. I think this may have been more of an elitist view than a reflection of a philosophical difference, since my sense is that in the Harvard world of the day there was not much of a consciousness about there being such a thing as a liberal versus conservative divide.)

As Dorsen makes amply clear, Friendly was not a down-the-line conservative. Beyond whatever Harvard Law School had done to shape my thinking about how to approach the law, I saw Friendly as a person who looked carefully and thoughtfully for the right answer—certainly for the right answer on the law but also for the right answer in relation to the facts presented in individual cases where there was some give in the law.

This is borne out very clearly in the book. The matter of Philip Kerner is a case in point.⁸ Kerner had been denied Social Security disability benefits.⁹ It would have been easy to affirm the district court’s award of summary judgment to the government. Such outcomes are a daily occurrence. But Friendly, digging into the case, became convinced that Kerner was being unjustly treated and that the evidence in the record did not support the conclusion that Kerner could still perform substantial gainful activity in the economy and was therefore not disabled.¹⁰ Whether Friendly knew it or not, his legal analysis challenged the routine approach to such cases. He said “[m]ere theoretical ability to engage in substantial gainful activity is not enough” and then said “the evidence as to employment opportunities was even less.”¹¹ The government must have been considerably less than pleased at that formulation. If those words had been in a Supreme Court decision, advocates for the disabled would have been overjoyed.

But Friendly wasn’t trying to make law. He was trying to do justice for Philip Kerner. He subsequently wrote to a friend that “[t]he way Kerner got polished off was utterly disgraceful.”¹² And then, worried that Kerner would lose on remand, he reached out to an acquaintance at a cardiac rehabilitation center to see if he could arrange for Kerner to get medical help. The story goes on, but what I have already said makes the point. He was quite susceptible to getting engaged in the equities of the facts of cases about

8. *Id.* at 174–76.

9. *Kerner v. Flemming*, 283 F.2d 916, 918 (2d Cir. 1960); DORSEN, *supra* note 1, at 174.

10. DORSEN, *supra* note 1, at 174–75.

11. *See id.* (quoting from *Kerner*, 283 F.2d at 921).

12. *Id.* at 175.

ordinary people and then walking more than the last mile to pursue a just result.

Every law student learns about *Bivens* torts. In *Bivens v. Six Unknown Agents of Federal Bureau of Narcotics*,¹³ the Supreme Court held that the Constitution creates a cause of action for damages against federal officials who violate the civil rights of private individuals.¹⁴ Friendly's role in the case is a major example of why it is difficult to pigeonhole him ideologically and, as well, of his occasional proclivity to act outside of the usual judicial boundaries.¹⁵

Friendly was sitting as the motions judge one day when he came across Webster Bivens's motion for leave to appeal *in forma pauperis*. Bivens had sued for damages after federal officers arrested him in his home without a warrant and handcuffed him in front of his wife and children.¹⁶ The district court had dismissed the case, saying what had always been assumed to be the law: federal officials acting in the performance of their duties could not be sued in this kind of case.¹⁷ Friendly, knowing that state officials could be sued under similar circumstances because of 42 U.S.C. § 1983, which had been enacted in the wake of the Civil War, saw an injustice in the disparity between the accountability of state officials and federal officials. He dragooned a recent clerk and by-then Wall Street lawyer, Stephen Grant, to take Bivens's case.¹⁸

Grant lost. Friendly, who had not been on the panel on the merits, wrote Grant a brief note suggesting that he "take the matter further."¹⁹ Then followed another note suggesting the lines of an argument to make to the Supreme Court. Grant won.²⁰ Quite a story.

The book is well-stocked with other examples of Friendly's role in cases that piqued his interest on a human level as well as numerous examples of his significant role in cases that went on to the Supreme Court. He became a prolific writer of important books and articles on an array of legal matters, and of letters, sometimes for publication, expressing views on issues of the day. As with his work on the court, he was far from predictable, although there were certainly areas in which he had clear conservative views. He not only had an open mind across a spectrum of issues, but was willing to change his position on thinking about it more. *Baker v. Carr*,²¹ decided while I was Judge Friendly's clerk, is an example. I remember the outrage

13. 403 U.S. 388 (1971).

14. *Id.* at 395–96.

15. DORSEN, *supra* note 1, at 183–85.

16. *Bivens*, 403 U.S. at 389–90.

17. *Id.* at 390.

18. See DORSEN, *supra* note 1, at 183–84 (discussing how Grant initially stated that he was not a litigator but that Friendly ultimately assigned Grant to represent Bivens).

19. *Id.* at 184.

20. *Id.*

21. 369 U.S. 186 (1962).

he expressed to me about the decision, and I knew that he had written to Justice Frankfurter, calling his dissent "magnificent" and "one of your truly wise and great opinions."²² I thought at the time that his anger was misplaced and was interested to note in Dorsen's book that Friendly changed his mind about the case a few years later and said so publicly.²³ I was glad to see that, both in itself and for what it says about the man.

In an age in which moderation is increasingly rare in the conservative world, both judicially and politically (realms that increasingly overlap), Henry Friendly is a man to remember with special respect. Had he sat on the Court, I am sure I would have disagreed with many, although far from all, of his opinions and votes, but I know I would have respected his reasoning and scholarship. He was a man of reason, above all.

22. DORSEN, *supra* note 1, at 125.

23. *Id.*

Assembly Resurrected

LIBERTY'S REFUGE: THE FORGOTTEN FREEDOM OF ASSEMBLY. By John D. Inazu. New Haven, Connecticut: Yale University Press, 2012. 288 pages. \$55.00.

Reviewed by Ashutosh A. Bhagwat*

After a long period triggered by 9/11 and the Bush Administration's response to it, when constitutional law was focused on issues such as executive power and the Fourth Amendment, the First Amendment is back in the forefront of judicial and academic attention. In the past several years, the Supreme Court has issued a series of important, even path-breaking, decisions focused on the scope and limits of the freedom of speech.¹ At the same time, academic attention has turned to the role that First Amendment freedoms, including freedoms other than free speech, play in our society. Important examples include Timothy Zick's *Speech Out of Doors*,² which discusses the relationship between assembly, expression, and public places³ and Ronald Krotoszynski's *Reclaiming the Petition Clause*,⁴ which examines the role that the Petition Clause of the First Amendment can play in modern politics.⁵ We have also seen a flurry of recent law review articles examining the rights of association and assembly, and their relationship to democratic self-governance.⁶ These are, in short, exciting times for those interested in First Amendment freedoms and their place in the constitutional order.

* Professor of Law, U.C. Davis School of Law. Thanks to Ben Strauss for excellent research assistance and to Ralph Mayrell and the staff of the *Texas Law Review* for inviting me to write this Review.

1. See, e.g., *United States v. Alvarez*, 132 S. Ct. 2537, 2551 (2012) (holding that an act criminalizing false claims to military medals was a violation of free speech); *Snyder v. Phelps*, 131 S. Ct. 1207, 1220 (2011) (holding that nondisruptive antihomosexual picketing outside a funeral was protected free speech); *Brown v. Entm't Merchs. Ass'n*, 131 S. Ct. 2729, 2741–42 (2011) (holding that an act prohibiting sales of violent video games to minors was a violation of free speech); *United States v. Stevens*, 130 S. Ct. 1577, 1592 (2010) (holding that an act criminalizing the creation, sale, or possession of depictions of animal cruelty was overbroad and therefore facially invalid under the First Amendment protection of speech); *Citizens United v. Fed. Election Comm'n*, 130 S. Ct. 876, 913 (2010) (holding that political speech may not be suppressed “based on the corporate identity of the speaker”).

2. TIMOTHY ZICK, *SPEECH OUT OF DOORS: PRESERVING FIRST AMENDMENT LIBERTIES IN PUBLIC PLACES* (2009).

3. *Id.* at 5–6, 21–24.

4. RONALD J. KROTOSZYNSKI, JR., *RECLAIMING THE PETITION CLAUSE: SEDITION LIBEL, “OFFENSIVE” PROTEST, AND THE RIGHT TO PETITION THE GOVERNMENT FOR A REDRESS OF GRIEVANCES* (2012).

5. *Id.* at 14–19.

6. See generally, e.g., Ashutosh Bhagwat, *Associational Speech*, 120 YALE L.J. 978 (2011) (arguing that First Amendment rights are interrelated mechanisms that serve to advance democratic

John Inazu has jumped into this ferment with his book *Liberty's Refuge: The Forgotten Freedom of Assembly*.⁷ *Liberty's Refuge* is an excellent book with a dual agenda: one part descriptive and one part normative. The focus of the book is the right, delineated in the First Amendment, "of the people peaceably to assemble."⁸ Inazu begins by tracing the central role that the right of assembly played historically in political struggles and in public perceptions of the First Amendment, through the middle of the twentieth century.⁹ He then traces the gradual transformation of the right of assembly, explicitly listed in the text of the Constitution, into a nontextual right of "association" during the 1940s and 1950s, what he calls "the national security era,"¹⁰ as well as the narrowing of the right of association, combined with the complete abandonment of assembly as an independent right during the period beginning in the early 1960s, which he dubs "the equality era."¹¹ These chapters constitute the descriptive, historical part of *Liberty's Refuge*, and they tell a novel and fascinating story. Inazu concludes, however, normatively, by making the case for the revival of freedom of assembly as a robust, independent constitutional right that will provide substantial protection to the internal composition and dynamics of groups. He argues, referring to several Supreme Court cases, that the modern right of association fails to provide such protection and criticizes this development as inconsistent with both the history and the purposes of the First Amendment.¹²

self-government); Tabatha Abu El-Haj, *Changing the People: Legal Regulation and American Democracy*, 86 N.Y.U. L. REV. 1 (2011) [hereinafter El-Haj, *Changing the People*] (describing the extensive role of assembly and association in nineteenth-century elections and politics); Tabatha Abu El-Haj, *The Neglected Right of Assembly*, 56 UCLA L. REV. 543 (2009) [hereinafter El-Haj, *Neglected Right*] (characterizing public demonstrations as historically being integral to American democracy and describing the narrowing of the right of assembly today); John D. Inazu, *The Forgotten Freedom of Assembly*, 84 TUL. L. REV. 565 (2010) (discussing the importance of the right to freedom of assembly to democracy through a historical account of the right); John D. Inazu, *The Strange Origins of the Constitutional Right of Association*, 77 TENN. L. REV. 485 (2010) (describing the underpinnings of the right of association and its relationship to basic notions of democracy). This recent scholarly explosion builds on earlier work examining association, from both a legal and social science perspective. See generally MARK E. WARREN, *DEMOCRACY AND ASSOCIATION* (2001) (examining the interplay between associational life and democracy); FREEDOM OF ASSOCIATION (Amy Gutmann ed., 1998) (highlighting the individual and civic values of associational freedom in liberal democracies); Richard A. Epstein, *The Constitutional Perils of Moderation: The Case of the Boy Scouts*, 74 S. CAL. L. REV. 119 (2000) (discussing the balance between freedom of association and nondiscrimination in response to a case holding that the Boy Scouts had the right to dismiss a homosexual scout leader under the freedom of association); Jason Mazzone, *Freedom's Associations*, 77 WASH. L. REV. 639 (2002) (describing how freedom of association promotes popular sovereignty); Katherine A. Moerke & David W. Selden, *Associations Are People Too*, 85 MINN. L. REV. 1475 (2001) (describing essays that address the limits on freedom of association and the relationship of the government with religious associations).

7. JOHN D. INAZU, *LIBERTY'S REFUGE: THE FORGOTTEN FREEDOM OF ASSEMBLY* (2012).

8. U.S. CONST. amend. I.

9. INAZU, *supra* note 7, ch. 2.

10. *Id.* ch. 3.

11. *Id.* ch. 4.

12. *Id.* at 144–49.

Finally, Inazu concludes by setting forth a “theory of assembly,” which he argues would restore the freedom of assembly to its rightful place.¹³

There is much to admire in *Liberty’s Refuge*. The history that Inazu recounts, and the story of doctrinal transformation that he tells, are fascinating and well worth the read. In addition, Inazu sets forth a compelling argument that the modern association right has failed in its primary purpose of protecting the group autonomy that must exist for effective democratic self-governance. I agree with much of what Inazu has to say in this regard. In Parts I and II of this Review I will summarize Inazu’s thesis in more detail, pointing to its strengths as well as highlighting a few areas where I disagree. In Part III, I turn to another issue, which I believe is raised by aspects of Inazu’s argument though not particularly explored, which is the relationship between the freedom of assembly and *other* provisions of the First Amendment. In particular, I look at the problem of religious groups and their role as “associations” or “assemblies” protected by the First Amendment. I ask whether the religious character of a group has any implications for the types of protection it receives and what the interplay might be between the assembly and association rights, and the Religion Clauses of the First Amendment, in addressing this question. The relationship between the association right and the Religion Clauses came to the fore in the Supreme Court’s recent decision in *Hosanna-Tabor Evangelical Lutheran Church & School v. EEOC*,¹⁴ but has not been much explored in the literature. In these brief pages, I hope to begin that conversation.

I. The Gradual Demise of Assembly

At the heart of *Liberty’s Refuge* lies a historical narrative. In these chapters, John Inazu recounts the central role that freedom of assembly played in American politics and culture from the Revolutionary Era through the 1940s, and then describes the decline and eventual disappearance of assembly in constitutional and political discourse. This part of the book is a *tour de force*, weaving together historical, legal, political, and intellectual developments in a way that is both compelling and highly digestible even to those without a deep background in either constitutional history or political science. This historical story itself makes *Liberty’s Refuge* well worth the read.

Inazu’s story begins with the drafting history of the Assembly Clause in the First Congress in 1789.¹⁵ His description is extremely illuminating for a number of reasons. First, it leaves no doubt about the widespread agreement among the founding generation of the significance of the assembly right,

13. *Id.* ch. 5.

14. 132 S. Ct. 694 (2012).

15. INAZU, *supra* note 7, at 22–25.

despite the fact that the protection of assembly (unlike the petition right with which it is paired, on which more later) had no clear precedent in English law.¹⁶ Inazu traces this consensus to that generation's knowledge of and sympathy with the travails of the famous Quaker (and founder of Pennsylvania) William Penn in his struggles with the religious establishment of England.¹⁷ Notably, Inazu emphasizes that this history supports the proposition that the Framers understood the assembly right to fully encompass religious gatherings.¹⁸

Second, Inazu's drafting history clears up an important ambiguity about the scope of the assembly right resulting from the language of the First Amendment. The relevant text reads, "Congress shall make no law . . . abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances."¹⁹ Prominent scholars, including Jason Mazzone, have read the syntax of this closing portion of the Amendment to link assembly and petition, so that what the Constitution protects is a right of the people to assemble but only for the purpose of petitioning the government for a redress of grievances.²⁰ Inazu convincingly refutes this reading. He points out that the original proposals and drafts of what became the First Amendment stated two distinct rights: a right of the people "to assemble and consult for their common good," and a right to petition for a redress of grievances.²¹ The language of the "common good" was eventually dropped, but not in order to narrow the assembly right or link it to petitioning; instead, it was dropped to ensure that the reference to the common good was not invoked to try and narrow the range of protected assemblies.²² In short, Inazu argues, the history of the Assembly Clause reveals a desire on the part of the Framers to protect a right that is fundamental and extremely broad in scope.²³

From drafting history, Inazu proceeds to a broad summary of the role that the assembly right played in American political history in the century and a half following the First Amendment's ratification in 1791. The history is a fascinating one, rich and eye-opening. It encompasses such seminal moments as the debate over the Democratic-Republican Societies of the 1790s,²⁴ the use of public meetings as a form of democratic activism in the

16. James Gray Pope, *Republican Moments: The Role of Direct Popular Power in the American Constitutional Order*, 139 U. PA. L. REV. 287, 330 & n.185 (1990).

17. INAZU, *supra* note 7, at 24–25.

18. *Id.* at 25.

19. U.S. CONST. amend. I.

20. Mazzone, *supra* note 6, at 713–16.

21. INAZU, *supra* note 7, at 23.

22. *Id.* at 22–24.

23. *See id.* at 25 ("The text handed down to us thus conveys a broad notion of assembly in two ways. First, it does not limit the purposes of assembly to the common good. . . . Second, it does not limit assembly to the purposes of petitioning the government.").

24. *Id.* at 26–29.

Jacksonian era,²⁵ the efforts of southern states to suppress assemblies of slaves and free blacks throughout the antebellum period,²⁶ and the embracing of public assemblies in the North during this period by both the abolitionist and burgeoning women's rights movements.²⁷ Moreover, the right of assembly continued to play a central role in social movements well into the twentieth century, including the suffrage movement, the Civil Rights movement, and (most importantly) the radical labor movement epitomized by the Industrial Workers of the World (IWW).²⁸ The story Inazu tells about the importance of public assemblies to American politics throughout this period is, as I said, an engrossing one, and one which opens up a whole new perspective on the nature of American democracy before World War II inaugurated the modern era of suburbanization, disaffection, and national interest groups. If there is any criticism to be made of this part of Inazu's story, it is that it is incomplete. Because Inazu's primary focus (as we shall see) is on the postwar era and the decline of assembly, he fails to explore in depth a number of other episodes during the pre-modern era where associations and assemblies played an important part in political developments.²⁹ But this is a minor point—on the whole, Inazu successfully conveys the cultural significance of assembly in American democracy up to World War I, and his narrative sets the stage nicely for the heart of his story.

That story begins to take off when the Supreme Court enters the stage in the Red Scare prosecutions of the 1920s.³⁰ As Inazu notes, the interwar period was an odd one for the right of assembly. On the one hand, scholarly and political defenses of the right of assembly continued and if anything increased.³¹ On the other hand, the actual right of assembly was subject to unprecedented restrictions as part of, first, the federal government's efforts to silence critics of American involvement in World War I, and then, second, Red Scare suppression of communist movements.³² And throughout this period the Supreme Court consistently failed to provide any meaningful protection to dissident groups. Indeed, as Inazu discusses, in the seminal

25. *Id.* at 29–31.

26. *Id.* at 30–33.

27. *Id.* at 33–35.

28. *Id.* at 44–48.

29. See, e.g., El-Haj, *Neglected Right*, *supra* note 6, at 554–55 (discussing street meetings in the early Republic); *id.* at 561–69 (describing the liberal legal regime governing public assembly through most of the nineteenth century); Mazzone, *supra* note 6, at 642–44 (discussing women's clubs in nineteenth-century America); see also El-Haj, *Changing the People*, *supra* note 6, at 40–51 (highlighting the wide variety of festive street politics that persisted well into the nineteenth century).

30. See INAZU, *supra* note 7, at 50 (quoting Justice Brandeis's famous concurring opinion in *Whitney v. California*, 274 U.S. 357, 375 (1927) (Brandeis, J., concurring)).

31. See *id.* at 49 (noting that libertarian interpretations of the First Amendment and political references to free speech and assembly increased during the interwar years).

32. *Id.* at 49–50.

case of *Whitney v. California*,³³ a majority of the Court opined that Anita Whitney's decision to assemble with the Communist Party was *more* dangerous and less worthy of protection than the speech of individuals.³⁴ Justice Brandeis's seminal separate opinion, joined by Justice Holmes, did provide robust protection for free speech *and* assembly rights,³⁵ but it received only two votes out of nine.³⁶

Whitney v. California probably represents the nadir of First Amendment rights in the Supreme Court and in the nation as a whole. As a consequence of the election of Franklin Delano Roosevelt as President in 1932 and the enactment of his New Deal by a transformed Congress, the political tone of the country changed dramatically in the 1930s (these changes were themselves, of course, a product of the social upheaval triggered by the Great Depression).³⁷ Political support for assembly rights, especially for labor organizers, expanded greatly in this period.³⁸ And in 1937, in *De Jonge v. Oregon*,³⁹ a majority of the Supreme Court wholeheartedly embraced the idea of extending assembly rights even to those meeting under the auspices of the Communist Party.⁴⁰ The Court confirmed this view soon thereafter in *Herndon v. Lowry*,⁴¹ and most significantly, in 1939 a plurality of the Court endorsed the idea that the people have a right to assemble even on publicly owned land such as streets and parks.⁴² The public rhetoric of this period, some of which was triggered by the *Hague v. CIO*⁴³ litigation, saw the freedom of assembly enshrined in popular culture as one of the "Four Freedoms" underlying American democracy, co-equal with religion, speech, and the press.⁴⁴ As late as 1945, the Supreme Court was still according vigorous protection to the freedom of assembly, that time in the labor context.⁴⁵ Freedom of assembly, it would seem, had fully and finally taken its place at the center of our political liberties.

33. 274 U.S. 357 (1927).

34. *Id.* at 372.

35. *Id.* (Brandeis, J., concurring); see also Bhagwat, *supra* note 6, at 983–84 (noting the central role that assembly and association rights played in the *Whitney* case even though it is generally cited as a case about free speech).

36. *Whitney*, 274 U.S. at 372 (Brandeis, J., concurring).

37. See INAZU, *supra* note 7, at 51–52 (discussing the changes in political and labor rhetoric concerning assembly during the 1930s).

38. *Id.*

39. 299 U.S. 353 (1937).

40. *Id.* at 364–66.

41. 301 U.S. 242, 263–64 (1937).

42. *Hague v. CIO*, 307 U.S. 496, 515–16 (1939).

43. 307 U.S. 496 (1939).

44. INAZU, *supra* note 7, at 54–58.

45. *Thomas v. Collins*, 323 U.S. 516, 539–40 (1945) (finding that a Texas statute requiring a union official to obtain an organizer's card as a condition precedent to union activity is an unconstitutional restraint upon petitioner's rights of free speech and free assembly).

As it happens, things turned out otherwise. Within little more than a decade, freedom of assembly as a separate right was in decline, and within forty years, it had largely been interred. Telling the story of how this happened, and tying these legal developments to the larger political and intellectual history of the postwar era, constitutes the core of *Liberty's Refuge* and Inazu's most original contribution to our understanding of the First Amendment.

What happened to the freedom of assembly? In broad terms, Inazu argues, what happened was that assembly was “swept within the Court’s free speech doctrine.”⁴⁶ The specific path by which this occurred, however, has much to do with the rise of another, nontextual constitutional right: the right of association. As Inazu notes, the rise of the associational right in the Supreme Court in the 1950s is closely tied to two developments: McCarthyite persecution of communists and Southern persecution of civil rights activists.⁴⁷ It was in reviewing various legislative and executive attacks on communists that the Court first began to refer to a “right of association” implicit in the Constitution, albeit in the early days generally to reject the right.⁴⁸ But by 1957 the Court had relied on an association right in at least two cases to place limits on the power of the federal and state governments to punish mere affiliation with the Communist Party.⁴⁹ In discussing the McCarthy-era cases, Inazu makes much of what he sees as a doctrinal division among the Justices, between those (notably Justices Douglas and Black, but also Justice Brennan and Chief Justice Warren) who favored an *incorporation* approach, which rooted the associational right in the First Amendment as incorporated against the states in the Fourteenth Amendment, and those (notably Justices Frankfurter and Harlan) who favored a *liberty* approach, which rested on the Fourteenth Amendment alone with no particular reference to the First.⁵⁰ Inazu’s view seems to be that the association right would have been more secure if it had firmly been linked to the First Amendment. In light of later developments, I am somewhat unconvinced of the significance of this now largely defunct doctrinal division and find this part of Inazu’s doctrinal story therefore less convincing. But in any event, the main point is that the McCarthy-era cases set the stage for the

46. INAZU, *supra* note 7, at 63.

47. *See id.* at 64 (noting that the “primary political factor” in the rise of the associational right was “the historical coincidence of the Second Red Scare and the Civil Rights Movement”).

48. *Id.* at 65–73.

49. *Sweezy v. New Hampshire*, 354 U.S. 234, 249–50 (1957) (holding that placing a professor in contempt for refusing to answer questions regarding his knowledge of the Progressive Party constitutes an unconstitutional abridgment of his right to associate with others); *Wieman v. Updegraff*, 344 U.S. 183, 191–92 (1952) (holding that a statute requiring certain state employees to take an oath regarding their membership in or affiliation with certain proscribed organizations was unconstitutional).

50. INAZU, *supra* note 7, at 71–77.

next key step in the Court's jurisprudence in this area: its seminal 1958 decision in *NAACP v. Alabama ex rel. Patterson*.⁵¹

The issue in the *Patterson* case was whether the State of Alabama could require the NAACP—the preeminent civil rights organization in the nation—to disclose its membership lists,⁵² despite the fact that public disclosure of NAACP membership would undoubtedly have subjected members to economic and even physical retaliation. The Supreme Court unanimously held that it could not, because mandated disclosure violated NAACP members' "right to freedom of association."⁵³ Importantly, as Inazu notes, the majority opinion (by Justice Harlan) begins by citing the *De Jonge* and *Thomas* cases, and giving a nod towards freedom of assembly.⁵⁴ The opinion then proceeds, however, to rest primarily on a right of "association," a word that does not appear in the Constitution.⁵⁵ Moreover, the opinion ends up quite ambiguous about the link between the associational right and the First Amendment, including the Assembly Clause in particular.⁵⁶ Nonetheless, the right of association had definitively arrived, and in subsequent cases involving both the NAACP and communists, the Court continued to recognize a right of association while remaining obscure about its source and scope (and continuing to favor civil rights claimants while disfavoring communist claimants).⁵⁷

By the mid-1960s, the transformation of the textual assembly right into a nontextual association right was largely complete. As Inazu acknowledges, however, this transformation need not have had significant substantive implications. There was no apparent reason to believe that "association" would prove a narrower right than assembly, and as Inazu also notes, scholars of this period, while recognizing the doctrinal developments, did not generally attribute much significance to them.⁵⁸ It is at this point that Inazu makes what to my mind is his most valuable contribution to our understanding of legal change. Inazu does so by tying doctrinal changes in the Court's jurisprudence to the broader intellectual climate, and in particular the rise to dominance in the postwar period of pluralist political theory as epitomized by the work of Robert Dahl.⁵⁹ At its heart, the pluralist vision of American society was an extremely positive and optimistic one, envisioning society as constituted by a harmonious balance among interest groups,

51. 357 U.S. 449 (1958).

52. *Id.* at 451.

53. *Id.* at 462.

54. *Id.* at 460; INAZU, *supra* note 7, at 81.

55. *See Patterson*, 357 U.S. at 460 ("It is beyond debate that freedom to engage in association for the advancement of beliefs and ideas is an inseparable aspect of the 'liberty' assured by the Due Process Clause of the Fourteenth Amendment, which embraces freedom of speech.")

56. *See id.* (recognizing "the close nexus between the freedoms of speech and assembly").

57. INAZU, *supra* note 7, at 84–93.

58. *Id.* at 94–96.

59. *Id.* at 96–114.

mediated through the democratic process.⁶⁰ Far from being dirty words, such as Madison's "factions" and modern "special interests," pluralistic interest groups were the vehicles through which citizens could meaningfully participate in politics.⁶¹ This vision seemed a natural response to state-centered fascism, but also to excessive individualism. It provided a logical intellectual foundation for the protection of associational rights, since interest groups had to be permitted to organize and exist in order to play their proper, benevolent role in society. But in the assumptions underlying pluralism lay a grave threat. As Inazu perceptively emphasizes, pluralist theory accepted the legitimacy only of groups which themselves accepted the basic premises of American democracy.⁶² Groups outside of that broad consensus had no useful role to play, and so could even be suppressed.⁶³ Inazu argues convincingly that this view "was bereft of either authority or tradition in American political thought,"⁶⁴ and certainly his earlier history of public assemblies bears out this view. In particular, the pluralist vision of groups operating within a consensus completely ignores the role that groups can play in resisting the "tyranny of the majority," in Tocqueville's words.⁶⁵ And though the influence of pluralist theory declined in response to the turbulence of the Vietnam War era, its impact on the rights of association and assembly, Inazu argues, continued.⁶⁶

These developments bring Inazu to the final chapter in his historical story (though not in *Liberty's Refuge*): what Inazu calls the "transformation of association" into a narrow and stunted right, and the concomitant abandonment of assembly as an independent right altogether. To understand the arc of Inazu's story, it is useful to begin where Inazu ends, with his *bête noire*, the Supreme Court's 2010 decision in *Christian Legal Society Chapter of the University of California, Hastings College of the Law v. Martinez (CLS)*.⁶⁷ *CLS* is a complicated case, raising issues too convoluted to fully

60. See ROBERT A. DAHL, *A PREFACE TO DEMOCRATIC THEORY* 132–33 (1956) (arguing that a foundational consensus among political participants necessarily underlies a functioning democratic system).

61. See *id.* at 137, 145–46, 150–51 (1956) (arguing that "[a] central guiding thread of American constitutional development has been the evolution of a political system in which all the active and legitimate groups in the population can make themselves heard at some crucial stage in the process of decision").

62. INAZU, *supra* note 7, at 105–06.

63. *Id.*

64. *Id.* at 106.

65. *Id.* at 114.

66. See *id.* at 116 ("[T]he largely unquestioned pluralist consensus that gave the Court its baseline for acceptable forms of association in the late 1950s and early 1960s opened the door for the egalitarianism that emerged in the 1970s and placed certain discriminatory associations beyond its contours.").

67. 130 S. Ct. 2971 (2010). Full disclosure: I was a member of the faculty at U.C. Hastings College of the Law, the defendant in this litigation, both when the events at issue occurred and during the litigation. I, therefore, of course personally know all of the individuals on the

explore here.⁶⁸ Briefly, however, the case arose when U.C. Hastings College of the Law, a public law school located in San Francisco, denied “registered student organization” status to a student organization consisting of Christian students.⁶⁹ The reason was that the organization, the Christian Legal Society or CLS, required its members and officers to sign a “Statement of Faith,” which among other things stated adherence to certain Christian doctrines and also condemned sexual activity outside of heterosexual marriage.⁷⁰ Hastings concluded that these provisions discriminated against potential members on the basis of religion and sexual orientation, and so violated a Hastings policy which required student organizations to accept “all comers”—i.e., any student who wished to join.⁷¹ The Court, by a 5–4 vote, upheld the Hastings policy.⁷² Crucially, the Court’s analysis focused almost entirely on free speech doctrine; the majority explicitly declined to analyze separately CLS’s “freedom of association” claim, concluding that it had little independent significance because, in essence, from the majority’s perspective CLS’s association rights only had significance in so far as they were linked to its speech rights.⁷³ How could this have come to pass, where a claim by a private group to control its own membership would be analyzed as a *free speech* issue, with association relegated to secondary status and the freedom of assembly not even mentioned? It is this doctrinal (and cultural) transformation that Inazu traces and seeks to explain, once again telling a compelling and complex story.

The trigger for the “transformation” of the associational right was the birth of what Inazu calls the “equality era,” with the enactment of key civil rights legislation in 1964, as well as judicial decisions in the 1960s interpreting Reconstruction-era legislation to bar private racial discrimination.⁷⁴ Until these developments, the significance of association to civil rights was to protect the autonomy of civil rights organizations such as the NAACP.⁷⁵ With the enactment of legislation banning private discrimination, however, associational rights potentially became a barrier to civil rights, if private groups could successfully invoke associational rights to resist racial integration. This problem first came to the Court in 1976 in

defendants’ side and indeed many of the plaintiffs as well. I did not, however, have any personal involvement in those events.

68. For a fuller examination of the litigation and its implications, see generally Symposium, *The Constitution on Campus: The Case of CLS v. Martinez*, 38 HASTINGS CONST. L.Q. 499 (2011).

69. *CLS*, 130 S. Ct. at 2980–81.

70. *Id.*

71. *Id.*

72. *Id.* at 2978, 2995, 2998, 3000.

73. *Id.* at 2984–86; see INAZU, *supra* note 7, at 147–48.

74. INAZU, *supra* note 7, at 120–21.

75. See, e.g., *NAACP v. Alabama ex rel. Patterson*, 357 U.S. 449, 466 (1958) (holding that the NAACP was protected under the Fourteenth Amendment to pursue its “lawful private interests privately and to associate freely with others”).

Runyon v. McCrary.⁷⁶ The primary holding in that case was that the Civil Rights Act of 1866 barred racial discrimination in admissions by a private, nonsectarian school.⁷⁷ Along the way, however, the Court also rejected an associational claim raised by the school, though on grounds that were doctrinally far from clear.⁷⁸ *Runyon* was nonetheless significant in clarifying that ideologically motivated, private discrimination could be regulated consistent with the right of association.⁷⁹

The key, next step in the evolution of association, and the foundational case for modern association analysis, is *Roberts v. United States Jaycees*.⁸⁰ At issue in *Roberts* was whether the Jaycees, a national organization dedicated to “promoting the interests of young men,”⁸¹ had a constitutional right to exclude female members, in violation of state law.⁸² The Court held (unanimously) that it did not.⁸³ In analyzing the Jaycees’ associational claim, Justice Brennan’s majority opinion draws a critical distinction between two rights of association: a right of “intimate association” protected by the Due Process Clause,⁸⁴ and a right to associate for expressive purposes (since described as a right of “expressive association”)⁸⁵ protected by the First Amendment.⁸⁶ The majority (reasonably) found no intimate-association issue because the Jaycees are not an intimate group even on the most generous definition.⁸⁷ Its rejection of expressive association, however, was more problematic. The Court held that the purpose of expressive association was solely to protect associations who advance expressive goals, and because the inclusion of women into the Jaycees would not “impede the organization’s ability to . . . disseminate its preferred views,” there was no constitutional violation.⁸⁸ In one fell swoop, the Court completed the process of converting what had been a freestanding, textual right of assembly into a nontextual and ancillary right of association for expressive purposes. It should be noted that this transition occurred even though the Court rooted this right squarely in the First Amendment (suggesting that Inazu’s concerns

76. 427 U.S. 160 (1976).

77. *Id.* at 172–74.

78. INAZU, *supra* note 7, at 123–24.

79. See *Runyon*, 427 U.S. at 176 (stating that the freedom of association protects the right of parents “to send their children to educational institutions that promote the belief that racial segregation is desirable, and that the children have an equal right to attend such institutions. But it does not follow that the *practice* of excluding racial minorities from such institutions is also protected by the same principle.”).

80. 468 U.S. 609 (1984).

81. *Id.* at 627.

82. *Id.* at 612.

83. *Id.* at 612, 631.

84. *Id.* at 617–18.

85. *Id.*; INAZU, *supra* note 7, at 135–40.

86. INAZU, *supra* note 7, at 135–40.

87. *Roberts*, 468 U.S. at 619–21.

88. *Id.* at 618, 627.

about “incorporation” versus “liberty” may be off the mark).⁸⁹ The difficulty was that instead of focusing on “assembly,” the Court focused on “speech” as the source of the associational right.

What intellectual forces produced this truncation of a formerly hallowed right? The pernicious influence of pluralism may have been the root cause, but Inazu traces the specific intellectual impetus to “The Rise of Rawlsian Liberalism.”⁹⁰ In Inazu’s view, the form of liberalism associated with John Rawls’s *Theory of Justice*, as expounded by later writers including notably Ronald Dworkin, built on pluralism by tying pluralist visions of harmony with specific commitments to equality and regard for others.⁹¹ This predisposition, Inazu suggests, naturally led lawyers and judges inculcated with the liberalism of the 1970s (including Justice Brennan) to prioritize equality principles over the autonomy of dissenting groups.⁹² I must confess that unlike Inazu’s pluralism story, which I find quite persuasive, his discussion of Rawlsian liberalism leaves me a bit cold. There is no doubt that Rawls and Dworkin represent a particular form of moderate-left thinking in the United States of the 1970s and 1980s. But were they, and especially legal thinkers like Dworkin, really shapers of opinion? Or were they merely rationalizers for a liberal consensus that was the outgrowth of the Civil Rights Movement and other social movements? Just as much of liberal jurisprudential writings from that period seem designed primarily to defend *Roe v. Wade*, one wonders if the embrace of equality over liberty was similarly designed to provide intellectual justification for a *fait accompli*—the legislative and judicial achievements of the civil rights era.

In any event, as Inazu points out, the *Roberts* reformulation of association has largely been adhered to since 1984.⁹³ The primary exception is *Boy Scouts of America v. Dale*,⁹⁴ in which the Court upheld the right of the Boy Scouts to exclude a gay assistant scoutmaster on the somewhat forced theory that inclusion of a gay assistant scoutmaster would interfere with the Boy Scouts’ ability to express a message of hostility to homosexuality, thereby violating the Scouts’ right of expressive association (the result would, of course, have been much easier to defend on a pure assembly or association theory).⁹⁵ But *CLS* retreated to some extent from that position,⁹⁶

89. See INAZU, *supra* note 7, at 74–75 (discussing the differences between the incorporation argument and the liberty argument).

90. *Id.* at 129–32.

91. See *id.* at 129 (“Pluralist political thought insisted on a consensus bounded by shared democratic values; Rawlsian liberalism presumed an ‘overlapping consensus’ in which egalitarianism rooted in an individualist ontology trumped and thus bounded difference.”).

92. See *id.* (“Like the pluralist assumptions that preceded them, the Rawlsian premises of consensus and stability pervaded political discourse and influenced the ways in which the equality era reshaped the right of association.”).

93. *Id.* at 142 (discussing *N.Y. State Club Ass’n v. City of New York*, 487 U.S. 1 (1988) and *Bd. of Dirs. of Rotary Int’l v. Rotary Club of Duarte*, 481 U.S. 537 (1987)).

94. 530 U.S. 640 (2000).

95. *Id.* at 655.

and in any event, given the confusions inherent in the expressive association doctrine, it remains far from clear what the exact scope of the *Dale* decision was and how it could be reconciled with *Roberts*. So for now, association remains a truncated right, limited to facilitating speech, and as Inazu notes, “The Court . . . has not addressed a freedom of assembly claim in thirty years.”⁹⁷

II. Inazu’s Theory of Assembly

Inazu’s historical story of doctrinal evolution ends with, as he sees it, the evisceration of any form of substantial group-autonomy rights in *CLS*. *CLS*, however, is in Inazu’s view not where the Court went truly wrong; it is instead the predictable fallout from earlier errors. The key error, Inazu argues, was the Court’s reformulation in *Roberts v. U.S. Jaycees* of the association right into dual, narrow rights of intimate and expressive association.⁹⁸ This left a gaping hole in protection of group rights. Intimate association protects small familial (and perhaps family-like) groups; and expressive association protects groups that are directed at speech (and perhaps other First Amendment activities such as petitioning the government or the exercise of religion).⁹⁹ But what about other groups, which are not familial in any meaningful sense and also not primarily expressive, but which nonetheless provide a critical space within which citizens can jointly develop their values and their capacity for self-governance? The *Roberts* reformulation, Inazu convincingly argues, leaves little or no protection for the internal autonomy of such groups, and therefore, leaves them at the mercy of tyrannical democratic majorities.¹⁰⁰

Enter assembly. The core of the normative argument in *Liberty’s Refuge* is that the time is ripe for a reinvigoration of the textual right of assembly in order to cure the deficiencies of the modern association doctrine. Inazu takes the position that interpretative theory fully supports a turn back to assembly as the key source of group rights.¹⁰¹ He also convincingly demonstrates that the history of group rights in this country fully supports a right of autonomy of dissenting, nonconformist groups,¹⁰² contrary to views of scholars such as Andrew Koppelman who argue that the “right to discriminate” recognized in *Boy Scouts v. Dale* was an historical

96. See generally *CLS*, 130 S. Ct. 2971 (2010) (deciding the case on other grounds, but noting that U.C. Hastings could condition the Christian Legal Society’s status as a registered student organization on its acceptance of persons of all religious beliefs, even though one of the Society’s purposes was to express solely Christian beliefs).

97. INAZU, *supra* note 7, at 62.

98. *Id.* at 135.

99. *Id.* at 140.

100. *Id.* at 135–41.

101. See *id.* at 5 (arguing that “[r]ecovering the vision of assembly remains an urgent task”).

102. See *id.* at 4 (arguing that the four principles of the history of assembly collectively counsel for the protection of groups “that dissent from majoritarian standards”).

aberration.¹⁰³ The Assembly Clause would, Inazu argues, protect dissident groups as well as nonexpressive social and religious groups in a way that association fails to do.¹⁰⁴

In addition to his interpretative and historical arguments, Inazu also presents a political theory of assembly, drawing upon the work of Sheldon Wolin¹⁰⁵ as a counterweight to the consensus-driven narrative of Dahlian Pluralism and Rawlsian Liberalism.¹⁰⁶ It is necessary, he argues, to protect *dissenting* and *political* assemblies, groups that reject certain consensus norms on a nonnegotiable basis, and that seek to engage in a form of politics outside of the accepted politics of state institutions.¹⁰⁷ Inazu also asserts that recognizing a vibrant assembly right will advance *expressive* goals, curing some of the shortcomings of expressive association by recognizing the variety and complexity of the ways in which groups can be expressive.¹⁰⁸ As I have argued elsewhere, I find this last argument less convincing.¹⁰⁹ It seems to me that one of the great advantages of supplementing “expressive association” with the textual right of assembly is precisely that it rejects the pernicious idea that groups deserve protection only to the extent that they are expressive. Even nonexpressive social and religious groups contribute to the goals of the First Amendment by protecting and advancing democratic self-governance in critical ways,¹¹⁰ and so lie fully within the coverage of the First Amendment. To emphasize the expressive nature of assemblies might undermine this critical point. At bottom, however, this is a relatively minor point of disagreement. There is no doubt that Inazu fully accepts the view that nonexpressive groups are entitled to constitutional protection,¹¹¹ and so

103. *Id.* at 162–66 (discussing ANDREW KOPPELMAN WITH TOBIAS BARRINGTON WOLFF, A RIGHT TO DISCRIMINATE? HOW THE CASE OF *BOY SCOUTS OF AMERICA V. DALE* WARPED THE LAW OF FREE ASSOCIATION (2009)).

104. *Id.* at 150–53.

105. *Id.* at 153–56 (discussing SHELDON S. WOLIN, POLITICS AND VISION: CONTINUITY AND INNOVATION IN WESTERN POLITICAL THOUGHT (2004)).

106. *Id.*

107. *Id.* at 156–60.

108. *Id.* at 160–62.

109. See Ashutosh Bhagwat, *Liberty's Refuge, or the Refuge of Scoundrels?: The Limits of the Right of Assembly*, 89 WASH. U. L. REV. 1381, 1383–84 (2012) (arguing that Inazu's emphasis on the expressive nature of assembly undermines the argument that the Assembly Clause is an “independent and co-equal” First Amendment right, and that assembly “should be protected not because it is expressive, but because it independently advances the goals of the First Amendment”). For Professor Inazu's response to my critique, see John D. Inazu, *Factions for the Rest of Us*, 89 WASH. U. L. REV. 1435, 1436 (2012) (replying that the emphasis on the inherent expressiveness of assembly was intended as a critique of the doctrinal distinction between expressive and nonexpressive associations and reaffirming that assembly is valuable because it facilitates “dissent, self-governance, and the informal relationships that make politics possible”).

110. For a more detailed discussion of the link between groups and democratic self-governance, see Bhagwat, *supra* note 6, at 991–99.

111. See Inazu, *supra* note 109, at 1436 (“[T]he expressive potential of a group is not the reason that we value assembly.”).

the space between his views and mine are primarily a question of rhetoric and emphasis.

Inazu concludes by setting forth in full-blown form his theory of assembly. He defines assembly as “a presumptive right of individuals to form and participate in peaceable, noncommercial groups.”¹¹² By adopting this broad view, Inazu seeks to avoid the limitations of the *Roberts* approach, and to affirm that the assembly right is a stand-alone right of group autonomy and not merely a handmaiden to other First Amendment liberties. But, inevitably, Inazu also is forced to recognize limits on the scope of assembly. The definition itself restricts protection to *peaceable* groups, a limitation which he acknowledges raises difficult boundary questions,¹¹³ and excludes commercial groups.¹¹⁴ Finally, and most significantly, Inazu excludes from protection groups which “prosper[] under monopolistic or near-monopolistic conditions.”¹¹⁵ As examples of such groups, he cites the famous Jaybird Association, which was the subject of the *Terry v. Adams*¹¹⁶ litigation, and a hypothetical student group “providing exclusive access to elite legal jobs.”¹¹⁷ Inazu urges a “contextual analysis,” focused on “how power operates on the ground,” in applying this exception,¹¹⁸ but ultimately he is clear that it is a narrow one. Inazu is a bit unclear about exactly why he would deny coverage to such “monopolistic” groups, but presumably the reason is that the social harm caused by the exclusion from such groups of individuals subject to discrimination outweighs the value of protecting the assembly right in such contexts.

All of the above points to some important questions raised but not answered by *Liberty's Refuge*. There is no question in my mind that Inazu's arguments do a great service in pointing out how ahistorical and theoretically problematic the *Roberts* reformulation and narrowing of group rights really was. I am also willing to accept Inazu's premise that this damage can be undone by resurrecting the textual assembly right from its premature demise—though one is left uncertain at the end of *Liberty's Refuge* why the same goals might not be accomplished by a broadening of the association right. Perhaps the answer lies in some combination of the fact that the doctrinal damage done by *Roberts* is at this point too entrenched to be reversed, and that the textual roots of assembly makes it a better repository for a stand-alone right of group autonomy.

112. INAZU, *supra* note 7, at 166.

113. *Id.* at 167. For a discussion of the ambiguities surrounding the exclusion of violent assemblies, see Bhagwat, *supra* note 109, at 1389–92.

114. INAZU, *supra* note 7, at 167.

115. *Id.* at 166.

116. 345 U.S. 461 (1953).

117. INAZU, *supra* note 7, at 172.

118. *Id.*

The unanswered questions raised by *Liberty's Refuge* concern Inazu's concept of dissenting political assemblies. Dissent is at the heart of the concept of assembly endorsed by *Liberty's Refuge*. And for Inazu, the quintessential example of a dissenting assembly is the Christian Legal Society, denied the right to define its own membership in *CLS*. But why is *CLS* a "dissenting" group? Certainly, in the doubly liberal environment of a law school located in San Francisco, a conservative Christian group opposed to homosexuality qualifies as "dissenting," in the sense of being out of the mainstream politically and socially. But similar groups, located in many, many social contexts in many, many parts of this country would fit comfortably in the mainstream, and it is LGBT groups that would be "dissenting." In those contexts, is defending the right of groups such as the Boy Scouts, unless they are "monopolistic," to exclude homosexuals truly advancing "dissent"? Similarly, consider the United States Jaycees. The Jaycees are a highly regarded, national group with a great deal of prestige. Is such a group, or the Rotary International (a defendant in similar litigation), truly a "dissenting" group, requiring judicial protection of their right to exclude women against a hostile, tyrannical majority? There is something distinctly odd about this picture.

This raises an even more basic question: *why* should we favor group autonomy even at the expense of other social values such as equality and social peace? That we have historically done so is a good starting point, but it does not provide a fully satisfactory answer, especially in light of the fact that we as a society have quite consciously and properly distanced ourselves from many of the exclusionary practices of the past. Inazu argues that the reason is to ensure that our society retains a true pluralism, rooted in differences in fundamental values.¹¹⁹ Moreover, despite the capaciousness of Inazu's theory and his commitment to group autonomy (which I do not for a moment question), the actual instances of conflict that he discusses in recent years overwhelmingly involve *religious* groups and values. I close my discussion by briefly considering why that might be so, and what a particularized focus on religious assemblies teaches us about assembly, association, and the role of the state. Lurking in the background here are two provisions of the First Amendment, the Establishment and Free Exercise Clauses, which get very little notice in *Liberty's Refuge*, but which I suggest may deserve more attention.

III. The Elephant in the Room: Religious Assemblies and the Religion Clauses

At the heart of *Liberty's Refuge* is a normative claim that for reasons both historical and theoretical it is important to grant constitutional

119. See *id.* at 11 (arguing against the political theory of consensus liberalism underwriting weakened group autonomy and resulting in the loss of meaningful pluralism).

protection to the internal autonomy of dissenting, nonconformist groups. Inazu is also clear about the sorts of groups that he has uppermost on his mind. One such group, as noted earlier, is the Christian Legal Society. Another group Inazu mentions is the Chi Iota Colony of the Alpha Epsilon Pi (AEPi) fraternity.¹²⁰ AEPi is a national social fraternity for Jewish college men, and the Chi Iota Colony was seeking to become an AEPi chapter at the College of Staten Island.¹²¹ The college denied Chi Iota's request to be granted official recognition (and access to funds) because Chi Iota refused to admit women.¹²² Chi Iota sued, but was unsuccessful because both its intimate and expressive association claims were weak.¹²³ Finally, Inazu clearly believes that the Supreme Court was correct in *Boy Scouts v. Dale* in upholding the Boy Scouts' right to exclude a gay assistant scoutmaster.

What do these groups have in common? On its face, it is the desire to exclude others. But that cannot be the end of it. Inazu, for example, seems quite sympathetic with the Court's decision in *Runyon* rejecting a private school's right to racially discriminate in admitting students.¹²⁴ Instead, CLS, the Boy Scouts, and, to a lesser degree, Chi Iota appear sympathetic because of the ideological, and in particular *religious* and *moral*, underpinnings of their actions. CLS is of course an explicitly religious organization, and the Boy Scouts themselves, even though not sectarian, clearly root their beliefs and actions in religious values—which is why the Scouts exclude not only homosexuals, but also atheists.¹²⁵ Chi Iota is the least obviously religious of these groups, but even its Jewish identity has a clear religious element—though Inazu tellingly suggests that Chi Iota's claim may well have been hurt by the fact that “[a]lthough [Chi Iota's] Jewish roots suggest religious freedom interests, most of its members were nonpracticing Jews.”¹²⁶ The plain implication is that an explicitly religious group's claims would (or should) be even more persuasive than Chi Iota's.

Nor is Inazu's concern with religiously oriented groups idiosyncratic. There was a time, in the McCarthy and Civil Rights eras, when associational rights were claimed primarily by nonconformist political groups such as the Communist Party, the NAACP, and other civil rights organizations. Later, during the 1970s and 1980s, associational issues arose in the context of eliminating race and gender segregation. In today's world, however, the battles over association, assembly, and group autonomy focus primarily on

120. *Id.* at 144–45.

121. *Chi Iota Colony of Alpha Epsilon Pi Fraternity v. City Univ. of N.Y.*, 502 F.3d 136, 142 (2d Cir. 2007).

122. *Id.*

123. *Id.* at 149 & n.2.

124. INAZU, *supra* note 7, at 123.

125. *See Barnes-Wallace v. City of San Diego*, 530 F.3d 776, 780 (9th Cir. 2008) (explaining that the Boy Scouts “maintain that agnosticism, atheism, and homosexuality are inconsistent with their goals and with the obligations of their members”).

126. INAZU, *supra* note 7, at 145.

religion. One line of cases pits religiously oriented groups seeking to exclude others on the basis of either religion or sexual orientation against state nondiscrimination policies.¹²⁷ In another line of cases, disputes have arisen over attempts by religious groups to meet—i.e., to assemble—on public property¹²⁸ or to obtain access to public benefits.¹²⁹

127. *See, e.g.,* *CLS*, 130 S. Ct. 2971, 2978 (2010) (pitting a law school chapter of the Christian Legal Society with membership requiring a statement of faith against the school's all-comers nondiscrimination policy); *Boy Scouts of Am. v. Dale*, 530 U.S. 640, 644 (2000) (placing the Boy Scouts of America, which maintained a policy against homosexuality, agnosticism, and atheism, against New Jersey's public accommodations law); *Truth v. Kent Sch. Dist.*, 542 F.3d 634, 637–41 (9th Cir. 2008) (pitting a school Bible Club seeking to exclude nonbelievers against school district's nondiscrimination policy), *overruled on other grounds by* *L.A. Cnty. v. Humphries*, 131 S. Ct. 447 (2010); *Christian Legal Soc'y v. Walker*, 453 F.3d 853, 857–58 (7th Cir. 2006) (pitting a Christian student organization seeking to exclude homosexuals against a university nondiscrimination policy); *Hsu ex rel. Hsu v. Roslyn Union Free Sch. Dist. No. 3*, 85 F.3d 839, 847–48 (2d Cir. 1996) (placing a high school Bible club seeking to exclude nonbelievers against the school's generally applicable nondiscrimination policy).

128. *See, e.g.,* *Good News Club v. Milford Cent. Sch.*, 533 U.S. 98, 102, 107 (2001) (finding a school's exclusion of a Christian children's club from meeting after hours at school, based on its religious nature, to be unconstitutional viewpoint discrimination); *Lamb's Chapel v. Ctr. Moriches Union Free Sch. Dist.*, 508 U.S. 384, 393–95 (1993) (finding a school district violated the First Amendment by denying a church access to school premises to exhibit film series on family and child-rearing issues); *Widmar v. Vincent*, 454 U.S. 263, 264–67 (1981) (finding a public university could not prohibit a registered religious group from use of university facilities which were generally available for use by other registered groups); *Bronx Household of Faith v. Bd. of Educ. of N.Y.*, 650 F.3d 30, 32–33 (2d Cir. 2011) (reversing an injunction against the city board of education and school district, which had excluded a church from religious worship practices on school grounds); *Faith Ctr. Church Evangelistic Ministries v. Glover*, 480 F.3d 891, 902, 918–19 (9th Cir. 2007) (reversing a preliminary injunction against a county excluding a religious nonprofit organization from holding worship services in the public library meeting room); *Donovan ex rel. Donovan v. Punxsutawney Area Sch. Bd.*, 336 F.3d 211, 214 (3d Cir. 2003) (finding a public high school's denial of permission for a religious club to meet on school premises during student activity period constituted viewpoint discrimination in violation of First Amendment); *Fairfax Covenant Church v. Fairfax Cnty. Sch. Bd.*, 17 F.3d 703, 704 (4th Cir. 1994) (finding a regulation allowing a school to charge churches an escalating rate for use of school facilities discriminated against religious speech in violation of the First Amendment); *Grace Bible Fellowship v. Me. Sch. Admin. Dist. No. 5*, 941 F.2d 45, 47–48 (1st Cir. 1991) (holding that by allowing other organizations to use facilities for expressive activities, the school district created a public forum from which it could not bar a religious organization); *Gregoire v. Centennial Sch. Dist.*, 907 F.2d 1366, 1369 (3d Cir. 1990) (allowing a religious group to conduct activities, not limited to those of a secular nature, in a high school auditorium).

129. *See, e.g.,* *Rosenberger v. Rector & Visitors of Univ. of Va.*, 515 U.S. 819, 822–23, 845–46 (1995) (holding that a state university's refusal to fund the printing of religious student publications while funding nonreligious publications violated the right to free speech); *Everson v. Bd. of Educ.*, 330 U.S. 1, 17 (1947) (holding that taxpayer-funded reimbursements for parochial school students' bus fares do not violate the First Amendment); *Badger Catholic, Inc. v. Walsh*, 620 F.3d 775, 776–78 (7th Cir. 2010) (holding that a public university's funding of student-group programs where prayer sessions occur does not violate the Establishment Clause); *Rocky Mountain Christian Church v. Bd. of Cnty. Comm'rs*, 612 F. Supp. 2d 1163, 1180 (D. Colo. 2009) (holding that the equal-terms provision of the Religious Land Use and Institutionalized Persons Act, as applied, does not violate the Establishment Clause); *Every Nation Campus Ministries at San Diego State Univ. v. Achtenberg*, 597 F. Supp. 2d 1075, 1078–79 (S.D. Cal. 2009) (holding that a public university's refusal to formally recognize Christian student groups that refuse to comply with the nondiscrimination policy does not violate the groups' First Amendment rights); *Roman Catholic Found., UW-Madison, Inc. v. Regents of the Univ. of Wis. Sys.*, 578 F. Supp. 2d 1121, 1133 (W.D.

And even outside of the courtroom, the most prominent modern examples of groups claiming autonomy and the right to choose their membership selectively also tend to involve religious groups. It is, for example, inconceivable (and of course illegal) for any significant commercial entity to exclude women from leadership positions, and even most noncommercial entities appear to have admitted women since the battles of the 1980s.¹³⁰ Yet it remains true that major religious sects, including the Catholic Church,¹³¹ Orthodox Jewish congregations,¹³² and the Mormon Church,¹³³ continue to exclude women from the clergy. In short, in the modern world, the epitome of the “dissenting, political” assembly that Inazu seeks to defend is the religious assembly.

It is also worth noting that the linkage between assembly—or for that matter speech—rights and religion is not merely a modern one. In *Liberty’s Refuge*, Inazu himself points to the importance of the tradition of religious nonconformity associated with William Penn and Roger Williams in helping to develop American ideas of free expression and assembly.¹³⁴ He also notes that during the actual debates in the First Congress over the Assembly Clause, a specific reference was made to the English prosecution of William Penn for holding a *religious* assembly of Quakers which did not comply with the strictures of the established Church of England.¹³⁵ Elsewhere, Inazu has more explicitly explained and explored the religious roots of the very term “assembly,” noting that going back to the early Christian era the term (and its Greek predecessor *ekklesia*) always had political *and* religious connotations.¹³⁶ Similarly, Akhil Amar has noted that during the antebellum era among abolitionists “the core right of assembly at issue seems to be the right of blacks ‘to assemble peaceably on the Sabbath for the worship of [the]

Wis. 2008) (holding that the Establishment Clause does not compel a public university to categorically refuse funding for a student group’s “worship, proselytizing or sectarian religious instruction”).

130. Including in 1991 the epitome of the “Old Boys Club,” the Skull and Bones secret society at Yale, though not without a fight. Dennis Hevesi, *Shh! Yale’s Skull and Bones Admits Women*, N.Y. TIMES, Oct. 26, 1991, at 21; see also *Yale Alumni Block Women in Secret Club*, N.Y. TIMES, Sept. 6, 1991, at B2 (reproducing an AP report that the Skull and Bones society “obtained a court order temporarily blocking the all-male club from admitting women”).

131. Ryan W. Jaziri, *Fixing a Crack in the Wall of Separation: Why the Religion Clauses Preclude Adjudication of Sexual Harassment Claims Brought by Ministers*, 45 NEW ENG. L. REV. 719, 721 n.17 (2011) (citing MARCI A. HAMILTON, *GOD VS. THE GAVEL: RELIGION AND THE RULE OF LAW* 190 (2005)).

132. Ilana S. Cristofar, *Blood, Water and the Impure Woman: Can Jewish Women Reconcile Between Ancient Law and Modern Feminism?*, 10 S. CAL. REV. L. & WOMEN’S STUD. 451, 462 (2001).

133. Elisabeth S. Wendorff, *Employment Discrimination and Clergywomen: Where the Law Has Feared to Tread*, 3 S. CAL. REV. L. & WOMEN’S STUD. 135, 140 (1993).

134. INAZU, *supra* note 7, at 12–13 & 13 n.28.

135. *Id.* at 24–25.

136. John D. Inazu, *Between Liberalism and Theocracy*, 33 CAMPBELL L. REV. 591, 601 & n.44 (2011).

Creator.”¹³⁷ There is thus good precedent for the modern centrality of religious groups and religious speech in First Amendment disputes.

When one recognizes the central role that religious groups play in modern association/assembly disputes, however, a conundrum arises: why do these cases typically turn on the Speech and Assembly Clauses of the First Amendment, and the related right of association, rather than on the First Amendment provisions which expressly address religion—the Establishment and Free Exercise Clauses? One might think that these provisions, whose very purpose is to protect religious autonomy, would provide greater protection to religious groups than the generic rights of assembly or association. But that is not the case. The Christian Legal Society did in fact join a Free Exercise claim to its primary speech and association claims in the *CLS* litigation, but the Court dismissed the argument in a casual footnote, citing its decision in *Employment Division v. Smith*¹³⁸ for the proposition that because Hastings’ “all-comers” policy was a generally applicable rule that did not target religion, it raised no free exercise issues.¹³⁹ Nor is the *CLS* decision an aberration in this regard. Lower courts have also relied upon *Smith* to conclude that the Free Exercise Clause grants less protection to the associational rights of religious groups than does expressive association.¹⁴⁰

Decisions such as *CLS* would seem to suggest that the Religion Clauses play second fiddle to speech, assembly, and association claims by religious groups. The truth, however, is rather more muddled, as demonstrated by the Supreme Court’s recent, important decision in *Hosanna-Tabor Evangelical Lutheran Church & School v. EEOC*. The issue in *Hosanna-Tabor* was whether the First Amendment created a “ministerial exception” to antidiscrimination statutes, which shielded religious institutions from antidiscrimination claims brought by ministers and other employees (the litigation arose when a teacher at a religious school brought a lawsuit under the Americans with Disabilities Act).¹⁴¹ The Court held that the Religion Clauses required such an exemption.¹⁴² The government and the plaintiff argued to the Court that instead of turning to the Religion Clauses, the Court should look to the right of association as the source of any such exemption, but the Court rejected this argument as “untenable,” and indeed,

137. AKHIL REED AMAR, *THE BILL OF RIGHTS: CREATION AND RECONSTRUCTION* 245 (1998) (quoting JACOBUS TENBROEK, *EQUAL UNDER LAW* 124–25 (1965)).

138. *Emp’t Div. v. Smith*, 494 U.S. 872 (1990).

139. *CLS*, 130 S. Ct. 2971, 2995 n.27 (2010) (citing *Smith*, 494 U.S. at 878–82).

140. *Salvation Army v. Dep’t of Cmty. Affairs of N.J.*, 919 F.2d 183, 194–96 (3d Cir. 1990); *Wiley Mission v. N.J. Dep’t of Cmty. Affairs*, Civil No. 10-3024, 2011 WL 3841437, at *13 (D.N.J. Aug. 25, 2011); *Jews for Jesus, Inc. v. Port of Portland, Or.*, No. CV04695HU, 2005 WL 1109698, at *15 (D. Or. May 5, 2005).

141. *Hosanna-Tabor Evangelical Lutheran Church & Sch. v. EEOC*, 132 S. Ct. 694, 700–01, 705–06 (2012).

142. *Id.* at 705–06.

“remarkable.”¹⁴³ The difficulty with this argument, the Court said, was that it would grant religious organizations no more autonomy than secular associations, and that was inconsistent with the fact that the First Amendment, through the Religion Clauses, “gives special solicitude to the rights of religious organizations.”¹⁴⁴ In other words, the *Hosanna-Tabor* Court read the Religion Clauses as granting religious associations *greater* protection than the general association right. And again, there are lower court cases consistent with this view.¹⁴⁵

Consider the *CLS* and *Hosanna-Tabor* cases, which were decided less than two years apart. Both involved attempts by religious groups to exclude individuals—in *CLS* from membership and in *Hosanna-Tabor* from employment. In both instances, the exclusion was religiously motivated. Yet *CLS* was litigated primarily, and unsuccessfully, as a freedom of association/free speech case, while *Hosanna-Tabor* was litigated successfully as a religion case. *Hosanna-Tabor* was a unanimous decision, and while the Court divided sharply in *CLS*, not even the dissenting justices invoked the Religion Clauses as a basis for protecting *CLS*’s autonomy. This is not to say that the results in the two cases are necessarily inconsistent. *CLS* was different from *Hosanna-Tabor* in that it did not involve a flat attempt by the State to regulate a religious entity. It involved only denial of official recognition and benefits (including funding and use of government property), and everyone seemed to acknowledge that the government could not have simply required *CLS* to admit members it wished to exclude. But the question does remain why in one case the Religion Clauses provided powerful protection for religious autonomy, while in the other they were brushed off as irrelevant. And more generally, the question raised by these cases is whether the religious nature of an association matters in determining the level of constitutional protection to which it is entitled.

It should be noted, moreover, that the uncertain lines between the Religion Clauses and the rest of the First Amendment are not limited to the associational context. In a separate line of modern cases, the Supreme Court has analyzed exclusion of religious groups from public property or public benefits as a species of viewpoint discrimination, violating the Free Speech Clause.¹⁴⁶ As my colleagues Vik Amar and Alan Brownstein have pointed out, however, this move and the concomitant failure of the Court to analyze these cases under the Religion Clauses is highly problematic and raises nontrivial questions about the general viability of laws banning

143. *Id.* at 706.

144. *Id.*

145. See, e.g., *Irshad Learning Ctr. v. Cnty. of DuPage*, 804 F. Supp. 2d 697, 717–18 (N.D. Ill. 2011) (holding that the allegations adequately alleged that the county violated free exercise rights under the First Amendment and the Illinois Constitution).

146. *Good News Club v. Milford Cent. Sch.*, 533 U.S. 98, 107–12 (2001); *Rosenberger v. Rector & Visitors of Univ. of Va.*, 515 U.S. 819, 828–46 (1995); *Lamb’s Chapel v. Ctr. Moriches Union Free Sch. Dist.*, 508 U.S. 384, 392–96 (1993).

discrimination on the basis of religion.¹⁴⁷ The truth is that while the Court pays occasional attention to the relationship between speech and religion, at a systematic level it seems blissfully unaware of the complexities here.

A full answer to these difficult questions is far beyond the scope of this Review, even limited to the problem of association. Any exploration, however, must begin with the question that, as I noted earlier, is largely elided in *Liberty's Refuge*: Why the First Amendment protects group autonomy, and for that matter, religious freedom. Part of the answer, Inazu suggests, lies in the need to protect dissent, including moral and religious dissent. I think, however, that this can only be part of the answer. Another part of the answer must lie in distrust of the state. The Constitution is, after all, at heart a structural document, and the limitations it places on state power, including those in the Bill of Rights, reflect structural concerns about misuse of that power. And those concerns are in turn rooted in the need to ensure that the sovereign people remain in charge of their government.¹⁴⁸ In other words, dissent is valuable precisely because it is an essential component of popular sovereignty and democratic self-governance. The scope of constitutional protection for assemblies and associations turns not on general principles regarding the proper role of private groups in our society, but rather on the appropriate relationship between such groups and the state.

Here, I think, is where the limits of freedom of association, or as Inazu would have it the Assembly Clause, become apparent. If the issue we are exploring is the proper relationship between religious groups and the state, those bodies of law are unlikely to provide useful answers because they do not distinguish between religious and other groups. But religion *is* different, a point that the Constitution recognizes in the Religion Clauses, especially the Establishment Clause. Exactly how religious assemblies differ from secular ones, however, is far from easy to pin down. Perhaps *Hosanna-Tabor* is correct in suggesting that government interference in the internal structure of religious groups is more constitutionally problematic than interference in secular groups. But on the flip side, it is also true that governmental benefits flowing to religious groups raise difficult constitutional questions that benefits to secular groups do not. This is not to say that the inclusion of a group like CLS in a general, neutral scheme of governmental benefits such as the Hastings Registered Student Organization program would violate the Establishment Clause—under current doctrine it

147. Alan Brownstein & Vikram Amar, *Reviewing Associational Freedom Claims in a Limited Public Forum: An Extension of the Distinction Between Debate-Dampening and Debate-Distorting State Action*, 38 HASTINGS CONST. L.Q. 505, 537–39 (2011).

148. For more detailed examinations of these themes, see generally AMAR, *supra* note 137 (chronicling the changing interpretation of the Bill of Rights throughout history) and ASHUTOSH BHAGWAT, *THE MYTH OF RIGHTS: THE PURPOSES AND LIMITS OF CONSTITUTIONAL RIGHTS* (2010) (arguing that the primary purpose of constitutional rights is to restrict governmental power, thereby maintaining the proper structural balance between individuals and the state).

almost certainly would not.¹⁴⁹ But such benefits can raise difficult problems, especially if they come with conditions. Consider the fact that governments regularly condition benefits or funds on recipients agreeing to restrict their conduct in particular ways, including commonly surrendering the right to discriminate.¹⁵⁰ No one seems to seriously believe that such conditions generally raise constitutional concerns. But what about when the recipient is a religious organization? I would posit that at a minimum we should be concerned about such state intrusion into the inner workings of religious groups, even if we would not be concerned about secular groups, and that the source of such concerns is not the Assembly Clause of the First Amendment but the Religion Clauses.

149. See *Zelman v. Simmons-Harris*, 536 U.S. 639, 652–53 (2002) (holding that a program, which provides tuition aid for students to attend participating public or private schools of their choosing, does not offend the Establishment Clause, even though governmental aid reaches some religious institutions indirectly through the program); *Rosenberger*, 515 U.S. at 845–46 (holding that a public university does not violate the Establishment Clause when it provides funding for a wide range of student organizations, even if some are religious organizations).

150. See, e.g., Education Amendments of 1972 §§ 901, 904, 20 U.S.C. §§ 1681, 1684 (2006) (barring discrimination based on sex or blindness); Age Discrimination in Employment Act of 1967, 29 U.S.C. §§ 621–634 (2006) (barring age-based discrimination); Rehabilitation Act of 1973 § 504, 29 U.S.C. § 794 (2006) (barring disability-based discrimination); Civil Rights Act of 1964 § 601, 42 U.S.C. § 2000d (2006) (barring discrimination based on “race, color, or national origin”); Americans with Disabilities Act of 1990 § 102, 42 U.S.C. § 12112 (2006) (barring disability-based discrimination in employment).

Recovering the Assembly Clause

LIBERTY'S REFUGE: THE FORGOTTEN FREEDOM OF ASSEMBLY. By John D. Inazu. New Haven, Connecticut: Yale University Press, 2012. 288 pages. \$55.00.

Reviewed by Timothy Zick*

Introduction

I recall driving to work one day several years ago and listening to a radio program on which listeners were invited to call in and test their basic knowledge of the First Amendment. The challenge was to name four of the freedoms listed in the First Amendment, or alternatively to identify the last names of four characters from the animated television show *The Simpsons*. It was a small sample, to be sure, but to both my amusement (as a commuter) and horror (as someone who teaches and writes about the First Amendment) every caller was far more successful naming *Simpsons* characters than identifying First Amendment freedoms.

As I recall, not a single caller mentioned the right “peaceably to assemble.”¹ After reading John Inazu’s book, *Liberty’s Refuge: The Forgotten Freedom of Assembly*, the reasons for this collective memory loss are clearer. As Inazu explains, the freedom of assembly has languished in exile for many decades. Inazu takes the reader on the Assembly Clause’s fateful journey, from its prominence in the early republic,² to its 1939 New York World’s Fair glory,³ to its eventual desuetude.⁴ He expertly recounts how historical, political, intellectual, and jurisprudential forces transformed a seemingly clear constitutional guarantee into an also-mentioned right that occasionally plays second fiddle to freedom of speech. Inazu complains that the once-venerable “freedom of assembly” has been eclipsed and replaced by a judicially constructed, and doctrinally constricted, freedom of “expressive association.”⁵ As Inazu notes, the Supreme Court has not explicitly based a decision on the Assembly Clause in three decades.⁶

In *Liberty’s Refuge*, Inazu ably comes to assembly’s defense. His account sheds new light on the history and constitutional metamorphosis of a

* Professor of Law, William & Mary Law School.

1. U.S. CONST. amend. I.

2. JOHN D. INAZU, LIBERTY’S REFUGE: THE FORGOTTEN FREEDOM OF ASSEMBLY 29–34 (2012).

3. *Id.* at 55–57.

4. *Id.* at 61–62.

5. *Id.* at 2–3.

6. *Id.* at 62.

critical but now largely forgotten First Amendment freedom. That alone makes the book well worth reading. However, there is much more in the book than exegesis and excavation. Inazu seeks not only to rediscover assembly, in the sense of explaining what happened to it, but also to recover it in a manner that gives it contemporary relevance and force. He argues that a robust freedom of assembly ought to protect the formation, composition, and expression of groups.⁷ Inazu makes some provocative claims, in the best sense of that term. He pushes back against prevailing equality norms and principles that tend to cast groups like the Boy Scouts of America and the Christian Legal Society as illiberal villains.⁸ He forces readers to grapple with some uncomfortable questions regarding the limits of group autonomy in a liberal democracy. He asks whether a truly robust freedom of peaceable assembly ought to shelter even some racially exclusionary groups.⁹

I share Inazu's desire to return the freedom of peaceable assembly to something like its former glory. In *Liberty's Refuge*, however, Inazu's focus on the rise of expressive association and its relation to a few notable groups dominates the analysis to such an extent that the full import of a rediscovered freedom of assembly may remain somewhat obscured. My principal suggestion is that we try to recover assembly in the fullest and most robust possible sense. To that end, although I will make some critical observations, my Review will also clarify and amplify several of Inazu's central claims. If we can think of the Assembly Clause as an artifact or relic, Inazu has unearthed and exposed it to the light of day. While praising this effort, I want to suggest how we might pull the Assembly Clause fully from the ground.

Part I describes Inazu's account of the freedom of assembly and his central claims. In Part II, I address some concerns regarding interpretive methodology and the substantive implications of the book's principal focus on illiberal and potentially dangerous assemblies. Part III focuses on some of the positive, personal, and public aspects of freedom of assembly, which receive somewhat limited attention in the book. Part IV concludes with a discussion of the implications of a fully recovered right of assembly for traditional forms of public protest, demonstration, and dissent.

7. *See id.* at 2 (“The central argument of this book is that something important is lost when we fail to grasp the connection between a group’s formation, composition, and existence and its expression.”).

8. *See id.* at 168–72 (arguing that the protections of assembly should apply to groups like the Boy Scouts and the Christian Legal Society).

9. *See id.* at 13 (noting that one of the most difficult issues in balancing the right of assembly with antidiscrimination laws “is whether the right of assembly tolerates racial discrimination by peaceable, noncommercial groups”).

I. Recovery and Refuge

In *Liberty's Refuge*, Inazu presents compelling historical, intellectual, and jurisprudential narratives in order to further two primary goals. First, he seeks to recover the right to peaceable assembly by tracing its roots and explaining its eventual transformation into a right of expressive association. Second, Inazu articulates a theory of freedom of assembly under which the First Amendment would provide greater refuge to various aspects of group autonomy and liberty.

Inazu begins his examination with what, in retrospect, was clearly assembly's halcyon period. As Inazu explains in Chapter 2, in the early republic citizens routinely invoked and exercised the freedom to peaceably assemble by joining together in societies, civic organizations, public marches, religious rituals, and community festivals.¹⁰ In a fascinating historical account, Inazu demonstrates that the freedom of peaceable assembly has deep social, political, and constitutional roots. He describes how society members, abolitionists, women's suffrage proponents, labor agitators, and civil rights activists all invoked the freedom to peaceably assemble.¹¹ Inazu effectively narrates assembly's glory days as one of the "Four Freedoms" celebrated at the 1939 New York World's Fair and as a constitutional freedom touted by public figures and the general public.¹² Chapter 2 ends, rather abruptly, with a very brief discussion of what Inazu refers to as the "demise of assembly."¹³ As Inazu notes, "by the end of the 1960s, the right of assembly in law and politics was largely confined to protests and demonstrations."¹⁴ By the early 1980s, even this aspect of the right of assembly had been subsumed by First Amendment free speech doctrine.¹⁵

As Inazu observes, the merger of freedom of assembly and freedom of speech tells only part of the story. Something more momentous and transformative occurred with regard to the Assembly Clause. In Chapters 3 and 4, Inazu demonstrates that during what he calls the "National Security" and "Equality" eras the freedom of assembly was transformed into a right of association.¹⁶ These chapters represent the heart of Inazu's volume and offer its most intriguing insights.

10. *Id.* at 29–30.

11. *See id.* at 34–44 (describing the abolitionists' use of assemblies and noting that during the Progressive era, the women's movement, the labor movement, and African-Americans all invoked the freedom of assembly).

12. *Id.* at 55–57.

13. *Id.* at 61–62.

14. *Id.* at 61.

15. *See id.* at 62 ("[E]ven cases involving protests or demonstrations could now be resolved without reference to assembly.").

16. *Id.* chs. 3, 4.

Most scholarly attention has focused on the path of freedom of speech during these critical eras. As Inazu explains, however, during these periods the right of individuals to assemble in pursuit of common causes was directly challenged by government and ultimately legitimized in the courts.¹⁷ Inazu carefully examines the political, jurisprudential, and theoretical factors that led to the transformation and eventual interment of assembly. In Chapter 3, he points to the intersection of anticommunist sentiment and the civil rights movement, doctrinal disagreements among Supreme Court Justices, and the influence of pluralist political theorists like Robert Dahl.¹⁸ In Chapter 4, he highlights civil rights activists' challenges to segregationists' claims for group autonomy, the development of the constitutional right to privacy, and the rise of Rawlsian liberalism.¹⁹

Inazu's central claim is that the combination of these influences produced a weak associative right based upon principles of liberal congruence and consensus. It is difficult to gauge the degree of influence that political events and philosophers have on the process of constitutional interpretation. The right of expressive association appears to have been constructed through a type of common law constitutional interpretation.²⁰ Having first (wrongly) tethered the right of assembly to the right to petition and later ventured into the realm of constitutional privacy, the Supreme Court eventually arrived at the nontextual and ancillary (to speech) right of association. Nonetheless, in terms of the substance of expressive association Inazu's political and theoretical narratives support his conclusion that the right the Court ultimately recognized "depoliticizes and disembodies expression in order to neutralize dissent."²¹ Inazu characterizes the association right as an "enfeebled" version of assembly that restricts group autonomy, suppresses dissent, and pushes groups toward conformity and congruence.²² In sum, he argues that the "forgetting of assembly and the embrace of association . . . marked the loss of meaningful protections for the dissenting, political, and expressive group."²³

As part of his restorative project, in Chapter 5 Inazu articulates a "political theory of assembly."²⁴ He finds intellectual support for this theory in the work of Sheldon Wolin. Wolin criticized Rawls and other consensus

17. See, e.g., *NAACP v. Alabama ex rel. Patterson*, 357 U.S. 449, 466 (1958) (protecting the NAACP from state scrutiny of its membership lists).

18. INAZU, *supra* note 2, at ch. 3.

19. *Id.* at ch. 4.

20. See generally DAVID A. STRAUSS, *THE LIVING CONSTITUTION* (2010) (arguing that many areas of constitutional doctrine, including freedom of speech, developed according to common law methods and principles).

21. INAZU, *supra* note 2, at 155.

22. *Id.* at 4.

23. *Id.*

24. See *id.* at 153–57 (citing SHELDON S. WOLIN, *POLITICS AND VISION: CONTINUITY AND INNOVATION IN WESTERN POLITICAL THOUGHT* (2004)) (discussing Sheldon Wolin's scholarship).

theorists for demonizing dissent and disagreement and for falsely equating conformity and politeness with civic reasonableness.²⁵ Wolin argued that dissent, social conflict, and nonconformity are necessary *destabilizing* components of a healthy democracy.²⁶ With Wolin and against pluralist and liberal theorists, Inazu argues for a conception of assembly “that resists the state’s push for consensus and control.”²⁷ Inazu claims that robust protection for group autonomy allows individuals to create distance between individuals and the state. Rather than having democracy’s substance and limits dictated by a monist state, he argues that assembly empowers groups to experiment with various democratic forms and practices.²⁸ Inazu’s political defense of group autonomy offers a strong counternarrative to that relied upon by antidiscrimination proponents (most notably Andrew Koppelman).²⁹

Although he anticipates that a variety of civic, religious, and other groups would benefit from a recovered freedom of assembly, Inazu is particularly concerned with extending protection to groups that act or wish to act contrary to what is commonly perceived to be the “common good.”³⁰ As Inazu envisions it, a robust freedom of assembly would provide “strong protections for the formation, composition, expression, and gathering of groups, especially those groups that dissent from majoritarian standards.”³¹

Although he discusses other aspects of group autonomy, Inazu focuses primarily on protection for group membership decisions. Thus, according to Inazu’s account, the biggest losers in the gradual disappearance and transformation of assembly into expressive association are groups that resist or fail to comply with pluralist and liberal norms relating to inclusion and equality.³² Throughout the book, Inazu focuses primarily on groups like the Jaycees, the Boy Scouts (who have recently affirmed their policy against openly gay Scouts or adult Scout Masters),³³ the Christian Legal Society, and all-male fraternities.³⁴ Invoking equality principles and antidiscrimination laws, plaintiffs and governments pressed such organizations to open their doors to all comers.³⁵ Courts have mainly, although not uniformly, held that

25. See *id.* at 154–56 (citing JOHN RAWLS, *A THEORY OF JUSTICE* (1971) and WOLIN, *supra* note 24) (discussing theories of Dahl, Rawls, and Wolin).

26. *Id.* at 156.

27. *Id.* at 162.

28. *Id.* at 5–6.

29. *Id.* at 162–66.

30. *Id.* at 152–53.

31. *Id.*

32. See *id.* at 171 (arguing that under one popular theory of expressive association, “every group that challenged antidiscrimination law” would be subjugated to the state if the state determined that “discrimination is central to the group’s core expression”).

33. Erik Eckholm, *Boy Scouts to Continue Excluding Gay People*, N.Y. TIMES, July 17, 2012, http://www.nytimes.com/2012/07/18/us/boy-scouts-reaffirm-ban-on-gay-members.html?_r=0.

34. INAZU, *supra* note 2, at 132–46.

35. See, e.g., *Boy Scouts of Am. v. Dale*, 530 U.S. 640, 645 (2000) (“The complaint alleged that the Boy Scouts had violated New Jersey’s public accommodations statute and its common law

antidiscrimination principles trump group autonomy.³⁶ In contrast, Inazu envisions a “meaningful pluralism” that countenances “all-male fraternities, all-male Jaycees, and all-Christian student groups,” as well as “all-female sororities, all-female health clubs, and all-gay social clubs.”³⁷ Perhaps most controversially, Inazu’s conception of group autonomy might be broad enough to grant some First Amendment protection to the exclusionary policies of some private groups that exclude individuals on the basis of race.³⁸

Inazu does not address in detail how courts would actually enforce a recovered right of assembly. He defines it as a “presumptive right of individuals to form and participate in peaceable, noncommercial groups.”³⁹ Inazu briefly considers the textual limitation that is suggested by the adjective “peaceably.” He suggests that this may exclude such things as “[c]riminal conspiracies, violent uprisings, and even most forms of civil disobedience.”⁴⁰ Inazu also posits a nontextual limitation, namely that *commercial* groups are not entitled to protection under the Assembly Clause.⁴¹ For groups that are presumptively protected by the Assembly Clause, Inazu proposes that courts apply a “contextual” analysis that considers “how power operates on the ground.”⁴² Where private groups overreach, as for example when they exercise monopoly power with respect to certain goods or services, the state may be able to rebut the presumptive protection afforded under the Assembly Clause.⁴³ However, in most cases, Inazu expects that the presumption will prevail against governmental interference with groups’ autonomous decision making.⁴⁴

Liberty’s Refuge is an important contribution to the First Amendment literature. It provides a thick, careful, and intellectually rigorous account of a freedom that has languished for too long and which judges, lawyers,

by revoking [the Plaintiff’s] membership based solely on his sexual orientation.”); Bd. of Dirs. of Rotary Int’l v. Rotary Club of Duarte, 481 U.S. 537, 541–42 (1987) (seeking an injunction on the grounds that an international Rotary Club’s revocation of one of its members’ local charters, because the local club had admitted women members, violated the Unruh Civil Rights Act).

36. See, e.g., *Rotary Int’l*, 481 U.S. at 547 (holding that “application of the Unruh Act to local Rotary Clubs does not interfere unduly with the members’ freedom of private association”). But see, e.g., *Dale*, 530 U.S. at 659 (holding that the First Amendment prohibits the state from requiring that the respondent be readmitted to the Boy Scouts through the application of its public accommodations law, which does not justify such a severe intrusion on the Boy Scouts’ right to freedom of expressive association).

37. INAZU, *supra* note 2, at 11.

38. *Id.* at 14.

39. *Id.* at 166.

40. *Id.* at 167.

41. *Id.* at 167–68. For a critique of this specific limitation, see generally Robert K. Vischer, *How Necessary Is the Right of Assembly?*, 89 WASH. U. L. REV. 1403 (2012).

42. INAZU, *supra* note 2, at 172.

43. *Id.*

44. See *id.* at 169 (arguing that “in almost all cases, the protections of assembly should prevail”).

scholars, and citizens have paid far too little attention to over the past several decades. Inazu's book also tells a cautionary tale about constitutional meaning and textual transformation, and demonstrates the importance of giving full effect to the entirety of the First Amendment's text. *Liberty's Refuge* does not purport to provide a final answer or set of answers regarding the scope and limits of the freedom of assembly. Having recovered the Assembly Clause, Inazu merely points us in the direction of its future enforcement.

II. Interpreting Assembly

The question of interpretive methodology is an important one, particularly as it relates to a constitutional provision that has been in exile for decades. Having mistakenly abandoned assembly, the Supreme Court could conceivably resurrect it by providing a new substantive account. The recent treatment of the Second Amendment is instructive in this regard. Inazu's account raises several interpretive concerns. What sources ought to be consulted in re-interpreting the right of peaceable assembly? What justifications are there for adopting a distinctly political theory of assembly that focuses primarily on protecting the autonomy of dissenting groups? Should the interpretive model be atomistic, in the sense that it focuses on a single First Amendment provision, or holistic, in the sense that it synthesizes assembly and other rights? Finally, does Inazu's primary focus on dissent and nonconformance risk offering too much protection for illiberal and violent groups? Although these are serious concerns, I think Inazu has offered some convincing responses. I want to amplify a bit on those responses, and to suggest some additional support for them.

A. *Eclectic and Atomistic Methodologies*

The extent to which the Assembly Clause protects the sort of group autonomy Inazu identifies is not clear from its text. Perhaps assembly is a temporal right—meaning that it applies only to *temporary* groupings or affiliations, which must remain peaceable for their duration. If so, longstanding organizations like the Boy Scouts would find no refuge under the Assembly Clause. Further, we could interpret the requirement that assemblies be “peaceable” as a requirement that they respect equality rights. Under this interpretation, peaceable activity is activity that conforms to certain consensus norms regarding public order and social tranquility. Or, in terms of external limits, one might argue that the Fourteenth Amendment's Equal Protection Clause was intended to modify or limit the First Amendment's protection for freedom of assembly.

As noted earlier, Inazu claims that the Assembly Clause ought generally to protect groups against imposition of consensus norms.⁴⁵ He argues that the substantive meaning of the Assembly Clause can be derived in part from political and philosophical principles of dissent and nonconformity. Is this theoretical account attractive because it is consistent with the original understanding? Because it comports with a structural interpretation of the Bill of Rights? Or is Inazu's interpretation simply the best answer given all of the available historical and other evidence we have regarding freedom of assembly?

Inazu acknowledges the importance of interpretive methodology. His approach is refreshingly transparent. Inazu states that he is using an eclectic interpretive model, which is to say that no particular methodology (i.e., originalism, textualism, living constitutionalism) propels his interpretation of the Assembly Clause.⁴⁶ Thus, Inazu engages in a textualist approach when he renders a close reading of the text and (correctly, in my view) decouples freedom of assembly from the right to petition government for a redress of grievances.⁴⁷ He makes copious use of history, structural arguments, prudential principles, and various other constitutional "modalities" in examining the Assembly Clause. Inazu's political theory of assembly is consistent with these sources; to a large extent, it follows from them.

Eclecticism is a defensible mode of constitutional interpretation. Indeed, for a rights guarantee like the Assembly Clause that has been dormant for so long it may be the best method of recovering meaning.⁴⁸ The freedom of assembly is, as Inazu ably demonstrates, a product of historical, social, and political events and influences. Its meaning has been forged over time in the courts, in public debate, in national celebrations, and even in international diplomacy. Inazu's eclectic and interdisciplinary approach rightly takes account of all of these contexts and sources.

Given the centrality of group discrimination to his account, Inazu might have paid somewhat more attention to the intersection of the First Amendment and the Fourteenth Amendment. Moreover, he might have avoided framing the question as one involving a choice between Dahl and Rawls, on the one hand, and Wolin on the other. We are not actually choosing among political theorists or political theories, but among plausible interpretations of constitutional text. However, Inazu's account seems to be consistent with all of the available historical, structural, and other evidence relating to the freedom of assembly. He offers substantial evidence to

45. See *id.* at 155 (arguing, contrary to the view of consensus theorists, that groups with different, unpopular views should be protected).

46. *Id.* at 17–19.

47. *Id.* at 23–25.

48. Cf. STRAUSS, *supra* note 20, at 55 (arguing that "the text and the original understandings of the First Amendment are essentially irrelevant to the American system of freedom of expression as it exists today").

support his interpretation, and suggests reasons to doubt alternative interpretive accounts—including Andrew Koppelman’s historical narrative, which Inazu claims is incomplete and privileges equality concerns over group autonomy and liberty.⁴⁹ In light of all of this evidence, as Inazu correctly notes, the burden rests on others to come forward with a more plausible account.

Inazu’s interpretive methodology is both eclectic and atomistic. By atomistic I mean that it focuses intently on a single clause or rights provision and examines it mostly in isolation from other constitutional text. Other constitutional scholars, including some who have examined First Amendment freedoms, have adopted a similar approach.⁵⁰ There are both benefits and costs associated with this kind of atomistic methodology.

On the considerable plus side, scholars engaging in atomistic interpretation are able to offer deep historical and intellectual accounts of constitutional rights and other provisions. By zeroing in on the Assembly Clause, Inazu is able to offer a granular, detailed, and intellectually thick account of the right to peaceably assemble. Like eclecticism, atomistic interpretation may be particularly well suited to contexts in which constitutional text has been exiled or significantly transformed over time.

On the cost side, atomistic interpretation can lead to a degree of myopia. Inazu’s approach is situated at the opposite extreme from works like Thomas Emerson’s iconic *The System of Freedom of Expression*.⁵¹ Emerson treated the First Amendment’s expressive liberties—speech, press, assembly, and petition—as part of an interrelated system that served core functions such as individual fulfillment, the search for truth, and self-governance.⁵² Emerson incorporated a discussion of the right to peaceably assemble into this systematic account.⁵³ He interpreted assembly and other First Amendment rights as protections against regulating belief, coercing orthodoxy, and insisting on congruence and conformity.⁵⁴

These are essentially the same core values that Inazu ascribes to the freedom of assembly. Thus, there is apparently some connective tissue that

49. See INAZU, *supra* note 2, at 162–66 (citing ANDREW KOPPELMAN WITH TOBIAS BARRINGTON WOLFF, A RIGHT TO DISCRIMINATE? HOW THE CASE OF *BOY SCOUTS OF AMERICA V. JAMES DALE WARPED THE LAW OF FREE ASSOCIATION* 1–24 (2009)) (summarizing differences between Inazu’s historical narrative and Koppelman’s narrative regarding association).

50. The most notable recent example, which examines the First Amendment’s Petition Clause, is RONALD J. KROTOSZYNSKI, JR., RECLAIMING THE PETITION CLAUSE: SEDITIOUS LIBEL, “OFFENSIVE” PROTEST, AND THE RIGHT TO PETITION THE GOVERNMENT FOR A REDRESS OF GRIEVANCES (2012).

51. THOMAS I. EMERSON, *THE SYSTEM OF FREEDOM OF EXPRESSION* 15 (1970).

52. See *id.* at 6–7 (stating that the system of freedom of expression is an essential means of assuring individual self-fulfillment, advancing knowledge and discovering truth, and providing for participation in decision making by all members of society).

53. See *id.* at 286–92 (discussing the vital role that the “various modes of public assembly and petition play in a modern system of free expression”).

54. See *id.* at 292–388 (discussing rights of peaceable assembly and petition).

binds the First Amendment's provisions together. One of the weaknesses of Inazu's atomistic interpretation is that it treats the Assembly Clause as an island of liberty rather than as part of an interlocking and mutually supportive system. This makes it more difficult to determine how freedom of assembly relates to or intersects with other freedoms. Thus we learn from Inazu's account that "assembly is a form of expression" and that it protects groups from state-enforced conformity and congruence.⁵⁵ What is less clear, though, is how the freedom of assembly might differ from, support, or operate within the First Amendment's system.

Atomistic interpretation makes it more difficult to determine what marks the freedom of assembly as distinctive or unique relative to other neighboring First Amendment rights. In the context of a public parade or protest, for example, citizens may be engaging simultaneously in freedom of speech, petition, and assembly. What, if anything, is distinctive about the freedom of assembly in this context? What distinguishes it, in either form or substance, from the rights of expression and petition? Early in his account, Inazu notes that assembly overlaps with religious freedoms. Indeed, freedom of assembly's roots can be traced back to the trial of William Penn, a Quaker who was infamously charged with assembling for religious purposes.⁵⁶ As Inazu's examples involving Christian campus organizations and Jewish fraternities show,⁵⁷ in some important respects the connection between assembly and religious free exercise remains close today. What is distinctive about the Assembly Clause in the context of religious assemblies? Why ought it, rather than the Free Exercise Clause, apply when adjudicating formation and composition questions relating to religious groups?⁵⁸

Other holistic or synthetic interpretive questions occurred to me as I read *Liberty's Refuge*. For instance, might a fully recovered freedom of assembly correct some of the errors, ambiguities, or weaknesses of free speech doctrine? The social pressure to conform to majority norms and to avoid social conflict is quite strong. First Amendment protection for some anonymous speech offers only a partial antidote to privacy concerns.⁵⁹ As Inazu suggests, the freedom of assembly provides refuge from state interference with group formation.⁶⁰ Perhaps freedom of assembly, rather

55. See INAZU, *supra* note 2, at 4–5 (highlighting how the right of expressive association provides strong protection for the formation, composition, expression, and gathering of groups and enables meaningful dissent from majoritarian standards).

56. *Id.* at 24–25.

57. *Id.* at 144–45.

58. See generally Ashutosh A. Bhagwat, *Assembly Resurrected*, 91 TEXAS L. REV. 351 (2012) (considering the question of how the Religion Clauses should interact with the Assembly Clause).

59. Cf. *McIntyre v. Ohio Elections Comm'n*, 514 U.S. 334, 349–53 (1995) (recognizing in dicta that a state's interest in preventing fraud and libel might justify a limited identification requirement).

60. See INAZU, *supra* note 2, at 4 (arguing that the four principles of the history of assembly collectively counsel for the protection of group formation).

than or in addition to freedom of speech, provides a substantive basis for protection against certain forms of state surveillance. If so, then the relevant First Amendment question would not be whether the state's actions have "chilled" speech in some tangible way, but rather whether they have interfered with a private group's autonomy regarding formation and composition.⁶¹

To be clear, I am not suggesting that Inazu's interpretation is illegitimate because it lacks Emersonian breadth. Like the eclectic model, the choice to delve deeply and thickly with respect to a right or clause rather than more holistically or comparatively is a valid interpretive scholarly choice. Inazu acknowledges that more systematic work must be done. As he states in the book's conclusion, "if courts were to reaffirm the continued importance of the freedom of assembly, then they would need to explain its doctrinal framework and outline the relationship of assembly to other First Amendment freedoms."⁶² But perhaps in this instance what Inazu views as the cart ought to come before the horse. If we were able to more fully recover and explain what is distinctive about the freedom of assembly, we might have more success convincing courts that they ought to reaffirm this forgotten right.

B. *Recovering Assembly's Darker Side*

Below I discuss some of the more positive social and political functions of assembly. In interpreting the Assembly Clause, Inazu's focus is elsewhere. He is particularly concerned with protecting the membership decisions of nonconforming groups. This orientation could create the impression that a recovered right of assembly will be useful primarily to society's most illiberal and dangerous assemblies. Why recover a right that benefits mobs and troublemakers? As Professor Bhagwat asks in a recent symposium contribution, is the freedom of assembly a refuge for constitutional liberty or a refuge for "scoundrels"?⁶³ Bhagwat is rightly concerned that the limits of the freedom of assembly be clearly defined, in particular with regard to potentially violent groups. Both in the book itself and in subsequent commentary,⁶⁴ Inazu offers some tentative responses to readers' concerns about assembly's darker side. Here, again, I want to elaborate on these responses and to offer some additional observations about the importance of protecting dissent and social conflict as manifested in

61. See *Laird v. Tatum*, 408 U.S. 1, 13–14 (1972) (holding that plaintiffs who objected to U.S. Army surveillance had not established standing to challenge the data-gathering program because they had not shown any regulatory effect on their expressive activities).

62. INAZU, *supra* note 2, at 186.

63. Ashutosh Bhagwat, *Liberty's Refuge, or the Refuge of Scoundrels?: The Limits of the Right of Assembly*, 89 WASH. U. L. REV. 1381, 1381 (2012).

64. See, e.g., John D. Inazu, *Factions for the Rest of Us*, 89 WASH. U. L. REV. 1435, 1438–40 (2012) (responding to concerns about the line between peaceable and violent assembly).

assemblies. In Part III, I will focus on the more positive aspects of freedom of assembly that receive less attention in Inazu's account.

Inazu argues that freedom of assembly ought to protect against certain forms of state-enforced orthodoxy.⁶⁵ In most cases, the freedom of peaceable assembly ought to bar coercive attempts by government to control the internal norms and practices of private assemblies. In a society that celebrates individualism but generally expects group conformity with regard to certain social norms and practices, a conception of pluralism that actually facilitates difference is indeed critically important. Inazu singles out a few organizations such as the Boy Scouts and the Christian Legal Society, which have been involved in recent high-profile disputes.⁶⁶ However, this sort of protection is also important to a host of other groups. Among these are American Muslims, Wiccans, Occupy Wall Street protesters, "Birthers," conspiracy theorists, medical marijuana advocates, Tea Party members, day laborers, labor strikers, gun advocates, and other individuals who join together and share creeds, causes, or conditions that many do not view as serving the common good.

It is not easy to be a dissenter or a nonconformist in America. That may strike some as an odd assertion. After all, Americans celebrate countercultural trends and actions. Indeed, they sometimes make heroes of nonconformists. However, it is still far easier to get along if one goes along with prevailing social and political norms. Dissenters and nonconformists face considerable pressures, both from government regulators and prevailing cultural forces, to get on board or in line.⁶⁷ Members of the dissenting and other out groups mentioned above can certainly attest to the pressure placed upon them to conform to majority religious, social, and political norms. They are frequently labeled discriminators, bigots, outsiders, weirdos, whackos, whiners, freeloaders, and closed-minded ideologues.⁶⁸ Whether they take the form of public protest movements, group memberships, or fringe causes, dissent and nonconformity can still use all the assistance they can get. Dissenting and nonconforming groups are not threats to democracy;

65. See *supra* note 44 and accompanying text.

66. See *supra* note 8 and accompanying text.

67. See, e.g., Max Abelson, *Occupy Plans 'S17' Wall Street Tie-Up*, WASH. POST, Aug. 30, 2012, at A3 (detailing plans for a demonstration to mark the one-year anniversary of the Occupy Wall Street movement, despite challenges posed by protester "burnout," and recounting how "governments around the world used concussion grenades, gas, riot gear, pepper spray and arrests to disband camps and protests"); Tina Susman & Andrew Tangel, *Protesters March Back to Wall Street*, L.A. TIMES, Sept. 18, 2012, at A8 (noting that more than 180 protesters were arrested during the one-year anniversary demonstration and describing popular criticisms of the movement for its "lack of focus" and its "failure to . . . adopt specific issues").

68. See, e.g., Editorial, *Occupy Plus One Year*, N.Y. POST, Sept. 17, 2012, at 24 (characterizing Occupy Wall Street protestors as "obnoxious outliers" and a "ragtag assemblage of stragglers, radicals, moochers, trust-fund sophists, bums, rapists, drug-dealers, petty criminals and cop-car poopers").

they are central components of our political and constitutional system. One of Inazu's signal contributions is to remind us of this easily forgotten fact.

Of course, there is a darker side to freedom of assembly. Some groups may actually be dangerous. As Professor Bhagwat has observed, a broad freedom of assembly might facilitate the formation and activities of violent groups.⁶⁹ Here, though, we must be careful not to adopt a common fallacy. During far too many periods of American history, including the current era, public officials and the public at large have equated assemblies with angry and destructive mobs.⁷⁰ Although his historical account is otherwise thick, Inazu underemphasizes this part of assembly's narrative.

Groups that reject consensus norms and occupy positions at the fringe of American culture ought not to be, for that reason alone, considered threats to national security or public safety. Of course, it is true that as collective enterprises, assemblies can be more dangerous than individual actors. None of the individual perpetrators of the September 11, 2001 attacks on the United States could have done as much damage acting alone. Many other dangerous networks, groups, and associations, including separatists and neoracists, currently reside in the United States. As Inazu notes, however, the Assembly Clause protects only "peaceable" forms of assembly. That clearly excludes individuals who assemble for the common purpose of engaging in acts of violence. Freedom of assembly offers no First Amendment immunity or defense for participants in criminal conspiracies such as the September 11 attacks.

Beyond this point, Inazu has conceded that he "lack[s] a clear sense of where the peaceability line ought to be drawn."⁷¹ I do not think this is an acute problem. With regard to violent conspiracies and the like, as Inazu has noted, the First Amendment is essentially irrelevant.⁷² This is true whether we are talking about freedom of speech or a recovered version of freedom of assembly. With regard to other out groups that do not intend to or actually engage in violent activities, the presumption of protection ought to apply. As I discuss below, the "peaceably" limitation would seem to present the most acute interpretive difficulties as applied to assemblies engaged in civil disobedience and other nonconforming, but nonviolent, activities. Even here the danger of an expansive right of assembly will likely be minimal. The assemblies at issue are likely to form or act in the open, on public streets and

69. Bhagwat, *supra* note 63, at 1394–96.

70. See, e.g., Carolyn Jones, *Oakland's Top Administrator Tough Enough for City She Loves*, S.F. CHRON., July 8, 2012, at A1 (profiling Oakland City Administrator Deanna Santana, who issued the final eviction notice to Occupy Wall Street protestors in Frank Ogawa Plaza on the basis of "safety issues," and quoting Santana expressing her concern that "if this place went up in flames, it'd be on me"); Andrew Tangel, *At 1 Year, Occupy's Effect Is Still Hard to Gauge*, L.A. TIMES, Sept. 15, 2012, at A1 ("Polls have shown that the public generally supports Occupy[] [Wall Street's] message but not its disruptive tactics.").

71. Inazu, *supra* note 64, at 1438.

72. *Id.* at 1440 & n.29.

in public parks where regulations define what is and is not lawful in terms of public protests and other forms of outdoor social conflict.

Perhaps Inazu's most provocative claim relates not to violent groups but to private assemblies that engage in racially or ethnically discriminatory practices. As Inazu forthrightly acknowledges, the suggestion that some such groups ought to receive refuge under the Assembly Clause is the most troubling and tentative in his volume.⁷³

I am not sure that we ought to protect the membership and other decisions of such assemblies—even if we currently allow them to use the public streets to engage in protest and other forms of expression. I do not think that it suffices to say, as Inazu has in defending this part of his analysis, that some degree of overprotection of freedom of assembly follows ineluctably from the logic of overprotection of freedom of speech.⁷⁴ The fact that some offensive and even vile expression is protected as part of the price for a robust freedom of speech does not necessarily answer the question whether we ought to protect discriminatory conduct by private groups or tolerate hateful organizations. Whether the First Amendment ought to protect degrading and hateful expression remains a matter of significant and ongoing debate.⁷⁵ Further, the Supreme Court's observation that free speech "may indeed best serve its high purpose when it induces a condition of unrest, creates dissatisfaction with conditions as they are, or even stirs people to anger" does not necessarily map well onto the sorts of private decision making Inazu discusses in the book.⁷⁶ The costs of exclusion and the societal dynamics associated with discriminatory groups may well require a different calculus and some distinct limitations.

Here, though, is another place where examining the ties to other First Amendment rights might bear some fruit. Might there be, for example, some notion of "counter-assembly" under which groups that are offensive to even the most deeply held societal norms are countered by groups that accept such norms?⁷⁷ Single-sex educational institutions compete with coeducational ones. Groups espousing traditional heterosexual marriage are countered by numerous gay rights groups. Male-only fraternities coexist on campuses across the country with female-only sororities. The National Rifle Association regularly spars with countless gun control groups. And civil rights groups keep tabs on and challenge racist organizations. Or perhaps we

73. INAZU, *supra* note 2, at 14.

74. Inazu, *supra* note 64, at 1437.

75. See generally JEREMY WALDRON, *THE HARM IN HATE SPEECH* (2012) (arguing for more regulation of hate speech, contrary to the mainstream position of overprotection).

76. *Terminiello v. Chicago*, 337 U.S. 1, 4 (1949).

77. See *Whitney v. California*, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring) (noting that, except in emergencies, the remedy for exposing falsehoods in speech is more speech, not repression), *overruled in part by* *Brandenburg v. Ohio*, 395 U.S. 444 (1969).

ought to develop a theory of tolerance that is uniquely related to assembly.⁷⁸ We might also borrow from the pluralist approach that has developed under the First Amendment's religion clauses.⁷⁹ So long as there are meaningful rights of entry and exit, and the group has no monopolistic power or characteristics, the state really ought to remain neutral with regard to the formation and composition of assemblies. Again, I am not sure that we ought to provide these or other justifications for protecting the autonomy of illiberal assemblies. But if we are to do so, more theoretical thought and effort must be devoted to producing a justification for extending assembly so far.

Inazu is undoubtedly correct that if the Assembly Clause is revived in the manner he suggests, we will have to think very carefully about the amount of breathing space we want to create for certain kinds of assemblies. In terms of managing this concern, Inazu has cast significant doubt on the expressive association doctrine. Determining how the problem of invidious discrimination by groups ought to be resolved under the Assembly Clause is a matter that requires further reflection.

III. The Forms and Functions of Peaceable Assembly

Liberty's Refuge offers a compelling argument that institutional autonomy, in particular with respect to membership decision making, is a critical aspect of freedom of assembly.⁸⁰ However, a fully recovered freedom of assembly would protect a diverse array of groups and would serve important functions, some of which Inazu addresses only briefly. For the purpose of amplification, and toward the end of taking assembly's fullest possible measure, this Part examines more closely the forms and functions of assembly.

A. *Assembly's Diverse Forms*

What is an "assembly"? Although Inazu is an otherwise careful textualist,⁸¹ he does not offer a basic definition of this term (as opposed to a definition of the *right* of assembly itself). An assembly is "a group of people gathered together in one place for a common purpose."⁸² The shared space

78. See generally LEE C. BOLLINGER, *THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA* (1986) (advancing the theory that societies that are tolerant of ideas that are legitimately unworthy of protection are strengthened by that tolerance).

79. See generally Michael W. McConnell, *The Origins and Historical Understanding of Free Exercise of Religion*, 103 HARV. L. REV. 1409 (1990) (discussing the Free Exercise Clause's origins in religious pluralism).

80. INAZU, *supra* note 2, at 152–53.

81. For example, Inazu convincingly argues that freedom of assembly and freedom to petition government for a redress of grievances are independent and freestanding rights. *Id.* at 23–25. Inazu is also careful to note that "peaceably" limits the scope of the right of assembly. *Id.* at 166–67.

82. NEW OXFORD AMERICAN DICTIONARY 95 (2001).

may be physical or virtual.⁸³ The common purposes may be social, political, religious, cultural, or educational. Although he acknowledges other forms, Inazu focuses primarily on groups or assemblies that are longstanding, organized institutions. As noted earlier, Inazu's interpretive and normative accounts treat the primary function of freedom of assembly as preserving autonomous space for dissenting and nonconforming organizations and institutions.⁸⁴

In Chapter 2's historical narrative, Inazu describes an extraordinary variety of assemblies. He discusses societies, institutions, congregations, organizations, rituals, feasts, protests, parades, and demonstrations. These types of gatherings have long been a critical part of American social, civic, and political culture. Indeed, they remain so today. After describing this rich history, however, Inazu's analysis conveys the impression that "assembly" and "organization" are synonymous terms and that the core of a recovered freedom of assembly is protection for group autonomy—particularly for certain well-organized, illiberal groups that face public disapproval and discrimination lawsuits. This orientation is in large part owing to Inazu's following the path forged by the Supreme Court, which led ultimately to recognition of the right of expressive association.

As Inazu clearly recognizes, however, assemblies take many forms. Assemblies can be quite small or very large. They can have private or public orientations. Historically, the right to assemble has protected the formation and composition of a diverse array of private groups including social clubs and churches. Some of these private groups are formed with the intention of making public claims, while others seek generally to maintain a more private existence and profile. Indeed, some groups form with the expectation that they and their members will remain completely anonymous.

As the discussion in Chapter 2 also shows, assemblies can be formally or informally organized. We might think of them as being situated on a continuum, ranging from longstanding institutions to spontaneous and casual gatherings. Assemblies may be organized with regard to a specific message or ideology, or they may be looser forms of alliance. They may be heavily regulated, as in the case of political parties, or they may operate mainly beyond and outside the state's control. Assemblies may be aligned against the state, or in some cases constituted specifically to support current public laws and policies.

Finally, assemblies have both collective and individual characteristics. They protect both organizational and individual interests. In his recuperative account, Inazu does not entirely ignore the individual dimension of assembly. But as I discuss below, for the most part he appears to conceptualize freedom of assembly as a form of protection for groups and specifically for their

83. Inazu examines virtual assemblies in a forthcoming paper. See John D. Inazu, *Virtual Assembly*, 98 CORNELL L. REV. (forthcoming 2013).

84. See *supra* note 44 and accompanying text.

organizational autonomy. However, as a *personal* freedom, the right to peaceably assemble belongs to each of the individuals who choose to participate in the formation and activities of the common venture.

In sum, assemblies in various forms are everywhere and all around us. Indeed, wherever two or more people gather in a common space an assembly has taken place.

The ability to gather in public has been a particularly important aspect of the freedom of assembly. As American history demonstrates, less structured and even spontaneous gatherings were in many cases the principal beneficiaries of a freedom of peaceable assembly. The freedom of assembly has facilitated traditional public displays such as pickets, demonstrations, parades, and protests. In contrast to the civic and religious organizations Inazu focuses on in the book, this is assembly's core dimension.

As Inazu briefly mentions early in the book, the Assembly Clause protects "the occasional, temporal gathering that often takes the form of a protest, parade, or demonstration."⁸⁵ Indeed, I think this is not only the traditional but perhaps also the most natural reading of the First Amendment's Assembly Clause. On the infrequent occasions when it has mentioned assembly, the Supreme Court seems to have agreed. Writing for all but one Justice in *Edwards v. South Carolina*⁸⁶ in 1963, Justice Stewart described a civil rights demonstration by 187 students on the State Capitol grounds as the exercise of free speech, free assembly, and freedom to petition for redress of grievances "in their most pristine and classic form."⁸⁷ The classic assembly consisted of a group of citizens gathered in the public square for a peaceful and temporary demonstration. These individuals were, and as I will explain, in some sense remain, most in need of the refuge of freedom of assembly.

At the end of his historical narrative in Chapter 2, Inazu notes with evident disappointment that by the 1960s the Supreme Court appeared to have limited freedom of assembly to public assemblies, protests, and demonstrations.⁸⁸ The real disappointment, as Inazu only briefly mentions, is that within the next two decades the Court buried even this "pristine and classic" form of assembly under an ever-expanding free speech doctrine.⁸⁹

Of course, if the Assembly Clause does not protect the most obvious and traditional associative endeavors, then it could be difficult to establish that it provides refuge for the formation, composition, and expression of civic and other organizations that are highly structured and do not exist to make public claims. Perhaps Inazu believes that protection for the more traditional forms of assembly such as protests, parades, and demonstrations is

85. INAZU, *supra* note 2, at 2.

86. 372 U.S. 229 (1963).

87. *Id.* at 235.

88. INAZU, *supra* note 2, at 61.

89. *Edwards*, 372 U.S. at 235; INAZU, *supra* note 2, at 61–62.

meaningfully assured under the Free Speech Clause, or that such protection will simply be a natural byproduct of the recognition of group autonomy he espouses. At least in the specific sense Inazu describes and analyzes the concept, group autonomy has not been the central concern of traditional assemblies. Given the challenges to traditional assembly, which included vigilante responses as well as official forms of suppression and abuse, restrictions on formation and composition were subordinate concerns. If we are to fully recover the Assembly Clause, we need to reconceive how it applies to more traditional forms and functions. In other words, the recovery effort ought to begin at assembly's roots.

I do not mean to argue that the freedom of assembly cannot be extended beyond traditional public gatherings, or that its meaning is frozen in time in some originalist sense. As Inazu observes, groups of individuals who have historically joined under an organizational umbrella or operated as hierarchical institutions have long *claimed* to be engaged in acts of assembly.⁹⁰ Although most of these groups used repertoires like demonstrations and protests, not all of them did. This history is certainly some evidence of the American public's own interpretation of assembly.⁹¹ Moreover, as a matter of simple definition, the groups whose autonomy Inazu is most concerned with protecting qualify as "assemblies." My concern is not that Inazu has wrongly or illegitimately interpreted the First Amendment's text, but rather that in his effort to transform "association" back into "assembly" Inazu may have given an inordinate amount of attention to a specific subset or type of assemblies, or to a specific problem created by the Supreme Court's interpretive adventurism. After Chapter 2, the more traditional forms of public assembly fade from view. In Part IV, I will examine how a recovered Assembly Clause might facilitate more traditional forms of public contention and dissent.

B. *Assembly's Functions*

As I have noted, Inazu is principally concerned with demonstrating how and why group autonomy has been harmed by the First Amendment doctrine of expressive association. Under his account, the primary beneficiaries of a recovered freedom of assembly would be dissident, exclusionary, and

90. I am less certain whether the assembly label applies in cases like *Holder v. Humanitarian Law Project*, 130 S. Ct. 2705 (2010). In that case, American citizens sought to engage in peaceful expressive activities such as the teaching of international law to designated foreign terrorist organizations. *Id.* at 2716. These individuals sometimes occupied common space and worked for a common purpose. *See id.* (describing the types of activities in which plaintiffs intended to engage, including training, offering legal expertise, and engaging in advocacy on behalf of the designated foreign terrorist organizations). In that sense, they meet the definition of an assembly. I am not certain how Inazu believes a recovered right of assembly would have assisted the plaintiffs in *Humanitarian Law Project* or altered the Court's analysis. Inazu seems to use the case primarily to demonstrate the ambiguity of the right of "expressive association." INAZU, *supra* note 2, at 4–6.

91. *See generally* JACK M. BALKIN, *LIVING ORIGINALISM* 17–18 (2011) (emphasizing the citizenry's understanding of constitutional provisions as an aspect of constitutional interpretation).

nonconforming organizations or groups. Inazu is particularly concerned with preserving space in which such groups can participate in self-governance (relatively) free from state interference. This is especially important for groups that engage in dissent, fail to conform to consensus norms and practices with regard to such things as political organization and rational discourse, and form alliances whose particular message may not be apparent to outsiders (including, in particular, judges).⁹² On Inazu's negative reading, the freedom of assembly allows private groups to resist the state's efforts to impose what Inazu claims are majority norms of consensus, congruence, and conformity.

Inazu addresses some of the most important defensive or negative attributes of a right of peaceable assembly. He argues that assembly is "most relevant when its exercise is challenged by the state."⁹³ But Inazu's focus on the struggle between certain private groups and the state's mechanisms of control downplays some of the more positive and personal aspects of the freedom of assembly. If we are to fully recover and restore the freedom of assembly, we must exhume not only its various forms but also its diverse functions. Moreover, we ought to consider those functions not from the perspective of the associative right the Supreme Court has recognized, but in light of the recovery of a freestanding, distinct, and robust Assembly Clause that this substitute has replaced.

Again, my goal here is more amplification than criticism. Inazu has a very brief discussion at the beginning of the book concerning what he calls the "social vision of assembly."⁹⁴ In addition to enabling meaningful dissent, he notes that the right of assembly "provides a buffer between the individual and the state" and contributes to "the shaping and forming of identity."⁹⁵ As Inazu wryly observes, "We lose more than the shared experience of cheese fries and cheap beer when we bowl alone."⁹⁶

I wish Inazu had elaborated on this "social vision."⁹⁷ If we accept Inazu's account, then it follows that the collective forgetting of the freedom of assembly has imposed significant social and political costs on American society. In some sense, it is true that constitutional rights are most important when the activities they protect are being directly challenged by the state.

92. See INAZU, *supra* note 2, at 156–62 (discussing dissenting, political, and expressive assemblies). Although Inazu raises some legitimate concerns regarding the interpretation of group messages, I am more optimistic regarding courts' ability to assess meaning in this and other contexts. See Timothy Zick, *Cross Burning, Cockfighting, and Symbolic Meaning: Toward a First Amendment Ethnography*, 45 WM. & MARY L. REV. 2261, 2375–79 (2004) (discussing judicial interpretation of group membership).

93. INAZU, *supra* note 2, at 156.

94. *Id.* at 5.

95. *Id.*

96. *Id.*

97. As Inazu indicates, the "social vision of assembly" he describes is based upon the work of scholars such as Robert Putnam, Alasdair MacIntyre, Charles Taylor, and Michael Sandel. *Id.* at 5 n.10.

However, it is also the case that the mere existence and recognition of an enforceable and robust constitutional right, such as the right to peaceably assemble with others, can serve critical functions which *precede* such challenges—and, indeed, may even prevent them from ever occurring. Further, even apart from any direct challenges by the state, the right of assembly can serve a variety of positive functions.

First and perhaps foremost, freedom of assembly provides a degree of safety and comfort in numbers. It is true, as discussed above, that this same attribute may increase the danger arising from assemblies. However, let us assume for the moment that we are talking only about “peaceable” assemblies. It may be difficult for contemporary Americans to appreciate the fear those accused in the 1950s of being Communists, or fellow travelers, experienced when they engaged in the simple act of meeting with others in private or public settings.⁹⁸ The ability to freely assemble or join with others fortifies individuals. It emboldens them to come forward, and to participate in social and political activities. In addition to creating space for group activities and group autonomy, the freedom of assembly facilitates a variety of *individual* acts of defiance, contention, and expression.

Freedom of assembly also serves various emotional and psychological functions. The act of assembly creates a sense of solidarity or common cause. It excites and energizes individuals, whether they gather to knit scarves, play soccer, pray, or participate in marches or protests. It fosters personal and civic pride by providing outlets and venues for the pursuit of common causes. Freedom of assembly does not simply allow individuals to develop their own identities. It allows otherwise marginalized individuals to be present with others and to communicate specific *identity claims* to the state and to the general public. For many individuals, this is a critical aspect not only of self-governance but also of personal self-esteem. In sum, a robust freedom to peaceably assemble with others facilitates full participation in and enjoyment of communal life.

In political terms, the freedom of assembly encourages and facilitates forms of local engagement. It provides foundation and structure for social and political projects. The ability to join with like-minded others allows citizens to form political associations and encourages them to contemplate future endeavors and initiatives. This may lead to new and unique institutions, including new political organizations and parties. Further, freedom of assembly strengthens and amplifies individual voices. It forces

98. See *Communist Party v. Subversive Activities Control Bd.*, 367 U.S. 1, 103 (1961) (holding that the right to assemble is secondary to the right of Congress “to bring foreign-dominated organizations out into the open where the public can evaluate their activities informedly against the revealed background of their character, nature, and connections”); David E. Bernstein, *The Red Menace, Revisited*, 100 NW. U. L. REV. 1295, 1301 (2006) (reviewing MARTIN H. REDISH, *THE LOGIC OF PERSECUTION: FREE EXPRESSION AND THE MCCARTHY ERA* (2005)) (providing background information about the Smith Act and other restrictions aimed at communists that limited the right of assembly).

officials and other members of the community to take notice by providing a rough depiction of individual preferences. In these representative and democratic senses, assembly acts as an informal method of voting or casting preferences—a way of marking or identifying oneself, often through public affiliations, as supportive of a particular position, cause, or side. Note that assembly serves this particular function whether individuals form a group at the fringes of societal norms or one situated within a majority consensus.

Inazu suggests that the reason for protecting groups' membership and leadership choices is that "the existence of a group and its selection of members and leaders are themselves forms of expression."⁹⁹ This obviously raises the question whether the freedom of assembly he espouses is cut from the same speech cloth as the right of expressive association.¹⁰⁰ I don't think that it is. However, had Inazu placed more emphasis on the individual social and political benefits of assembly, the separation would have been much clearer. Many of these functions are nonexpressive. They are a form of social sustenance and a critical part of our political structure. On this view, the fact that assembly protects the Boy Scouts' ability to express its preferences through exclusion is not the central point. The critical aspects of assembly lie beneath the surface of that public message; they are antecedent to the state's challenge to it.

Inazu's account of freedom of assembly is primarily political rather than sociological. However, elaborating somewhat on the positive and personal functions of assembly would have clarified the extent of assembly's independence from speech. More importantly, it would have allowed for a fuller recovery and explication of the variety of functions served by the freedom of assembly.

IV. Assembly and Outdoor Contention

As I noted earlier, perhaps the most natural interpretation of the Assembly Clause is that it protects an individual's right to gather with others for some limited period in a public place in order to pursue some common cause. Thus, whenever and wherever two people gather in a public place where they have a right to be, for lawful and peaceful purposes, the Assembly Clause ought to protect their right to do so. As citizens of authoritarian nations will attest, this is not some secondary or minimal constitutional concern.¹⁰¹ Where the freedom of assembly is recognized and

99. INAZU, *supra* note 2, at 152.

100. See Bhagwat, *supra* note 63, at 1383 (questioning Inazu's account insofar as it relies upon expressive values).

101. See, e.g., David M. Herszenhorn, *New Russian Law Assesses Heavy Fines on Protesters*, N.Y. TIMES, June 8, 2012, <http://www.nytimes.com/2012/06/09/world/europe/putin-signs-law-with-harsh-fines-for-protesters-in-russia.html> (reporting on enactment of a new Russian law restricting street demonstrations); Jim Yardley, *China Sets Zones for Olympics Protests*, N.Y. TIMES, July 24, 2008, http://www.nytimes.com/2008/07/24/sports/olympics/24china.html?_r=3&ref=world

enforced, authorities cannot without good cause require individuals to disperse, desist, disband, or move along. This right to be and to remain in public places lies at the core of the right of peaceable assembly.

Inazu has offered convincing reasons for recognizing other forms of assembly. However, a recovered Assembly Clause would be as or even more important to outdoor politics than to the indoor membership decisions of civic organizations and private businesses.¹⁰² Admittedly, restrictions on public protest and assembly were not Inazu's *raison d'être*. However, as I suggested earlier, a full recovery of the Assembly Clause will not be possible without some consideration of its relation to traditional forms of public assembly and contention. Inazu's account may offer some important insight with regard to this more traditional dimension of freedom of assembly. I want to make this contribution more explicit, and to raise some issues that require further consideration by Inazu and others who are interested in more public forms of dissent and contention.

In my own work, I have emphasized the necessity of adequate physical resources for the effective exercise of public speech, assembly, and petition rights.¹⁰³ I have argued that over time, a variety of societal, political, and jurisprudential forces have reduced the supply of public space that is available to individuals and groups who wish to engage in expression and politics out of doors. In brief, these and other forces have produced a significantly diminished public square. In addition, even in the remaining public spaces, individuals who wish to engage in speech, assembly, and petition activities are too often displaced by a variety of regulatory mechanisms, including the construction of "speech zones."¹⁰⁴

Had it been published prior to my own, Inazu's book would have provided welcome support for my thesis regarding access to public spaces, particularly public forums. According to the Supreme Court, these are places such as public streets and parks, which have "time out of mind" been available "for purposes of assembly, communicating thoughts between citizens, and discussing public questions."¹⁰⁵ As Inazu notes, in the early 1980s the Supreme Court "swept the remnants of assembly within the ambit of free speech law."¹⁰⁶ There assembly's remnants were combined with increasingly anemic public speech and petition rights, which were

&pagewanted=print (discussing China's repression of free speech assembly rights and the country's attempt to appear less repressive by creating "free speech zones" during the 2008 Olympics).

102. For a recent treatment of this aspect of assembly, see generally Tabatha Abu El-Haj, *The Neglected Right of Assembly*, 56 UCLA L. REV. 543 (2009).

103. TIMOTHY ZICK, *SPEECH OUT OF DOORS: PRESERVING FIRST AMENDMENT LIBERTIES IN PUBLIC PLACES* 3-4 (2009).

104. Timothy Zick, *Speech and Spatial Tactics*, 84 TEXAS L. REV. 581, 636 (2006).

105. *Hague v. Comm. for Indus. Org.*, 307 U.S. 496, 515 (1939); see also *Perry Educ. Ass'n v. Perry Local Educators' Ass'n*, 460 U.S. 37, 45-46 (1983) (describing public forum categories).

106. INAZU, *supra* note 2, at 61.

themselves hemmed in by an increasingly restrictive system of bureaucratic regulations.¹⁰⁷

Although Inazu focuses primarily on internal group autonomy, his account of the Assembly Clause has important implications for public assembly and contention. To examine some of these implications, I want to consider Inazu's account against the background of the Occupy Wall Street (OWS) demonstrations. The demonstrations, which occurred across the United States (and indeed spread to several foreign nations) during the fall of 2011,¹⁰⁸ are contemporary examples of the sort of public contention that was common during assembly's robust abolitionist, labor, and civil rights periods. In part for the reasons Inazu points to in his book, public discussions and litigation involving the OWS protests focused almost exclusively on free speech concerns.¹⁰⁹ This was so even though the Assembly Clause's language most closely captures the OWS signature repertoire—gathering in public for a common purpose or purposes.

According to Inazu, "The right of assembly is a presumptive right of individuals to form and participate in peaceable, noncommercial groups."¹¹⁰ OWS is clearly a noncommercial group, and thus entitled to presumptive protection under the Assembly Clause.¹¹¹ In addition, the OWS demonstrations served all of the core functions of assembly. They provided critical outlets for dissenters, nonconformists, and dissidents. OWS demonstrations allowed and perhaps emboldened individuals to challenge consensus norms.¹¹² The assemblies facilitated public dissent, politicized group activity, and provided channels for expression. They created space within which citizens could resist governmental control. The ability to assemble with others in common public spaces provided incubation space for a potential social movement. Further, the OWS assemblies allowed individuals to experiment with unique forms of democratic organization.¹¹³

107. See generally ZICK, *supra* note 103 (discussing the restriction of public speech rights under the First Amendment's public forum and time, place, and manner doctrines).

108. *Occupy Wall Street Protests Spread, But Can the Movement Gain Critical Mass?*, WASH. POST, Oct. 13, 2011, http://www.washingtonpost.com/business/occupy-wall-street-protests-spread-but-can-the-movement-gain-critical-mass/2011/10/13/gIQAzOM2hL_print.html.

109. See, e.g., *Occupy Minneapolis v. Cnty. of Hennepin*, Civ. No. 11-3412 (RHK/TNL), 2011 WL 5878359, at *4 (D. Minn. Nov. 23, 2011) (holding that sleeping or erecting tents on public property by Occupy protesters is protected free speech).

110. INAZU, *supra* note 2, at 166.

111. See Shelley DuBois, *Occupy Wall Street: Yes, There is Organization*, CNN MONEY (Dec. 7, 2011, 11:35 AM), <http://management.fortune.cnn.com/2011/12/07/occupy-wall-street-yes-there-is-organization/> (describing the "grassroots," noncommercial organization of the Occupy Movement).

112. Cf. INAZU, *supra* note 2, at 5 (discussing the importance of informal group assembly to democracy).

113. See Meredith Hoffman, *Protestors Debate What Demands, If Any, to Make*, N.Y. TIMES, Oct. 16, 2011, <http://www.nytimes.com/2011/10/17/nyregion/occupy-wall-street-trying-to-settle-on-demands.html?ref=occupywallstreet&r=moc.semityn.www> (describing the democratic process for decision making).

OWS became well known not just for its outward displays of commandeering and camping in public places, but also for its internal methods of communication and unique approach to governance by consensus.¹¹⁴ Finally, as have other public assemblies, the OWS demonstrations “disrupt[ed] social norms and consensus thinking.”¹¹⁵ They initiated a national and international conversation concerning issues like social equality, fairness, capitalism, and political representation.¹¹⁶

Perhaps the most frequently commented-upon aspect of the OWS demonstrations, at least in the mainstream media, was the apparent lack of a coherent message associated with the demonstrations or the group itself.¹¹⁷ Here Inazu offers a key insight. The First Amendment does not protect assembly solely for the purpose of communicating some identifiable, coherent message. Assembly is protected in its own right; it stands on its own bottom. The act of assembling is thus itself the relevant constitutional event. If individuals want to assemble for the purpose of snapping their fingers, chanting in tongues, or simply showing solidarity or strength through numbers, then they have a First Amendment right to do so (subject, of course, to any permitting and other requirements). Under this approach to freedom of assembly, no further explication of the specific content of OWS’s message would be required.¹¹⁸ This is a critical point, for public assemblies can often be disorganized, spontaneous, cacophonous, and incoherent.

In the context of the OWS demonstrations, we can more fully appreciate the value of a freestanding freedom of assembly. Thus, perhaps the most significant move Inazu makes in his volume turns out to be textual. By divorcing assembly and petition, he allows for the development of a distinct freedom of assembly. This freedom grants the people the right to be present in and to use certain public places. They may of course do so to speak or to petition government officials. But these activities and rights are distinct from the right to peaceably assemble.

Thus, a full recovery of the Assembly Clause clarifies the extent of the government’s trust obligation regarding public places under its control. It highlights the scope of the “easement” the people possess when they occupy

114. N.R. Kleinfeld & Cara Buckley, *Wall Street Occupiers, Protesting Till Whenever*, N.Y. TIMES, Sept. 30, 2011, <http://www.nytimes.com/2011/10/01/nyregion/wall-street-occupiers-protesting-till-whenever.html?pagewanted=all>.

115. INAZU, *supra* note 2, at 3.

116. Paul Krugman, Op-Ed., *Money and Morals*, N.Y. TIMES, Feb. 9, 2012, <http://www.nytimes.com/2012/02/10/opinion/krugman-money-and-morals.html?gwh=1B8B872410A8FFE1376708CD918AFF25>.

117. See Andrew Ross Sorkin, *Occupy Wall Street: A Frenzy That Fizzled*, DEALBOOK, N.Y. TIMES (Sept. 17, 2012, 8:51 PM), <http://dealbook.nytimes.com/2012/09/17/occupy-wall-street-a-frenzy-that-fizzled/> (lamenting the lack of a coherent message from the movement).

118. See INAZU, *supra* note 2 at 161–62 (observing that assembly itself is expression and multiple interpretations of an assembly are possible).

and use public forums.¹¹⁹ There has long been some level of discomfort relating to the idea that the First Amendment imposes an affirmative obligation on officials to provide space or other resources for the peoples' exercise of constitutional rights. However, if the First Amendment protects not only discrete activities like speech and petition, but also simple *presence* in public places, then it begins to look very much as if the First Amendment contemplates a degree of affirmative support. After all, assembly had to take place somewhere, and the most natural or obvious place would be something like a public square. Interpreting the Assembly Clause as an independent form of refuge for public dissent fortifies the argument that the First Amendment was intended, at least in part, to facilitate public presence and outdoor politics.

Indeed, recovery of the Assembly Clause might alter or clarify a number of First Amendment doctrines and principles relating to public protests, demonstrations, and other forms of outdoor politics like the OWS demonstrations. Let me highlight just a few examples.

The Supreme Court has attempted to explain how a parade with no clearly identifiable message nevertheless constitutes either a form of expressive conduct or an expressive association.¹²⁰ However, once the parade is properly characterized and analyzed as an assembly, courts need not attempt to interpret such gatherings. This insight applies to a variety of public gatherings. For example, where individuals have gathered in a public park for the purpose of feeding the homeless, the fact that no particularized message would be discernible to the public would not make any difference under the Assembly Clause.¹²¹ These and other unique but nonexpressive gatherings could find refuge under the Assembly Clause even if protection is not available to them under the Free Speech Clause or the expressive association doctrine.

The Court has also indicated that picketing on a public sidewalk near a person's residence may be entitled to less protection under the Free Speech Clause because the protesters did not seek to communicate with a broad public audience.¹²² That observation, and potential limitation, is simply irrelevant in the context of the freedom to peaceably assemble on a public sidewalk—the actual activity in question. Further, resort in some cases to the Assembly Clause, which by its terms protects a form of conduct, could reduce some of the considerable pressure the courts have placed on the speech–conduct distinction. Indeed, recovery of the Assembly Clause might

119. Harry Kalven, Jr., *The Concept of the Public Forum: Cox v. Louisiana*, 1965 SUP. CT. REV. 1, 13.

120. See *Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Boston, Inc.*, 515 U.S. 557, 568 (1995) (discussing the expressive nature of parades).

121. See *First Vagabonds Church of God v. City of Orlando*, 610 F.3d 1274, 1292 (11th Cir. 2010), *vacated*, 616 F.3d 1229 (11th Cir. 2010), *reinstated in part*, 638 F.3d 756 (11th Cir. 2011) (upholding permit requirements as applied to the feeding of homeless in public parks).

122. *Frisby v. Schultz*, 487 U.S. 474, 486–88 (1988).

at long last elevate demonstrating, marching, and labor picketing to the status of fully protected First Amendment activities rather than allowing them to be consigned to the lesser-protected rung of expressive conduct.¹²³

In many public protest cases decided after the 1960s, including several involving protests near abortion clinics, the Court has used free speech and time, place, and manner doctrines to examine the constitutionality of limits on public contention and dissent.¹²⁴ The primary concern in those cases was to what extent speakers should have a meaningful opportunity to engage with their intended audiences.¹²⁵ Indeed, in numerous contexts, courts have reviewed regulatory requirements that implicate freedom of assembly, including permit and insurance provisions, as if they affect only the freedom of speech.¹²⁶ However, these regulations may have separate and significant effects on assembly rights. Suppose that courts refocused the inquiry in such a way that assembly rather than speech became the primary concern. It is possible that something like the time, place, and manner doctrine would develop in this context. However, it is also possible that different considerations would lead to distinct doctrinal formulations and perhaps even to an expansion of public protest rights.

Let me return a final time to the OWS demonstrations. As noted earlier, the Assembly Clause contains a textual limitation. It recognizes a right “peaceably to assemble.” Inazu does not offer a definitive interpretation of this text. It is clear that the Assembly Clause does not protect riotous mobs. Certainly an assembly that engages in vandalism or violent acts can be suppressed. Further, under free speech doctrine authorities may impose basic limitations on public demonstrations for the purpose of ensuring public order and safety.¹²⁷

123. This would require revisiting statements by the Supreme Court in civil rights-era cases to the effect that the First Amendment provides less protection to acts such as assembly than it does to pure speech. See *Cox v. Louisiana*, 379 U.S. 559, 564 (1965) (suggesting that the freedom to peaceably assemble was linked to expression and inferior to its purest forms); *id.* at 555 (same). Justice Black had even less regard for marching, picketing, and parading. Although he often claimed to be a strict textualist, Justice Black was confident that the state could absolutely bar such activities on the public streets. *Id.* at 581 (Black, J., concurring).

124. See *Hill v. Colorado*, 530 U.S. 703, 719–20 (2000) (examining the constitutionality of a Colorado statute using free speech and time, place, and manner doctrines); *Madsen v. Women’s Health Ctr., Inc.*, 512 U.S. 753, 762–64 (1994) (same).

125. See *Hill*, 530 U.S. at 716 (“The right to free speech, of course, includes the right to attempt to persuade others to change their views, and may not be curtailed simply because the speaker’s message may be offensive to his audience.”); *Madsen*, 512 U.S. at 774 (“[I]t is difficult . . . to justify a prohibition on *all* uninvited approaches of persons seeking the services of the clinic, regardless of how peaceful the contact may be, without burdening more speech than necessary to prevent intimidation and to ensure access to the clinic.”).

126. See, e.g., *Thomas v. Chi. Park Dist.*, 534 U.S. 316, 326 (2002) (upholding permit requirement for activities in public parks as a valid regulation of speech).

127. See *Hill*, 530 U.S. at 713–14 (explaining that protecting the safety of individuals is a legitimate government interest).

The OWS demonstrations pressed the boundaries of these limits.¹²⁸ Insofar as OWS participants were unlawfully present in private spaces, let us assume that the freedom of assembly offered them no refuge. But in many cases, protesters sought to permanently occupy public forums and other public venues.¹²⁹ Were these “peaceable” assemblies? As I noted earlier, one could argue that the original or traditional understanding was that the Assembly Clause contemplated the formation and relatively brief presence of the people in public places. However, there is nothing in the Assembly Clause itself that suggests any kind of temporal limitation. There is nothing violent or unpeaceable about the mere act of assembly or even of occupation. So long as the occupation does not disrupt the flow of pedestrian or other traffic, violate any time restriction, or violate noise ordinances and the like, what basis is there for requiring the assembly to disperse?¹³⁰

It seems that at least two fundamental questions must be answered. The first, as I have already suggested, is whether we ought simply to incorporate all of the various time, place, and manner requirements that are not deemed generally to abridge freedom of speech¹³¹ into the assembly context. In that case, courts would likely equate “peaceably” with lawfully. This would essentially mean that in public places where individuals have a right to congregate, the freedom of assembly is coextensive with the freedom of speech. However, this would be inconsistent with recognition of a distinct and separate freedom of peaceable assembly. Second, and perhaps more fundamentally, we need to address whether the Assembly Clause provides some refuge for certain forms of civil disobedience.¹³² Since freedom of assembly was not seriously considered in the OWS litigation, the courts never reached these issues.

Like the outer bounds of group autonomy Inazu discusses, none of the foregoing issues has yet received any significant attention in connection with the Assembly Clause. If or once they do, however, we may find that the First

128. See *In re Waller v. City of New York*, 933 N.Y.S.2d 541, 545 (N.Y. App. Div. 2011) (holding that OWS failed to show a right to a temporary restraining order that would restrict the city’s ability to promote health and safety); James Barron & Colin Moynihan, *City Reopens Park After Protesters Are Evicted*, N.Y. TIMES, Nov. 15, 2011, http://www.nytimes.com/2011/11/16/nyregion/police-begin-clearing-zuccotti-park-of-protesters.html?pagewanted=all&_moc.semityn www (describing the events surrounding the denial of the order).

129. Joel Banner Baird, *To Be Occupied: Burlington’s City Hall Park*, BURLINGTON FREE PRESS, Oct. 25, 2011, at A1; Jimmy Vielkind, *A Permanent Occupation?*, TIMES UNION, Oct. 31, 2011, <http://www.timesunion.com/local/article/A-permanent-occupation-2243717.php>.

130. See El-Haj, *supra* note 102, at 578 (noting that, historically speaking, “the government was considered justified in restricting public assemblies only when they created public disorder, because only then were the assemblies no longer within the protection of the constitutional right”).

131. See *Hill*, 530 U.S. at 719–20 (upholding a content-neutral statute designed to protect the access and privacy of patients by prohibiting speech-related conduct within 100 feet of the entrance of any health care facility); *Madsen*, 512 U.S. at 762–64 (holding that certain restrictions imposed on antiabortion protesters were not directed at the content of speech, and thus were permissible as protecting the health and well-being of patients).

132. INAZU, *supra* note 2, at 167.

Amendment affords some additional measure of refuge for traditional forms of public protest and contention. Inazu's partial recovery of the Assembly Clause ought to motivate civil rights litigators, scholars, and courts to start thinking more carefully about assembly's implications in the more traditional contexts of public protest and demonstration.

Conclusion

Liberty's Refuge is an enlightening account of a First Amendment freedom that has for too long languished in the shadow of freedom of speech and under the weight of a judicially conceived right of expressive association. The Assembly Clause may never again be feted at something like a World's Fair. As Inazu shows, the more immediate impact of its recovery would be felt more locally. Private, nonconforming groups would gain a fuller measure of autonomy from a recovered freedom of assembly. In addition, as I have argued, individuals would enjoy the social and political benefits of a robust and recovered freedom of assembly. Finally, as I have also suggested, traditional public assemblies would occupy firmer constitutional ground. We owe a debt to Inazu for his exhumation of a once—and still—fundamental constitutional liberty. Inazu has invited us to participate in a conversation about a long-forgotten freedom, and has provided compelling reasons to accept this invitation. I look forward to reading his future work and to future discussions regarding the *recovered* freedom of assembly.

What Is the Essential Fourth Amendment?

MORE ESSENTIAL THAN EVER: THE FOURTH AMENDMENT IN THE TWENTY-FIRST CENTURY. By Stephen J. Schulhofer. New York, New York: Oxford University Press, 2012. 216 pages. \$21.95.

Reviewed by Christopher Slobogin*

I. Introduction

To the average American, the Fourth Amendment probably brings to mind a jumbled notion of warrants, probable cause, and exclusion of illegally seized evidence. Compared to the First Amendment, *Miranda*'s right to remain silent,¹ the jury trial guarantee,² and the Equal Protection Clause's prohibition on racial discrimination,³ the right to be secure from unreasonable searches and seizures is not well understood by most of the populace, either in its precise scope or its rationale.

Some confusion about specific Fourth Amendment prohibitions is tolerable and understandable. After all, it is the job of the police and judges, not Joe Q. Citizen, to apply search and seizure law, and even these government actors are more than occasionally flummoxed by the rules. Public ignorance about the Amendment's rationale is perhaps just as excusable, but it is much more unfortunate. People do not always understand why the law appears to prefer a judge's opinion over that of the streetwise cop, why a person who has nothing to hide should care about official surveillance, or why a person who does have something to hide should be able to exclude evidence of guilt because the police violated some arcane rule. As a result, citizens are often outraged by judicial opinions that free defendants on "technicalities,"⁴ and seldom are bothered by those court decisions—much more prevalent in the past several decades—that curtail liberty and privacy in the name of crime control and national security.

Stephen Schulhofer sees this as a problem, and in *More Essential Than Ever: The Fourth Amendment in the Twenty-First Century*⁵ he tries to redress it. Pitched toward a general audience rather than the legally trained, the book

* Milton Underwood Professor of Law, Vanderbilt University Law School.

1. *Miranda v. Arizona*, 384 U.S. 436, 444 (1966) ("Prior to any questioning, the person must be warned that he has a right to remain silent . . .").

2. U.S. CONST. amend. VI.

3. *Id.* amend. XIV, § 1.

4. See generally William A. Geller, *Is the Evidence in on the Exclusionary Rule?*, 67 A.B.A. J. 1642, 1645 (1981) (discussing the public policy debate over the exclusionary rule).

5. STEPHEN J. SCHULHOFER, *MORE ESSENTIAL THAN EVER: THE FOURTH AMENDMENT IN THE TWENTY-FIRST CENTURY* (2012).

provides a passionate defense of the “essential” Fourth Amendment that, as Schulhofer would have it, the Founders intended but the current Supreme Court has ignored. Much of what is said in this book will not be new to Fourth Amendment scholars. But the work’s straightforward eloquence provides a strong, popularized brief for interpreting the Fourth Amendment as a command that judicial review precede all nonexigent police investigative actions that are more than minimally intrusive. Schulhofer argues that this interpretation is not only consistent with the intent of the Framers, but remains a crucial means of discouraging government officials from harassing innocent people, promoting citizen cooperation with law enforcement efforts, and protecting the speech and association rights that are indispensable to a well-functioning democracy.⁶

Schulhofer’s liberal take on the Fourth Amendment is largely persuasive. This Review points out a few places where Schulhofer may push the envelope too far or not far enough. But, these quibbles aside, *More Essential Than Ever* is a welcome reminder for scholars and the public at large that the Fourth Amendment is a fundamental bulwark of constitutional jurisprudence and deserves more respect than the Supreme Court has given it.

II. Judicial Review as a Means of Protecting Privacy and Limiting Discretion

More Essential Than Ever is composed of eight chapters, the first two of which set up the rest of the book. Chapter 1 sketches out the thesis that was just described. In the course of doing so, Schulhofer describes his views on the core purpose of the Fourth Amendment. While he appears to accept the Supreme Court’s stance that the scope of the Fourth Amendment is defined primarily by reasonable expectations of privacy,⁷ he reminds us that the Amendment explicitly speaks not of privacy but of “the right of the people to be *secure* in their persons, houses, papers and effects.”⁸ Thus, he reasons, the Fourth Amendment is not about privacy in the sense of keeping secrets, but rather protects privacy as a means of ensuring people are secure in their ability to control information vis-à-vis the government.⁹ To the

6. See *id.* at 6 (“[The Fourth Amendment] offers a shelter from governmental intrusions that unjustifiably disturb our peace of mind and our capacity to thrive as independent citizens in a vibrant democratic society.”).

7. See *Kyllo v. United States*, 533 U.S. 27, 33 (2001) (stating that a Fourth Amendment search occurs if “the individual manifested a subjective expectation of privacy in the object of the challenged search,” and “society [is] willing to recognize that expectation as reasonable” (quoting *California v. Ciraolo*, 476 U.S. 207, 211 (1986))).

8. SCHULHOFER, *supra* note 5, at 7 (citation omitted) (emphasis in original).

9. See *id.* at 10 (arguing that it never would have occurred to Americans in the eighteenth century that “by entering into relationships with others, they had given *the government* unrestricted access to any information they revealed to trusted social and professional associates”). Schulhofer later clarifies that the Fourth Amendment is about “the right to control knowledge about our

argument that innocent people should have nothing to fear from law enforcement discovery of private information, especially when it can be discovered without physical intrusion, Schulhofer has the following riposte: “[S]urveillance can have an inhibiting effect on those who are different, chilling their freedom to read what they choose, to say what they think, and to join with others who are like-minded.”¹⁰ And when this occurs without justification, “[i]t undermine[s] politics and impoverish[es] social life for everyone.”¹¹

It has become fashionable to criticize the idea that Fourth Amendment search doctrine is meant to protect privacy. Critics claim that the Fourth Amendment is really about government power,¹² protecting property rights,¹³ or preventing coercion.¹⁴ But *all* of the guarantees in the Bill of Rights are about restricting government power. The Fourth Amendment focuses on protecting *particular* individual interests from *certain* types of government power, and Schulhofer is right that privacy, construed to mean control of information from unjustified government access, is the dominant focus of Fourth Amendment doctrine,¹⁵ at least as it applies to searches.¹⁶ The Fourth Amendment’s prohibition on unauthorized government monitoring of our activities, thoughts, and plans is a potent limit on official power that protects against trespass and official coercion but also protects against much more.

Chapter 2 provides a survey of the historical conflicts and cases that led to the Fourth Amendment. Schulhofer does a masterful job telling the story of the general warrant. He begins with the sagas of two Englishmen well-

personal lives, the right to decide how much information gets revealed to whom and for which purposes.” *Id.* at 130.

10. SCHULHOFER, *supra* note 5, at 13.

11. *Id.* at 14.

12. See, e.g., Paul Ohm, *The Fourth Amendment in a World Without Privacy*, 81 MISS. L.J. 1309, 1338 (2012) (“The new constitutional lodestar, power, is the Fourth Amendment’s third act [after property and privacy] Power seems to be the amendment’s essence, not merely a proxy for something deeper.”); Raymond Shih Ray Ku, *The Founders’ Privacy: The Fourth Amendment and the Power of Technological Surveillance*, 86 MINN. L. REV. 1325, 1326 (2002) (“The Fourth Amendment protects power not privacy.”).

13. Morgan Cloud, *A Liberal House Divided: How the Warren Court Dismantled the Fourth Amendment*, 3 OHIO ST. J. CRIM. L. 33, 72 (2005) (arguing for a Fourth Amendment “rooted in property theories” (emphasis added)).

14. William J. Stuntz, *The Substantive Origins of Criminal Procedure*, 105 YALE L.J. 393, 446 (1995) (contending that the Fourth Amendment is meant to limit “coercion and violence”).

15. See CHRISTOPHER SLOBOGIN, *PRIVACY AT RISK: THE NEW GOVERNMENT SURVEILLANCE AND THE FOURTH AMENDMENT* 23–26 (2007) (arguing that privacy is a central value protected by the Fourth Amendment).

16. Schulhofer confusingly supports his point about the importance of privacy in search cases by referring to cases involving seizures. SCHULHOFER, *supra* note 5, at 7 (describing cases involving the towing of a mobile home and arrests). Seizures are not governed by the expectation of privacy language used in search cases but rather are defined in terms of interference with property or movement. *Jacobsen v. United States*, 466 U.S. 109, 113 (1984) (seizure of property occurs when there is “some meaningful interference with an individual’s possessory interests”); *Florida v. Bostick*, 501 U.S. 429, 436 (1991) (holding that seizure of a person occurs when he would not “feel free to . . . terminate the encounter”).

known to Fourth Amendment scholars: John Wilkes, a member of Parliament whose office was ransacked by government officials seeking proof of seditious libel under a “nameless warrant,”¹⁷ and John Entick, also suspected of sedition, whose papers were seized pursuant to a warrant issued by an executive official rather than a judge and that failed to describe the items sought.¹⁸ Schulhofer also engagingly describes the hullabaloo in the colonies over the writs of assistance that allowed British officials to search any place they desired for evidence of unspecified offenses,¹⁹ and of course he includes an account of James Otis’s famous denunciation of the writs in 1761.²⁰ From this type of evidence, Schulhofer concludes that “there is no doubt that resistance to discretion lay at the heart” of the Fourth Amendment.²¹

Schulhofer is right about that. But he moves from that observation to the further conclusion that this resistance to the tyranny of every “common Officer” requires *ex ante* review by a judge for most searches and seizures.²² Making that connection takes more work. The *Entick* and *Wilkes* cases involved searches for and seizures of papers, and the writs of assistance were aimed primarily at customized goods held by colonial merchants. The Framers, mostly from the middle and upper classes, may not have cared very much about whether seizures of ordinary criminals and searches for evidence of “street crime” were anticipated by a warrant.²³ Schulhofer himself notes that warrantless arrests for routine felonies were permitted upon “reasonable cause”; that warrantless searches pursuant to arrest were routine; and that searches of ships, wagons, and other property outside the home at least “occasionally” took place without judicial authorization.²⁴ Even warrantless searches of homes occurred in colonial times.²⁵

So while the Framers hated the general warrant, they did not necessarily think specific warrants were or should be the primary means of regulating all types of government investigations. Schulhofer indirectly concedes this point,²⁶ but insists that modern-day resistance to executive discretion requires a preference for warrants even in situations in which they may not have been

17. SCHULHOFER, *supra* note 5, at 24–26.

18. *Id.* at 26–27.

19. *Id.* at 27–30.

20. *Id.* at 29.

21. *Id.* at 35.

22. *Id.* at 36.

23. Indeed, as Schulhofer points out, James Madison supported the Fourth Amendment because “he feared that popular majorities would enact legislation authorizing broad warrants, to the disadvantage of the new nation’s propertied elite.” *Id.* at 35.

24. *Id.* at 37.

25. Thomas Y. Davies, *Recovering the Original Fourth Amendment*, 98 MICH. L. REV. 547, 622 (1999) (stating that during the Framing Era “the initiation of arrests and searches commenced when a crime victim either raised the ‘hue and cry’ or made a sworn complaint,” although also noting that the hue and cry was probably relegated to “fresh” cases by the late eighteenth century).

26. SCHULHOFER, *supra* note 5, at 40–41.

required in colonial times.²⁷ He gives a number of reasons for this position, but the most prominent of them is the rise of organized police forces, aided by technological advances, that have vastly expanded government search and seizure capacity compared to that possessed by the lonely colonial constable.²⁸

More broadly, this huge shift in the relative power structure leads Schulhofer to argue for an analytic approach that focuses on original principles rather than original rules, which is an approach he dubs “adaptive originalism.”²⁹ On this last point, Schulhofer is in league with a number of scholars. For instance, Donald Dripps has recently argued that trying to tie modern rules to specific practices that existed in the eighteenth century makes no sense in a whole host of uniquely modern situations, including administrative searches, searches of private papers, investigative stops on less than probable cause, wiretapping, and the use of gunfire to effect the arrest of a fleeing felon.³⁰ Moreover, even the common law rules that can sensibly be applied today were in the process of changing in the eighteenth century and were not necessarily favored by the Framers.³¹ So, like Schulhofer, Dripps would ask whether and to what extent a search and seizure threatens “the priority of individual liberty and privacy, as against public security, that the founders aspired to.”³² The key question remains, however, whether adaptive or aspirational originalism requires the strong warrant requirement that Schulhofer favors.

III. A Critique of Modern Search and Seizure Rules

Chapters 3 through 7 of *More Essential Than Ever* try to answer that question. They address the Supreme Court’s jurisprudence in five general areas: the overarching rules governing searches and arrests; the special problems that arise in policing on the streets; the law governing administrative searches such as health and safety inspections, roadblocks and drug testing of school children; wiretapping and other electronic searches; and the dilemmas caused by national security concerns. The theme throughout these chapters is that, in generating current rules, the Supreme Court “has increasingly put police convenience above . . . original Fourth

27. See *id.* at 41 (arguing that though we should respect the Framers’ interpretations of searches and seizures under the Fourth Amendment, “that respect cannot take the form of an unreflective commitment to old rules that now have radically different effects in practice”).

28. See *id.* at 40 (arguing that eighteenth-century law enforcement was “a small, poorly organized, amateur affair, a far cry from the sizeable force of well-armed, full-time police who only a few years later became a constant presence on the streets of American cities and towns”).

29. *Id.* at 39–41.

30. See generally Donald A. Dripps, *Responding to the Challenges of Contextual Change and Legal Dynamism in Interpreting the Fourth Amendment*, 81 *MISS. L.J.* 1085 (2012) (proposing aspirational originalism).

31. *Id.* at 1089.

32. *Id.* at 1128.

Amendment priorities” and thus failed to curb sufficiently the executive branch’s discretion to invade privacy.³³

In Chapter 3, entitled “Searches and Arrests,” Schulhofer attacks the Court’s unwillingness to exclude evidence when police violate the rule governing no-knock entries,³⁴ driving home his point with descriptions of several incidents in which residents were killed or harmed when surprised by police.³⁵ He disagrees with the Court’s decisions allowing pretextual traffic stops and cajoled consents,³⁶ and partly as a way of undermining those decisions he appears to argue that the police should have to obtain a warrant for all nonexigent arrests, or at least for all nonexigent arrests for crimes that would have been misdemeanors at common law.³⁷ He also seems to think that warrants should be required for searches of cars in all but the most exigent circumstances, given the much-expanded use we make of vehicles in modern times.³⁸ Finally, he castigates two of the Court’s rationalizations for its retrenchment on the exclusionary rule—the increased professionalism of the police and the development of alternative remedies³⁹—by arguing that neither development has progressed far enough to justify the trust the Court places in law enforcement.⁴⁰ In Schulhofer’s mind, the suppression remedy is required in order to deter the police and ensure judicial integrity, and undercutting it as the Court has done breeds lawlessness.⁴¹

Chapter 4, “Policing Public Spaces,” tackles the special problems that arise in defining seizures of people and the scope of stop-and-frisk doctrine.⁴² In contrast to many commentators on the liberal end of the spectrum, Schulhofer would not reverse *Terry v. Ohio*,⁴³ the Court’s iconic case sanctioning stops and frisks on reasonable suspicion (a level of justification

33. SCHULHOFER, *supra* note 5, at 44.

34. *See generally* *Hudson v. Michigan*, 547 U.S. 586 (2006) (holding that violation of the knock-and-announce rule does not require exclusion of evidence seized as a result).

35. SCHULHOFER, *supra* note 5, at 46–47.

36. *Whren v. United States*, 517 U.S. 806, 817 (1996) (holding that the Fourth Amendment does not recognize pretext arguments when the police action is based on probable cause); *Schneckloth v. Bustamonte*, 412 U.S. 218, 249 (1973) (holding that individuals need not be told of their right to refuse consent).

37. *See* SCHULHOFER, *supra* note 5, at 52 (arguing that the common law exception permitting warrantless arrest for felonies “should be interpreted narrowly”).

38. Schulhofer states that “[m]ost Fourth Amendment experts find it hard to reconcile the warrant requirement for homes, suitcases, and paper bags with the no-warrant rule for cars,” and dismisses “the practical challenges involved in immobilizing cars on the roadside while waiting for a search warrant” by noting the availability of telephonic warrants. *Id.* at 57.

39. *Hudson*, 547 U.S. at 598–99.

40. *See* SCHULHOFER, *supra* note 5, at 67 (commenting that the premise that “executive officers can be trusted to exercise search-and-seizure powers fairly, in the absence of judicial oversight, is precisely the assumption that the Fourth Amendment rejects”).

41. *See id.* at 69 (“[T]he evidence shows that official disregard for fair procedure weakens public willingness to respect legal requirements and cooperate with law enforcement efforts to apprehend offenders.”).

42. *Id.* at 71–92.

43. 392 U.S. 1 (1968).

short of probable cause).⁴⁴ He states that “it is hard to imagine how the Court could have done better” in light of the need to give police flexibility in dealing with “fast-breaking police actions on the street.”⁴⁵ However, he believes that the Court’s subsequent application of *Terry* and related rules—ranging from declarations that seizures do not occur when police chase fleeing inner-city youth or confront factory workers and bus passengers⁴⁶ to its holding that reasonable suspicion exists when individuals in high-crime areas run from the police⁴⁷—“bears little relationship to social or psychological reality.”⁴⁸ These decisions, he argues, have acquiesced in the creation of racially tinged “police states” that “affect thousands of citizens every year, undermining their security, their respect for authority, their sense of acceptance in the wider community, and even their willingness to assist law enforcement efforts to control crime.”⁴⁹ He urges reversal of these decisions and commends the Court for striking down vagrancy laws that give police discretion to harass people pretextually.⁵⁰

Chapter 5, on “The Administrative State,” takes on the most difficult area of Fourth Amendment jurisprudence—searches and seizures that fall outside the paradigmatic investigation of street crime because they focus on garnering evidence for regulatory rather than criminal purposes (as with health and safety inspections of homes) or on special populations (such as drug testing of school children).⁵¹ In these situations the Court has either diluted the warrant requirement by permitting “area warrants” that are not based on individualized suspicion or has done away with the warrant and probable cause requirements altogether on the assumption that “special needs beyond those of ordinary law enforcement” are involved.⁵² Following the dissents in these cases, Schulhofer argues instead that departures from the judicial review requirement be permitted only when: (1) the objective of the government’s enforcement program is important; (2) normal investigative methods cannot achieve it; (3) the program is implemented through neutral

44. See SCHULHOFER, *supra* note 5, at 77 (arguing that the Court in *Terry* “established a pragmatic framework of relatively flexible powers in order to preserve police capacity to maintain order in public spaces”).

45. *Id.*

46. *United States v. Drayton*, 536 U.S. 194, 200 (2002) (holding that the Fourth Amendment is not implicated when police confront bus passengers and ask for consent to search their luggage); *California v. Hodari D.*, 499 U.S. 621, 626 (1991) (holding that chasing a fleeing person is not a seizure); *INS v. Delgado*, 466 U.S. 210, 218 (1984) (holding that questioning of factory workers is not a seizure).

47. *Illinois v. Wardlow*, 528 U.S. 119, 124–25 (2000).

48. SCHULHOFER, *supra* note 5, at 84.

49. *Id.* at 92.

50. See, e.g., *Chicago v. Morales*, 527 U.S. 41, 63–64 (1999) (striking down a statute criminalizing failure to disperse upon a police command).

51. SCHULHOFER, *supra* note 5, at 93–114.

52. See generally Eve Brensike Primus, *Disentangling Administrative Searches*, 111 COLUM. L. REV. 254 (2011).

criteria applicable to all; and (4) the primary purpose of the program is not “prosecutorial.”⁵³ Thus, for instance, Schulhofer believes the Court was correct in holding that a drug testing program aimed at political candidates was unconstitutional⁵⁴ (because the government interest was not substantial enough);⁵⁵ incorrect in upholding sobriety checkpoints,⁵⁶ suspicionless searches of probationers,⁵⁷ drug testing of students in nonathletic activities,⁵⁸ and spot inspections of junkyards for stolen parts⁵⁹ (because less intrusive investigative alternatives were available);⁶⁰ and correct in rejecting drug checkpoints⁶¹ and programs designed to test pregnant women for cocaine⁶² (because of their dominant prosecutorial purpose).⁶³ In contrast, health and safety inspections conducted according to neutral criteria⁶⁴ and airport checkpoints that monitor everyone do pass muster with Schulhofer.⁶⁵

“Wiretapping, Eavesdropping and the Information Age” is the title of Chapter 6. Schulhofer’s primary target here is the Court’s so-called “third-party doctrine,” which holds that when one knowingly exposes information to others one assumes the risk the government will acquire the information.⁶⁶ Relying on this rationale, the Court has concluded that the Fourth Amendment does not apply to government surveillance of travel on public roads and government acquisition of phone logs and bank records.⁶⁷ As have many others,⁶⁸ Schulhofer notes that under the Court’s third-party doctrine,

53. SCHULHOFER, *supra* note 5, at 97–98.

54. *Chandler v. Miller*, 520 U.S. 305, 323 (1997).

55. *See* SCHULHOFER, *supra* note 5, at 100–01 (praising the Court for assessing the significance of the State’s interest in drug testing political candidates and for determining that it was not substantial enough to outweigh the privacy interests at stake); *Chandler*, 520 U.S. at 318.

56. *Mich. Dep’t of State Police v. Sitz*, 496 U.S. 444, 455 (1990).

57. *Griffin v. Wisconsin*, 483 U.S. 868, 880 (1987).

58. *Bd. of Educ. v. Earls*, 536 U.S. 822, 838 (2002); *Vernonia Sch. Dist. 47J v. Acton*, 515 U.S. 646, 665 (1995).

59. *New York v. Burger*, 482 U.S. 691, 717 (1987).

60. SCHULHOFER, *supra* note 5, at 101.

61. *City of Indianapolis v. Edmond*, 531 U.S. 32, 48 (2000).

62. *Ferguson v. City of Charleston*, 532 U.S. 67, 76 (2001).

63. SCHULHOFER, *supra* note 5, at 108; *Ferguson*, 522 U.S. at 83.

64. *See Camara v. Mun. Court*, 387 U.S. 523, 534 (1967) (holding that the Fourth Amendment bars prosecution of a person who has refused to permit a warrantless code-enforcement inspection of his personal residence).

65. Schulhofer also appears to be comfortable with border searches and does not discuss checkpoints for licenses. SCHULHOFER, *supra* note 5, at 105. *Cf. Delaware v. Prouse*, 440 U.S. 648, 657 (1979) (permitting such checkpoints in dictum). Since these seizures might be said to have a dominant “prosecutorial purpose,” it is not as clear how they fare under his model.

66. *See generally* Orin S. Kerr, *The Case for the Third-Party Doctrine*, 107 MICH. L. REV. 561 (2009) (offering a defense of the often-criticized doctrine).

67. *Smith v. Maryland*, 442 U.S. 735, 745 (1979) (holding that there is no reasonable expectation of privacy in phone numbers dialed); *United States v. Miller*, 425 U.S. 435, 442 (1976) (holding that there is no reasonable expectation of privacy in information surrendered to banks).

68. *See* Erin Murphy, *The Case Against the Case for Third-Party Doctrine: A Response to Epstein and Kerr*, 24 BERKELEY TECH. L.J. 1239, 1239 (2009) (arguing against the “current

one cannot reasonably expect privacy from government discovery of information given to a third party even when the disclosure to that party occurs with the understanding it is confidential, is made for a specific purpose only, or is unavoidable if one wants to live in modern society.⁶⁹ Schulhofer's adaptive originalism leads him to reject this result.⁷⁰ He points out that "[t]he colonists who conferred with friends while planning the American revolution did not think that by sharing confidential information they had lost their right to exclude strangers,"⁷¹ and they certainly did not think they had thereby lost their right to exclude the government.⁷² Furthermore, he continues, the Court's equation of citizen or institutional third parties with government agents is nonsensical in the modern age.⁷³ Schulhofer points out that "we routinely deny government the power to pursue actions that are freely available to individuals"—such as practicing a particular religion—and, more importantly, "[t]he extraordinary resources available to the government give it unique power and unique potential to threaten the liberty and autonomy of individuals."⁷⁴

Thus, Schulhofer believes that the tracking of a car using a GPS device, as occurred in the recent case of *United States v. Jones*,⁷⁵ is a Fourth Amendment search that requires a warrant based on probable cause even when it is not effectuated by a trespass on the car⁷⁶ (the limitation on the definition of search endorsed by the majority in *Jones*).⁷⁷ He strongly endorses Justice Sotomayor's concurring opinion in that case voicing concern that even brief locational tracking can chill freedoms,⁷⁸ and he rejects the gist of Justice Alito's concurring opinion, which would apply the

configuration" of the third-party doctrine rule that holds that "information disclosed to third parties receives no Fourth Amendment protection").

69. SCHULHOFER, *supra* note 5, at 126–34.

70. See also Lawrence B. Solum, *Faith and Fidelity: Originalism and the Possibility of Constitutional Redemption*, 91 TEXAS L. REV. 147, 154 (2012) (noting that "[a]lmost all originalists agree that courts should view themselves as constrained by original meaning and that very good reasons are required for legitimate departures from that constraint").

71. SCHULHOFER, *supra* note 5, at 130.

72. *Id.*

73. See *id.* at 128–32 (critiquing the notion that citizens have the option of communicating by means other than the internet or telephone and arguing that those communications should be protected).

74. *Id.* at 136.

75. 132 S. Ct. 945 (2012).

76. See SCHULHOFER, *supra* note 5, at 139 (citing *Jones*, 132 S. Ct. at 960 (Alito, J., concurring)) (expressing agreement with Justice Alito's concurring opinion that the police tactics at issue in *Jones* were unacceptable interferences with privacy rights).

77. See *Jones*, 132 S. Ct. at 954 ("It may be that achieving the same result through electronic means, without an accompanying trespass, is an unconstitutional invasion of privacy, but the present case does not require us to answer that question.").

78. *Id.* at 956–57 (Sotomayor, J., concurring) (stating that "[a]wareness that the Government may be watching chills associational and expressive freedoms" and also stating "it may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties").

Fourth Amendment only to “prolonged tracking” and only as long as the public does not itself begin engaging in such tracking for convenience or security purposes.⁷⁹ Schulhofer would not always require a warrant when government seeks information from third parties or in every case of knowing exposure, however.⁸⁰ For instance, he endorses the practice of obtaining records via a subpoena, challengeable by the target.⁸¹ And even in the case of surveillance, Schulhofer would only dictate that a search has occurred when police use “technology that is not widely available,”⁸² suggesting that he believes nontechnological surveillance or surveillance with technology that is in “general public use” can escape Fourth Amendment regulation.⁸³

Chapter 7 deals with “The National Security Challenge,” a development that has threatened to undercut Fourth Amendment principles even further.⁸⁴ Schulhofer reminds us that we have come to deeply regret past overreactions to outside dangers and suggests we will similarly end up ruing post-9/11 phenomena such as the detentions in Guantanamo Bay, the Patriot Act’s sneak-and-peek warrants,⁸⁵ National Security Letters authorizing FBI agents to gather up any records that are useful in “criminal, tax, and regulatory matters,”⁸⁶ and the expansion of electronic surveillance powers under the Foreign Intelligence Surveillance Act.⁸⁷ To Schulhofer, these departures from the norm can actually have a negative effect on national security because they overwhelm the government with information, distract officials from more effective methods of protecting the country, and discourage cooperation by those groups in society most likely to have information about potential foreign threats.⁸⁸

79. *Id.* at 962–64 (Alito, J., concurring) (“New technology may provide increased convenience or security at the expense of privacy, and many people may find the tradeoff worthwhile . . . [or] reconcile themselves to this development as inevitable.”).

80. See *infra* notes 81–82 and accompanying text.

81. SCHULHOFER, *supra* note 5, at 134.

82. See *id.* at 142 (noting that “no one suggests that government data mining should be prohibited altogether” and that the Fourth Amendment is only intended to “assure that invasive methods of investigation are subject to oversight”).

83. The “general public use” nomenclature comes from dictum in *Kyllo v. United States*, 533 U.S. 27, 34 (2001).

84. SCHULHOFER, *supra* note 5, at 144–69.

85. 18 U.S.C. § 3103(a) (2006). Schulhofer would not object to all sneak-and-peek warrants, however. See SCHULHOFER, *supra* note 5, at 48.

86. 12 U.S.C. § 3414(a)(5)(A) (2006); 18 U.S.C. § 2709(b) (2006).

87. 50 U.S.C. §§ 1801–1885c (2006 & Supp. IV 2011).

88. See SCHULHOFER, *supra* note 5, at 168 (arguing that “[p]roposals . . . to relax Fourth Amendment requirements and ‘trade-off’ liberty for security . . . make counterterrorism efforts more difficult, not less”). He goes on to discuss the ways in which Muslim Americans are less likely to cooperate with authorities if they believe the police are targeting their communities without explanation. *Id.*

IV. A Critique of the Critique

Schulhofer makes a compelling case for privacy as the linchpin of Fourth Amendment protection and for making *ex ante* review of police search and seizure decisions the default regulatory stance. Also persuasive is his position that the Amendment should be viewed as a crucial means of preserving democracy, encouraging diversity of views, and promoting citizen respect for and cooperation with police work. Finally, adaptive originalism makes eminent sense in a country with a strong foundational document that is over two hundred years old. In short, I am in agreement with the broad strokes of the book. I'm not as sure about all the particulars.

For instance, many vibrant Western democracies have been able to control their police without the draconian remedy of exclusion.⁸⁹ Contrary to Schulhofer's assertion,⁹⁰ routine suppression of evidence found through a Fourth Amendment violation probably *delegitimizes* the legal system in the eyes of most citizens,⁹¹ and thus may contribute to the dissatisfaction with government that Schulhofer wants to avoid. Furthermore, in many situations—for instance, the violence and property damage that sometimes accompany illegal no-knock entries—monetary restitution is a more commensurate response than exclusion of evidence, as well as more satisfying when the victim of such acts is innocent of the crime and thus cannot resort to exclusion. Properly constructed, an action for damages⁹²—the only remedy for illegal searches available in colonial times⁹³—is more likely to accomplish all of the goals Schulhofer seeks: respect for government (because it punishes the true perpetrators of the illegality, not the prosecutor); deterrence of misconduct (especially in pretextual traffic and suspect drug possession cases, which wallet-conscious police will decide are not worth pursuing); improved professionalism (resulting from police departments literally having to pay the cost of bad training); and greater use of warrants (which police will realize immunizes them from liability).⁹⁴ While Schulhofer argues that an effective damages remedy would foreclose

89. See generally Craig Bradley, *Mapp Goes Abroad*, 52 CASE W. RES. L. REV. 375 (2001) (recounting resistance to, or significant limitations on, the exclusionary remedy in Europe, Australia, and Canada).

90. See SCHULHOFER, *supra* note 5, at 69 (arguing that “judicial tolerance for Fourth Amendment violations” creates problems for law enforcement because it “discourages law-abiding citizens from offering the cooperation needed to catch and convict offenders in future cases”).

91. As Schulhofer admits, “Fourth Amendment requirements often garner little public support [because] [t]hey seem like a gift to those bent on wrongdoing.” *Id.* at 171.

92. See, e.g., 42 U.S.C. § 1983 (2006) (providing a civil action for the deprivation of constitutional rights); *Bivens v. Six Unknown Agents of Fed. Bureau of Narcotics*, 403 U.S. 388, 390–97 (1971) (recognizing an action for damages when a plaintiff's injuries resulted from federal agents' violation of the Fourth Amendment).

93. See SCHULHOFER, *supra* note 5, at 67 (“[J]udicial oversight originally did not involve an exclusionary rule; the deterrent to an illegal search was the victim's ability to sue for damages”).

94. See generally Christopher Slobogin, *Why Liberals Should Chuck the Exclusionary Rule*, 1999 U. ILL. L. REV. 363, 445–46 (summarizing the advantages of a damages remedy).

just as many prosecutions as the exclusionary rule, he may be wrong on that score;⁹⁵ in any event, a damages remedy would not flaunt the costs of the Fourth Amendment in the delegitimizing way the rule does, or involve judges, lawyers, and juries in trials they know are charades. As an alternative to attacking police abuse of discretion on the street by vastly reducing arrests for minor crimes (which is the effect of Schulhofer's more stringent arrest warrant requirement), the exclusionary remedy might best be reserved in such cases for evidence not related to the purpose of the search and seizure, a move that should maximize deterrence of pretextual actions and spurious consents.⁹⁶

The procedural justice literature upon which Schulhofer relies to make many of his arguments may also undercut some of his conclusions, especially in connection with regulation of large-scale crime-control efforts.⁹⁷ Schulhofer is right that parts of our cities, especially those occupied by minority groups, mimic police states, and the Court's willingness to blink at this state of affairs is outrageous, as well as complicit in discouraging cooperation with the authorities. At the same time, these communities are rife with crime, and their efforts to deal with that problem—through appropriately limited loitering statutes, camera surveillance, drug checkpoints, and the like—should not be foreclosed when they are the product of local democratic deliberations.⁹⁸ After all, the Framers themselves passed statutes permitting suspicionless inspections and searches, some of which were aimed at obtaining evidence of crime.⁹⁹ The principal defect of most of the administrative search and seizure cases heard by the Supreme Court to date is that they involved ad hoc programs established by the executive branch.¹⁰⁰ If instead authorization from a representative legislative body is required, if the legislation does not single out a discrete

95. *Id.* at 444 (“With an effective deterrent in place, police who lack probable cause will not necessarily give up; the more reasonable assumption is that they will simply get more cause.”).

96. Ricardo J. Bascuas, *Lessons from the Highway and the Subway: A Principled Approach to Suspicionless Searches*, 38 RUTGERS L.J. 719, 787–90 (2007) (making this argument).

97. Schulhofer's most explicit work on this subject is Stephen J. Schulhofer et al., *American Policing at a Crossroads: Unsustainable Policies and the Procedural Justice Alternative*, 101 J. CRIM. L. & CRIMINOLOGY 335 (2011).

98. See Tracey L. Meares, *Norms, Legitimacy and Law Enforcement*, 79 OR. L. REV. 391, 410–13 (2000) (using loitering statutes to illustrate the importance of involving the community in devising effective law enforcement strategies in order to enhance legitimacy).

99. See Fabio Arcila, *The Death of Suspicion*, 51 WM. & MARY L. REV. 1275, 1304–10 (2010) (discussing various Revolutionary period statutes that permitted suspicionless searches).

100. See, e.g., *Ferguson v. City of Charleston*, 532 U.S. 67, 71–72 (2001) (scrutinizing a policy authorizing drug testing of pregnant women formulated by hospital officials and local police); *Mich. Dep't of State Police v. Sitz*, 496 U.S. 444, 447 (1990) (involving a highway sobriety checkpoint established by the police department); *City of Indianapolis v. Edmond*, 531 U.S. 32, 34–35 (2000) (reviewing a drug roadblock established by local police); *New York v. Burger*, 482 U.S. 691, 693–94 (1987) (examining a junkyard inspection program established by legislation but providing no limits on police discretion); *Skinner v. R'y Labor Execs. Ass'n*, 489 U.S. 602, 608–12 (1989) (analyzing a drug testing program for railway workers authorized by legislation that provided no standards for implementation).

and insular minority, and if it is implemented in a nondiscriminatory fashion (e.g., across-the-board or randomly), a better balance between crime control and individual rights might be achieved.¹⁰¹ Nullification of such legislation probably would have more community-denigrating effects than the Court's current jurisprudence.

The same types of points can be made about national security surveillance endeavors, often aimed at accumulating information about thousands or hundreds of thousands of people (virtually all of whom are innocent of any wrongdoing).¹⁰² If, before voting, legislators are required to imagine application of these programs to themselves and all of their constituents, they are not likely to approve 1984-type laws, as evidenced by Congress's resistance to post-9/11 efforts to expand wiretapping authority¹⁰³ and its defunding of the infamous Total Information Awareness data-mining program.¹⁰⁴ And while courts are capable of figuring out when the legislative process is defective or when the police are unfairly implementing a legislatively authorized program, they are not equipped to make the nuanced determination, required by Schulhofer's approach, as to which law enforcement techniques are the most effective, least intrusive, most feasible means of achieving government aims.¹⁰⁵ Schulhofer's added stipulation that prosecution not be the dominant purpose of these programs has the ironic consequence, as he acknowledges, of providing more privacy protection for those who may be engaged in criminal activity than those who are not.¹⁰⁶

Conversely, when law enforcement has targeted a specific individual, whether for prosecutorial or other reasons, the legislative process cannot work and judicial review before the search and seizure takes place is crucial. For this reason, Schulhofer's disdain for the third-party and knowing-exposure doctrines, which often work to vitiate *ex ante* review, is well-grounded. What is not as clear is why he would require probable cause for technologically sophisticated tracking of any length while permitting the government to obtain bank, credit card, and phone records with a subpoena (which at most requires a showing that the records are somehow relevant to

101. This approach, based on political process theory, was first proposed by Richard Worf in *The Case for Rational Basis Review of General Suspicionless Searches and Seizures*, 23 *TOURO L. REV.* 93, 197–98 (2007), and is developed further in Christopher Slobogin, *Government Dragnets*, 73 *LAW & CONTEMP. PROBS.* 107, 143 (2010).

102. See, e.g., Timothy B. Lee, *House Approves Another Five Years of Warrantless Wiretapping*, *ARS TECHNICA* (Sept. 12, 2012), <http://arstechnica.com/tech-policy/2012/09/house-approves-another-five-years-of-warrantless-wiretapping> (reporting on the FISA Amendment Act's goal of intercepting American citizens' international communication).

103. SCHULHOFER, *supra* note 5, at 158–59.

104. Department of Defense Appropriations Act of 2004, *PUB. L. NO. 108-87*, § 8131, 117 *Stat.* 1054, 1102 (2003).

105. See Slobogin, *supra* note 101, at 127–29 (explaining that while the Court can engage thoughtfully in strict scrutiny analysis in various contexts like time, place, and manner restrictions on speech, it is ill-equipped to analyze the efficacy and necessity of law enforcement techniques).

106. SCHULHOFER, *supra* note 5, at 95–96.

an investigation),¹⁰⁷ or why he would leave entirely unregulated even long-term surveillance with the naked eye or with generally available technology.¹⁰⁸ In terms of intrusiveness and the chilling effect on innocent activity—Schulhofer’s concerns—record acquisition would seem at least as intrusive as tracking.¹⁰⁹ Further, tracking with a GPS would seem to be no more inimical to these interests than monitoring travels with the human senses or technology in general public use.¹¹⁰ An alternative would be to permit both accessing of single-transaction records and short-term tracking—whether the police use naked-eye observation, primitive technology, or sophisticated devices—on reasonable suspicion, while requiring probable cause for acquisition of records containing substantial personal information and more prolonged surveillance.¹¹¹

It is also not clear how Schulhofer would treat undercover investigations, since he does not discuss the relevant case law in the book. Perhaps he would analogize this popular law enforcement technique to naked-eye and low-tech surveillance, in which case, consistent with Supreme Court decisions on the issue, it would be unregulated by the Fourth Amendment.¹¹² But the ability of undercover agents to insinuate themselves into personal lives can often result in much more intrusion than even long-term tracking, and thus ought to require at least as much justification (as the eighteenth-century disdain for undercover “thief-takers” suggests).¹¹³ Only

107. *United States v. R. Enters., Inc.*, 498 U.S. 292, 301 (1991) (stating that a subpoena should only be quashed on irrelevance grounds when “there is no reasonable possibility that the category of materials the government seeks will produce information relevant to the general subject of the grand jury’s investigation”); *United States v. Morton Salt*, 338 U.S. 632, 652 (1950) (stating that administrative subpoenas meet constitutional requisites even if they are meant only to satisfy “nothing more than official curiosity”).

108. Indeed, Schulhofer’s primary concern with data mining appears to be, not its breadth, but its use of technology not widely available to the public. *See* SCHULHOFER, *supra* note 5, at 142 (making the use of “technology that is not widely available” a critical element of a “search” under the Fourth Amendment).

109. *Cf.* Christopher Slobogin & Joseph E. Schumacher, *Reasonable Expectations of Privacy and Autonomy in Fourth Amendment Cases: An Empirical Look at “Understandings Recognized and Permitted by Society,”* 42 DUKE L.J. 727, 737 (1993) (reporting data indicating that perusal of bank records is considered more intrusive, by a significant margin, than tracking a car).

110. Schulhofer notes that, at common law, public movements were not considered private. SCHULHOFER, *supra* note 5, at 123. But research indicates that “conspicuously” following someone down the street is viewed as fairly intrusive, albeit not as intrusive as technological tracking of a car for three days. SLOBOGIN, *supra* note 15, at 112.

111. These points are developed further in Christopher Slobogin, *Making the Most of United States v. Jones in a Surveillance Society: A Statutory Implementation of “Mosaic Theory,”* DUKE J. CONST. L. & PUB. POL’Y (forthcoming 2012) and Christopher Slobogin, *Is the Fourth Amendment Relevant in a Technological Age?*, in CONSTITUTION 3.0: FREEDOM AND TECHNOLOGICAL CHANGE 11 (Jeffrey Rosen & Benjamin Wittes eds., 2011).

112. *See, e.g., Hoffa v. United States*, 385 U.S. 293, 302–03 (1966) (holding that the Fourth Amendment does not apply to evidence voluntarily disclosed to an informant).

113. *See* JOHN H. LANGBEIN ET AL., *HISTORY OF THE COMMON LAW: THE DEVELOPMENT OF ANGLO-AMERICAN LEGAL INSTITUTIONS* 677–81 (2009) (describing police and jury distrust of

when the third party is neither an agent of the government nor an impersonal entity like a bank should the third-party doctrine permit government to acquire the third party's information without any Fourth Amendment justification. In other words, the Fourth Amendment would be inapplicable in third-party scenarios only when the third party is independent of the government and can be said to possess a right (as an autonomous being) to disclose to the government any information he or she sees fit to reveal.¹¹⁴

Undoubtedly, Professor Schulhofer would have responses to all of these points. In any event, all of them only attack his thesis at the edges, without disturbing the crucial attributes of the Fourth Amendment's principles that he articulates and defends. *More Essential Than Ever* successfully captures the essence of the Fourth Amendment in a way that should bring home to everyone—not just lawyers and judges, but the “I've got nothing to hide” crowd, the “inner-city folks are all criminals” crowd, and the “government can be trusted” crowd—why it is so important.

thief-takers, who received rewards for turning in thieves that they often enticed into engaging in theft).

114. See Mary I. Coombs, *Shared Privacy and the Fourth Amendment, or the Rights of Relationships*, 75 CALIF. L. REV. 1593, 1643 (1987) (arguing that people in possession of information about others, even information that is “private” and obtained through an intimate relationship, have an “autonomy-based right to choose to cooperate with the authorities”).

Notes

Setting Examples, Not Settling: Toward a New SEC Enforcement Paradigm*

I. Introduction

Imagine a world where the only punishment for breaking the law is the payment of a negotiated fine. Imagine further that after paying this fine, you do not even have to admit to being guilty of the crime, no matter how staggering the evidence arrayed against you might be. Add to this the fact that your lawyer, negotiating the fine, used to work for the government entity prosecuting the crime. What is more, the government prosecutor sitting across the table from you will, in a few years time, want a job with the law firm now defending you. Finally, indulge in one more flight of fancy: imagine that you can pay these fines with other people's money. Now, ask yourself, exactly how much of a law-abiding citizen would you be?¹

Sadly, this is not a premise for a dystopian, science-fiction movie. This is the current enforcement paradigm of the Securities and Exchange Commission (the Commission), the federal agency predominately tasked with writing, overseeing, and enforcing the rules and regulations that govern the country's financial markets.² Over the last few decades, the once-feared Division of Enforcement—the law enforcement wing of the Commission—

* I would like to express my eternal gratitude to my family for continuing to believe in me far past the point when it became irrational to do so. Mom, I am probably never going to win an Oscar, so this is the best thank you that you are going to get. I would also like to thank the Volume 91 Notes Office—Monica Hughes, Lauren Ross, and Michael Selkirk—for their fabulous editing, and more importantly, for being fabulous people. Finally, I would like to thank our Editor in Chief, Parth Gejji. Helming a law review is a thankless job that exacts its toll in flesh and tears, but we are all better for the fact that you elected to do it.

1. Follow-up question: How often do you speed? While it is true that state and U.S. Attorneys can and do impose criminal sanctions on certain defendants, the vast majority of violators of securities laws are punished, if at all, via civil sanctions. See Eric Lichtblau, *Federal Cases of Stock Fraud Drop Sharply*, N.Y. TIMES, Dec. 24, 2008, <http://www.nytimes.com/2008/12/25/business/25fraud.html> (discussing a general decline in criminal prosecutions). There are many practical reasons for this, of course, the most obvious being the burden of proof. But that does not change the fact that our securities laws are primarily enforced via civil sanctions—and therefore through settlements. See Peter J. Henning, *Two More Setbacks in Securities Fraud Cases*, DEALBOOK, N.Y. TIMES (July 2, 2010, 9:30 AM), <http://dealbook.nytimes.com/2010/07/02/two-more-setbacks-in-securities-fraud-cases/> (explaining the difficulties of bringing a criminal securities fraud claim).

2. See *The Investor's Advocate: How the SEC Protects Investors, Maintains Market Integrity, and Facilitates Capital Formation*, U.S. SEC. & EXCHANGE COMMISSION, <http://www.sec.gov/about/whatwedo.shtml> (last modified July 30, 2012) (listing the responsibilities of the agency as issuing and amending rules, coordinating U.S. securities regulations, and enforcing those regulations).

has slowly warped into a meek and abiding Division of Settlement.³ This is a change that has and will continue to have sweeping consequences on the effectiveness of U.S. financial regulations, and through them, the stability of the overall global financial markets.

As a partial riposte to this rigged game—where corporate defendants violate the securities laws, are civilly sanctioned for these violations, and then proceed to violate the laws again in a veritable merry-go-round of fraud—the Honorable Jed S. Rakoff of the Southern District of New York rejected a proposed settlement and consent decree that the Commission had entered into with Citigroup Global Markets on November 28, 2011.⁴ In this unexpected and controversial move,⁵ Judge Rakoff took the Commission to task—not for the first time—calling such deals “unfair,”⁶ “unreasonable,”⁷ “suggest[ing] a rather cynical relationship between the parties,”⁸ “done at the expense . . . of the truth,”⁹ “worse than mindless [and] inherently dangerous,”¹⁰ aimed at “suppressing or obscuring the truth,”¹¹ “designed to provide the S.E.C. with the façade of enforcement,”¹² and perhaps most damningly, “engine[s] of oppression.”¹³ At the heart of his criticism, though, was Judge Rakoff’s belief that such settlements—which in this case imposed a \$285 million fine on Citigroup—amounted not to a sanction but to the imposition of a “cost of doing business” on the defendant, a mere inconvenience, and therefore failed to further the public interest in affecting tangible regulatory enforcement.¹⁴ This Note addresses Judge Rakoff’s concerns. While it is true that the Second Circuit will likely overturn Judge Rakoff’s decision in a matter of months,¹⁵ owing to the high levels of deference traditionally granted to agency determinations, his opinion

3. See Jean Eaglesham, *Weighing SEC’s Crackdown on Fraud*, WALL ST. J., April 11, 2012, <http://online.wsj.com/article/SB10001424052702304587704577333683615866446.html> (describing critics’ complaints that SEC settlements are “weak” and “need scrutiny”).

4. U.S. SEC v. Citigroup Global Mkts. Inc., 827 F. Supp. 2d 328, 335 (S.D.N.Y. 2011).

5. See M. Todd Henderson, *Impact of the Rakoff Ruling: Was the Judge’s Scuttling of the SEC/BofA Settlement Legally Pointless or Incredibly Important—or Both?*, WALL ST. LAW., Nov. 2009, at 1, 4 (calling the decision “a potential watershed moment”); Robert Khuzami, *Public Statement by SEC Staff: Court’s Refusal to Approve Settlement in Citigroup Case*, U.S. SEC. & EXCHANGE COMMISSION (Nov. 28, 2011), <http://www.sec.gov/news/speech/2011/spch112811rk.htm> (arguing that settlements like the Citigroup one had been “repeatedly approved for good reason by federal courts across the country”).

6. SEC v. Bank of Am. Corp., 653 F. Supp. 2d 507, 510 (S.D.N.Y. 2009).

7. *Id.*

8. *Id.* at 512.

9. *Id.*

10. U.S. SEC v. Citigroup Global Mkts. Inc., 827 F. Supp. 2d 328, 335 (S.D.N.Y. 2011).

11. *Id.*

12. *Bank of Am. Corp.*, 653 F. Supp. 2d at 510.

13. *Citigroup Global Mkts.*, 827 F. Supp. 2d at 335.

14. *Id.* at 333–34.

15. See U.S. SEC v. Citigroup Global Mkts. Inc., 673 F.3d 158, 160, 169 (2d Cir. 2012) (per curiam) (granting a stay against Judge Rakoff’s order and granting deference to the SEC’s conception of public interest).

nevertheless raises fundamental questions about the soundness of the Commission's settlement policy—questions that demand answers.

The Commission currently settles, in a manner not unlike the scenario described above, roughly 98% of its cases.¹⁶ This amounts to between 650 and 700 settlements per year.¹⁷ This means that the Commission settles cases at a higher rate than private parties¹⁸—even though the Commission's goal is not to vindicate its own private interests, as settlement often does, but to vindicate the interests of the public at large.¹⁹ The numbers themselves, then, suggest that there is something foundationally askew.

These settlements—termed consent judgments or consent decrees because the defendant is 'consenting' to an order being entered against it—have three constituent parts. First, the defendant typically agrees to pay a monetary fine and to disgorge any ill-gotten gains.²⁰ Second, the defendant agrees to the issuance of an injunction barring the defendant from violating the securities laws again in the future.²¹ Third, without admitting the allegations, the defendant agrees to an injunction barring it from denying those allegations in the future.²² While these three points may seem comprehensive, there is an elemental flaw with this system: namely, that it fails in its primary purpose. It fails to adequately deter violations of the securities laws.²³ The monetary sanctions, by themselves, are not sufficiently punitive to deter wrongdoing, especially when the defendant is a well-capitalized financial institution.²⁴ The brunt of a monetary sanction on a publicly traded corporation falls on shareholders, and to a lesser extent

16. The SEC settles roughly 650 to 700 cases a year. See ELAINE BUCKBERG ET AL., NERA ECON. CONSULTING, SEC SETTLEMENT TRENDS: 2H11 UPDATE 5 (2012), available at http://www.nera.com/nera-files/PUB_SEC_Trends_2H11_0612.pdf (listing the numbers for the last nine years, including that there were 682 settlements in 2011). The SEC prosecutes, at trial, roughly 14 cases a year. Jesse Eisinger, *Needed: A Cure for a Severe Case of Trialphobia*, THE TRADE, PROPUBLICA (Dec. 14, 2011, 4:10 PM), <http://www.propublica.org/thetrade/item/needed-a-cure-for-a-severe-case-of-trialphobia>. This proportion roughly corresponds to my own research, which shows that of the 44 cases filed in the Southern District of New York in 2008 that have been resolved, 38 were settled with consent judgments. My numbers ignore the entire pool of cases that are settled via administrative, rather than legal, means.

17. BUCKBERG ET AL., *supra* note 16, at 5.

18. See Marc Galanter & Mia Cahill, "Most Cases Settle": *Judicial Promotion and Regulation of Settlements*, 46 STAN. L. REV. 1339, 1339–40 (1994) ("Of-cited figures estimating settlement rates between 85 and 95 percent are misleading . . .").

19. See *The Investor's Advocate*, *supra* note 2.

20. *E.g.*, SEC v. Quadrangle Grp. LLC, Litigation Release No. 21,487, 98 SEC Docket 1088, 1088 (Apr. 15, 2010) (ordering defendants to pay a \$5,000,000 civil penalty).

21. *E.g.*, *id.*

22. *E.g.*, SEC v. Quadrangle Grp. LLC, No. 10-CV-1392, 2010 WL 1506633 (S.D.N.Y. Apr. 14, 2010).

23. See, *e.g.*, Submission of the SEC Addressing the Issues Identified in the Court's May 19, 2003 Order Concerning the Proposed Settlement of the Commission's Monetary Claims Against WorldCom, SEC v. Worldcom, Inc., Civ. No. 02-CV-4963 (JSR) (S.D.N.Y. June 6, 2003) (contending that the primary purpose of the Commission's penalties is to deter fraud).

24. See *infra* notes 122–126 and accompanying text.

creditors, not on the officers or executives who perpetuated the fraud.²⁵ Meanwhile, the injunctive relief barring future violations is toothless, or at least riddled with cavities, because the Commission almost never enforces its injunctions despite evidence of a raft of repeat offenders.²⁶

The shortcomings of this all-hands-on-deck settlement regime have broad ramifications. Over the last thirty years, the United States has experienced a virtually perpetual cycle of financial panics caused, in a not insignificant manner, by the violators of securities laws who were ineffectively deterred and detected by the Commission until it was too late.²⁷ The still-present and still-haunting Global Financial Crisis (GFC) is only the most recent example of this: Citigroup, Goldman Sachs, Merrill Lynch, *et al.*—all of the prime movers of the financial crisis were also the usual suspects for financial foul play in the decades preceding the crisis. Each of them had been repeatedly sanctioned for securities laws violations by the Commission prior to the GFC.²⁸ Little good it did. Rather than paying for the true, societal cost of their transgressions, these firms simply paid off the regulators in managed settlements.²⁹

The thesis of this Note is that the Commission must structurally rethink the way it enforces securities laws because the current system is radically inadequate. Part II discusses the history of this system. Part III details the incentives behind the settlement regime, and Part IV documents its rabid and uncompromising failures on both a theoretical and empirical level. Part V suggests two solutions—bringing more cases to trial and imposing individual, rather than corporate, liability—as methods for bringing securities laws' enforcement back to deterrence equilibrium, where the harm of violating the laws actually compares to the expected costs of those violations. Only at such an equilibrium can the public reasonably expect to see the number of financial frauds decrease to an acceptable level.

A final, important distinction needs to be elucidated. It is an oft-cited criticism of the Commission that it brings cases against corporations and not individuals.³⁰ In the aggregate, the empirical evidence does not bear this out. Of the 682 cases that the SEC settled in 2011, 484 of them named individuals as defendants.³¹ However, one place where this is patently not true is in the

25. *See infra* Part IV.

26. *See infra* notes 139–145 and accompanying text.

27. *See infra* Part II.

28. *See infra* notes 141–145 and accompanying text.

29. *See infra* notes 103–104 and accompanying text.

30. *See, e.g.*, Andrew Ackerman & Jean Eaglesham, *SEC Pushes to Toughen Penalties for Offenders*, WALL ST. J., Nov. 30, 2011, <http://online.wsj.com/article/SB10001424052970204262304577068281927469216.html> (mentioning a call for more suits of individuals by Sen. Grassley); Stavros Gadinis, *The SEC and the Financial Industry: Evidence from Enforcement Against Broker-Dealers*, 67 BUS. LAW. 679, 682 (2012) (finding that, in actions against large broker-dealers, the SEC targets an institution's corporate entity rather than its employees or executives).

31. BUCKBERG ET AL., *supra* note 16, at 5.

Commission's dealings with, and prosecutions of, large commercial and investment banks and financial institutions. Individuals who work at these white-shoe institutions are rarely, if ever, targeted for prosecution, even if the Commission brings suits against their employers.³² Since the stability of these institutions has the greatest impact on our financial system, and because the law-abidingness of an institution often reflects its stability, the focus of this Note will be on the securities laws and how they are applied to large financial institutions. Before evaluating how the securities laws and these institutions interact, however, it is important to understand the history of the federal securities laws.

II. A (Brief) History of (Modern) U.S. Securities Laws

There are a few recurrent truths about the evolution of the Securities and Exchange Commission. One is that its history has been plagued by a series of repeated failures to deter and discover financial fraud. Another is that the public outcry following each failure has materialized through an increase in the amount of fines the Commission is authorized to levy. This cyclical pattern has coincided with a linear one: a consistent trend towards settlement and cooperation with the entities the Commission is tasked with regulating. While these trends seem like they should be in conflict with each other, just the opposite is true. The forces are procyclical. The fines become higher, so the Commission becomes more and more comfortable with settling under the mistaken belief that the high fines are punishment enough.³³ At the same time, the regulated entities continue to be willing to accept these fines because, even at their pinnacle, they still represent little more than a tax on the overall cost of doing business.³⁴ Yet despite the evidence that these fines have almost no deterrent effect—the Commission's success in extracting even exorbitant fines has failed to prevent or even curb a repeated pattern of financial abuse by a repeated group of abusers—the response from the Commission and Congress has always been the same: more and higher fines.

32. Despite the fact that Goldman Sachs agreed to pay \$550 million and admitted making “mistakes” as part of its settlement with the SEC, no charges were brought against any high-level employees at Goldman Sachs—only against a low-level trader, Fabrice Tourre. See SEC v. Goldman, Sachs & Co., Litigation Release No. 21,489, 98 SEC Docket 1192, 1192 (Apr. 16, 2010) (detailing allegations in the consent judgment); see also Edward Wyatt, *Promises Made, and Remade, by Firms in S.E.C. Fraud Cases*, N.Y. TIMES, Nov. 7, 2011, http://www.nytimes.com/2011/11/08/business/in-sec-fraud-cases-banks-make-and-break-promises.html?_r=1 (explaining the failure of the SEC to prevent repeated violations by the largest financial companies).

33. Cf. James J. Park, *Rules, Principles, and the Competition to Enforce the Securities Laws*, 100 CALIF. L. REV. 115, 153–54 (2012) (“After a period of increased enforcement, the SEC found itself criticized for its imposition of significant penalties and responded by seeking to limit the situations where such penalties would be sought.”). Of the fifty-one repeat violations of securities laws over the last fifteen years, mostly by large financial institutions, all of the repeated sanctions were imposed on corporations, not individuals. See *infra* note 139 and accompanying text.

34. See *infra* subpart III(b).

The Insider Trading Sanctions Act of 1984 (ITSA) was a response to the rash of insider trading cases that flared up in the late 1970s and early 1980s.³⁵ ITSA authorized treble damages in insider trading cases³⁶ and increased the maximum penalty for violations of the Securities Exchange Act of 1934 to \$100,000.³⁷ When these sanctions proved insufficient to deter, Congress and the Commission doubled down with the Insider Trading and Securities Fraud Enforcement Act of 1988 (ITSFEA), which expanded the Commission's authority to impose penalties on managers who were in control of subordinates that traded on material, nonpublic information in violation of the law and further required the imposition of internal controls by broker-dealers and investment advisors to prevent the "misuse . . . of material, nonpublic information."³⁸ The statute increased the maximum fines for Exchange Act violations further to \$1 million for individuals and \$2.5 million for non-natural persons.³⁹

In 1990, Congress passed the Securities Enforcement Remedies and Penny Stock Reform Act of 1990 (Remedies Act), giving the Commission a multitude of new powers to bring enforcement actions and levy penalties.⁴⁰ In addition to giving the Commission more flexibility in seeking penalties, the Remedies Act granted the Commission the expansive power to collect fines, via an administrative law judge, on regulated entities such as broker-dealers, investment advisors, or others, whenever such a penalty would be in the "public interest."⁴¹ While it took a number of years for the Commission to fully flex its new enforcement muscle, since the enactment of the Remedies Act, the Commission has brought cases against dozens of issuers, collecting billions of dollars in fines.⁴²

However, even this broad expansion of the Commission's punitive scheme did not prevent a deluge of corporate scandals around the turn of the twenty-first century that saw corporate behemoths like Enron and Worldcom collapse in a mess of financial rubble. In response, Congress and the Commission responded as they always had in the past: the Commission was

35. See Paul S. Atkins & Bradley J. Bondi, *Evaluating the Mission: A Critical Review of the History and Evolution of the SEC Enforcement Program*, 13 *FORDHAM J. CORP. & FIN. L.* 367, 386–87 (2008) (stating that the growing number of insider trading cases led the Commission to believe that its existing tools "were inadequate to deter persons from trading on material, nonpublic information").

36. Insider Trading Sanctions Act of 1984, Pub. L. No. 98-376, § 2, 98 Stat. 1264, 1264 (codified as amended at 15 U.S.C. § 78u-1 (2006)).

37. *Id.* § 3. This raised the maximum penalty from \$10,000 to \$100,000. *Id.*

38. Insider Trading and Securities Fraud Enforcement Act of 1988, Pub. L. No. 100-704, § 3, 102 Stat. 4677, 4680 (codified as amended in scattered sections of 15 U.S.C.).

39. *Id.* § 4.

40. Securities Enforcement Remedies and Penny Stock Reform Act of 1990, Pub. L. No. 101-429, 104 Stat. 931 (codified in scattered sections of 15 U.S.C.).

41. *Id.* § 202.

42. Atkins & Bondi, *supra* note 35, at 394. As of 2008, the Commission had brought over sixty cases against issuers. *Id.* That number is surely higher now.

given even broader authority to enforce new and existing laws and to impose even greater penalties.⁴³ The Sarbanes-Oxley Act imbued the Commission with broad new injunctive powers by authorizing it to obtain director and officer bars in administrative proceedings, and to do so under a significantly lower standard of proof.⁴⁴ In addition to these injunctive tools, Sarbanes-Oxley contained a “Fair Funds” provision, which allowed the Commission to seek not just a disgorgement of profits from the wrongdoers, or even penalties commensurate with the broad public harm of an action, but also penalties to compensate individual, harmed shareholders or investors.⁴⁵ The import of this new penalty regime was easy to see. Prior to the enactment of Sarbanes-Oxley in 2002, the largest penalty that the Commission had ever obtained was \$10 million, in its settlement with Xerox.⁴⁶ This was considered exorbitant at the time.⁴⁷ In contrast, from 2003 to 2007, the SEC obtained roughly \$13.8 billion in disgorgement and civil penalties.⁴⁸

Despite this incredible ramp-up in penalties, the financial markets crashed again in 2008, exposing another spate of frauds, from Bernie Madoff’s and Allen Stanford’s Ponzi schemes to the mortgage frauds perpetrated by some of the largest and most respected banks on Wall Street. Clearly, this policy of enforcing the securities laws through huge penalties, extorted via settlement processes, was not having its intended effect.

Nevertheless, in 2010, Congress responded by passing the Dodd-Frank Act, which granted the Commission even broader equitable and legal powers, and further augmented its ability to seek larger and larger fines.⁴⁹ While the breadth of the Dodd-Frank Act is much too large to discuss in such space constraints, there are two relevant ways in which the Act expanded the Commission’s enforcement powers. First, the Dodd-Frank Act provides the SEC with authority to impose substantial *administrative* fines on all

43. Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745 (codified in scattered sections of 15 & 18 U.S.C.); *see also* Atkins & Bondi, *supra* note 35, at 395 (discussing how the Sarbanes-Oxley Act granted the SEC significant control).

44. *See* Sarbanes-Oxley Act of 2002 § 305(a)(1), 15 U.S.C. §§ 77t(e), 78u(d)(2) (2006) (changing the standard from substantially unfit to unfit).

45. *Id.* § 308; *see also* Press Release, U.S. Sec. & Exch. Comm’n, Goldman Sachs to Pay Record \$550 Million to Settle Charges Related to Subprime Mortgage CDO (July 15, 2010), <http://www.sec.gov/news/press/2010/2010-123.htm> (“Of the \$550 million to be paid by Goldman in the settlement, \$250 million would be returned to harmed investors through a Fair Fund distribution . . .”).

46. Barbara Black, *Should the SEC Be a Collection Agency for Defrauded Investors?*, 63 BUS. LAW. 317, 330 (2008).

47. *See id.* (stating that the penalty against Xerox flipped from being considered harsh when it was first imposed to being thought of as “antiquated” after Sarbanes-Oxley).

48. U.S. SEC. & EXCH. COMM’N, 2007 PERFORMANCE AND ACCOUNTABILITY REPORT 26 (2007), *available at* www.sec.gov/about/secpar/secpar2007.pdf.

49. Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub. L. No. 111-203, 124 Stat. 1376 (2010) (codified as amended in scattered sections of 2, 5, 7, 12, 15, 18, 22, 26, 28, 31, 42, and 44 U.S.C.).

persons—not just securities brokers or investment advisors.⁵⁰ While previously the Commission was required to seek an order from a federal district court in a civil action to impose such fines, Dodd-Frank allows the levying of such penalties entirely outside of a district court's jurisdiction.⁵¹ This means that the limited public oversight that Judge Rakoff sought to assert has been withered away for an even larger class of cases.⁵² For these administrative actions, Dodd-Frank “adopts the three-tiered penalty grid already contained in the Securities Exchange Act, but raises the [maximum] penalty amounts by [50%].”⁵³ Second, Dodd-Frank expands the ability for the Commission to bring secondary liability—or aiding and abetting—cases against investment advisors and brokerage firms for failing to adopt appropriate reporting and internal controls.⁵⁴ This lowers the bar for imposing civil penalties on corporations, requiring not that the corporations committed fraud, but only that they “knew” of or were “reckless” in failing to detect the violations.⁵⁵

The Dodd-Frank Act thus falls into the same trap as the regulatory schemes that came before it. It raises the penalties the Commission can seek and lowers the bar that must be cleared to seek the penalties. If history is any guide, this logic that monetary sanctions alone can deter corporate fraud will once again prove tragically flawed, and it will only be a matter of time before the vicious cycle, of frauds begetting fines, repeats itself.

III. Why Do They Settle?

This cyclical pattern of financial fraud and enhanced penalties followed by another fraud raises two questions: First, if the practice of allowing defendants to enter into consent judgments, rather than forcing them into trials, is not deterring wrongdoers—why does the Commission continue to practice it? Second, if the fines have grown so astronomical, why do defendants continue to settle roughly 98% of all cases?⁵⁶

A. *Why the Commission Settles Cases*

The Commission achieves a number of advantages, at seemingly little cost, by settling rather than trying a case. These advantages include

50. *Id.* § 929P (codified in scattered sections of 15 U.S.C.).

51. *The Dodd-Frank Act Reinforces and Expands SEC Enforcement Powers*, GIBSON DUNN (July 21, 2010), <http://www.gibsondunn.com/publications/pages/Dodd-FrankActReinforcesAndExpandsSECEnforcementPowers.aspx>.

52. Certain enforcement remedies may still only be imposed by a federal judge, e.g., issuing an order that prohibits a person from serving as an officer in a public company or an order that requires forfeiture of incentive- or equity-based compensation. *Id.*

53. *Id.*

54. Dodd-Frank Act § 929M (codified at 15 U.S.C. §§ 77o, 80a-47 (Supp. IV 2011)), § 929O (codified at 15 U.S.C. § 78t(e) (Supp. IV 2011)).

55. *Id.* § 929N (codified at 15 U.S.C. § 80b-9 (Supp. IV 2011)).

56. *Supra* note 18 and accompanying text.

extracting cooperation, maximizing Commission resources, and avoiding litigation risks. Further, the conventional wisdom is that the Commission obtains these advantages while at the same time imposing civil penalties and disgorgements of profits that at least equal what the Commission would otherwise hope to receive via a trial verdict. Finally, the revolving door that exists between the Division of Enforcement and the law firms that represent repeat defendants may discourage the vigorous prosecution of such defendants. Each of these potential factors will be evaluated below.

Equal Damages.—As mentioned above, conventional wisdom is that the Commission receives civil penalties that at least equal what they would otherwise hope to achieve at trial.⁵⁷ With so few cases actually going to trial, this information is difficult to verify. The circumstantial evidence, however, is impressive. In 2010, the Commission approved one of its largest fraud settlements to date, imposing fines of \$550 million on Goldman Sachs.⁵⁸ The Commission contends that since the onset of the financial crisis, it has received more than a whopping \$1.2 billion in penalties from the financial crisis.⁵⁹ This is not just a product of volume; the settlement values have increased as well. The *median* settlement value with companies almost doubled from \$800,000 in 2010 to \$1.47 million in 2011.⁶⁰ In other words, the information that does exist gives no reason to question the official line: that the Commission usually settles cases specifically because it believes it can obtain all the relief the Division of Enforcement deems appropriate.

Cooperation.—In recent years in particular, the Commission has instituted a renewed focus on cooperation. In its famous *Seaboard Report*, the Commission first laid out its approach to cooperation with defendants and the circumstances it would consider when determining whether to reduce sanctions against a defendant.⁶¹ Since then, the Commission has only increased its focus on cooperation, adding additional mechanisms, modeled

57. See David M. Becker, *What More Can Be Done to Deter Violations of the Federal Securities Laws?*, 90 TEXAS L. REV. 1849, 1860 (2012) (stating that while working at the Commission, he did not “recall a single instance in which the Division of Enforcement said it was recommending sanctions less severe than what it expected it would get in litigation (except in old or trivial cases)”); Danné L. Johnson, *SEC Settlement: Agency Self-Interest or Public Interest*, 12 FORDHAM J. CORP. & FIN. L. 627, 661 (2007) (“Arguably, the SEC, in a settlement, receives as much in terms of sanctions as it does in a contested proceeding . . .”).

58. Sewell Chan & Louise Story, *Goldman Pays \$550 Million to Settle Fraud Case*, N.Y. TIMES, July 15, 2010, <http://www.nytimes.com/2010/07/16/business/16goldman.html> (“[T]he [Goldman Sachs] settlement would rank among the largest in the 76-year history of the Securities and Exchange Commission . . .”).

59. Eisinger, *supra* note 16.

60. BUCKBERG ET AL., *supra* note 16, at 1.

61. See, e.g., Report of Investigation Pursuant to Section 21(a) of the Securities Exchange Act of 1934 and Commission Statement on the Relationship of Cooperation to Agency Enforcement Decisions, Exchange Act Release No. 44,969, 76 SEC Docket 220 (Oct. 23, 2001) [hereinafter *Seaboard Report*], available at <http://www.sec.gov/litigation/investreport/34-44969.htm> (detailing the criteria that the Commission would consider when determining how much to credit self-policing and cooperation by defendants).

on the investigative procedures of criminal prosecutors, to its enforcement manual.⁶² While the extent of cooperation, and its utility, are difficult to quantify, it is clear that the possibility of cooperation with a defendant in other ongoing investigations is and should be a factor when determining whether to settle.

Conservation of Resources.—The Commission’s caseload is extensive, and has only grown with each successive financial innovation and corresponding crisis. In addition to trying around 14 cases per year,⁶³ the Commission settles another nearly 700 cases.⁶⁴ It is universally accepted that litigation is both time and resource intensive. Trials often take years to see through to the end. Bringing more cases to trial—without a concomitant increase in the Commission’s budget—would necessarily lead to fewer total cases filed. The Division of Enforcement budget was already \$415 million in 2011, roughly one-third of the entire budget of the Commission.⁶⁵ That works out to roughly \$580,000 per case brought to final conclusion assuming that every last penny is spent on litigating cases. Given the difficulties of investigation,⁶⁶ the prolonged nature of trials,⁶⁷ and how cash-strapped the Commission already claims to be,⁶⁸ it is unthinkable that this rate of success could be maintained if the Commission brought even half of these cases to trial.

Litigation Risks.—The Commission impressively boasts that it resolves 92% of its cases successfully.⁶⁹ This statistic deceptively includes settled cases and cases that result in a default judgment, as well as cases at which the

62. See generally DIV. OF ENFORCEMENT, U.S. SEC. & EXCH. COMM’N, ENFORCEMENT MANUAL 119–37 (2012) (setting forth cooperation mechanisms such as proffer agreements, nonprosecution agreements, deferred prosecution agreements, and immunity requests). The Manual is revised from time to time and is available on the Commission’s website for downloading. *Id.* at 1.

63. Eisinger, *supra* note 16.

64. BUCKBERG ET AL., *supra* note 16, at 5.

65. See U.S. SEC. & EXCH. COMM’N, IN BRIEF: FY 2013 CONGRESSIONAL JUSTIFICATION 13, 51 (2012).

66. See, e.g., Peter J. Henning, *Closer Look at S.E.C.’s Mortgage Fraud Charges*, DEALBOOK, N.Y. TIMES (Dec. 19, 2011, 3:16 PM), <http://dealbook.nytimes.com/2011/12/19/closer-look-at-s-e-c-s-mortgage-fraud-charges> (noting that the Commission’s “investigation of Fannie and Freddie took a little more than three years”).

67. See Joshua Gallu, *SEC Trials Increase 50 Percent As Execs Fight Lawsuits*, BLOOMBERG (May 22, 2012), <http://www.bloomberg.com/news/2012-05-22/sec-trials-increase-50-percent-as-execs-fight-lawsuits.html> (“It’s not just an expenditure of resources in the near-term; these cases go on for years and years.” (quoting Mark Schonfeld, former chief of the SEC’s regional office in New York)).

68. See, e.g., Jonathan R. Macey, *The Distorting Incentives Facing the U.S. Securities and Exchange Commission*, 33 HARV. J.L. & PUB. POL’Y 639, 655–56 (2010) (discussing how an SEC official testified that there was no effort to obtain information on Bernie Madoff because the SEC, as a general matter, believes that obtaining and analyzing audit data is “too expensive and time-consuming”).

69. U.S. SEC. & EXCH. COMM’N, FY 2011 PERFORMANCE AND ACCOUNTABILITY REPORT 61 (2011), available at www.sec.gov/about/secpar/secpar2011.pdf.

Commission emerges victorious at trial.⁷⁰ One of the dangers of pursuing more trials, rather than settlements, is the risk that the Commission could lose. Commentators disagree over how successful the Commission could be if it brought more cases to trial⁷¹—but common sense dictates that even if the Commission only proceeds in “winnable cases,” that it will lose at least some of the cases at trial that would have otherwise settled.⁷² My own research shows that of the limited number of trials prosecuted by the SEC in cases filed in the Southern District of New York in 2008, the SEC was victorious in five and lost in one.⁷³ Other surveys of Commission cases have demonstrated that the Commission is most successful when it is pursuing cases within its core areas of enforcement competency, such as fraud and insider trading cases,⁷⁴ but less successful in litigating more peripheral doctrines, based on theories of market timing or aiding and abetting liability.⁷⁵ Not only is losing a case embarrassing publicly, but losing a case on a novel concept of law can have far-reaching collateral estoppel consequences.⁷⁶ The reluctance to bring more cases to trial could also stem from what one commentator has labeled the Commission’s preference to pursue “low-hanging fruit.”⁷⁷ The Commission may focus on settling because such a policy allows it to bring as many cases as possible, collect the largest number of fines possible, and thus secure the largest federal budget possible.⁷⁸ Having notable trial losses splattered across the front pages of the *Wall Street Journal* or *New York Times* surely would inhibit this goal.

70. *Id.*

71. Compare Becker, *supra* note 57, at 1860 (arguing that many defendants are more likely to get a favorable outcome “both on the merits and with respect to sanctions” if they litigate rather than settle claims), with Johnson, *supra* note 57, at 672–73 (claiming that the Commission fares well in trials, at least when it is bringing cases within its core competency—including fraud and insider trading).

72. Becker, *supra* note 57, at 1856 (claiming that in at least some cases defendants are pressured into settling “cases of dubious merit”).

73. See *supra* note 16. Recent trial results are more mixed. While the government has had some successes in prosecuting individuals whose actions were at the heart of the financial crisis, there have also been some notable failures. *E.g.*, Nathaniel Popper & Jessica Silver-Greenberg, *Money-Market Pioneer and Son Cleared of Fraud*, N.Y. TIMES, Nov. 12, 2012, http://www.nytimes.com/2012/11/13/business/bruce-bent-sr-and-son-cleared-of-fraud-charges.html?pagewanted=1&_r=0 (discussing a federal jury’s recent repudiation of a fraud charge brought against the inventor of the “money market fund” whose flagship fund failed in September 2008).

74. See generally Peter M. Saporoff et al., *Hitting Home Runs or Swinging and Missing? Examining How the SEC Fared in Recent District Court Litigation*, in ALI-ABA COURSE OF STUDY, SECURITIES LITIGATION: PLANNING AND STRATEGIES 103 (2007), available at SM015 ALI-ABA 711 (Westlaw) (finding SEC success in fraud and insider trading cases); see also Johnson, *supra* note 57, at 672–73 (same).

75. See Saporoff et al., *supra* note 74 (calling the Commission’s success in secondary liability cases a mixed bag).

76. Johnson, *supra* note 57, at 666–68.

77. Macey, *supra* note 68, at 654–57.

78. *Id.* at 646.

Regulatory Capture.—Finally and most sinisterly, it is possible that at least part of the Commission’s unwillingness to bring more cases to trial stems from a fear of biting the hand that feeds it. In recent years, the Division of Enforcement has been a head-hunting ground for the largest New York investment banks and law firms—those entities that most routinely serve as the defendants or defense counsel in the Commission’s biggest cases. The current Director of the Division, Robert Khuzami, was counsel at Deutsche Bank before taking the job at the SEC.⁷⁹ He is hardly the only one to switch sides at the negotiating table. From 2008 through the first nine months of 2009, sixty-six former Commission employees filed 168 letters with the Commission disclosing that they planned to represent a client before the Commission.⁸⁰ Such letters only need to be filed if the employee has left the Commission within the last two years.⁸¹ Besides this systemic use of the Commission as an externship, which gives off only the perception of impropriety, there has also been anecdotal evidence of particular influence wielded by former employees. In one example, an insider-trading investigation was halted within days of a phone call to the Director of Enforcement by a former U.S. Attorney, then in private practice.⁸² As Michael Lewis and David Einhorn put it, one “could be forgiven for thinking that the whole point of landing the job as the S.E.C.’s director of enforcement is to position oneself for the better paying one [as a lawyer] on Wall Street.”⁸³

Whether or not this fundamentally affects the Commission’s strategic decisions on an institutional basis is difficult to determine, but it seems unlikely that a counsel for the Commission, looking for a lucrative turn in private practice, would do too much to upset those on the other side of the negotiating table. This analysis assumes that the other side of the negotiating table—the defendant—sees settlement as the most desirable outcome once it has been charged with wrongdoing. Despite the fact that the civil penalties are expected to be just as large as they would be at trial, and the fact that by settling the defendant is forfeiting any discount to the expected value of the litigation by the chance that it might win at trial, the sheer propensity of settlements shows that they are just as favored by defendants as they are by the Commission.

79. Press Release, U.S. Sec. & Exch. Comm’n, Robert Khuzami Named SEC Director of Enforcement (Feb. 19, 2009), <http://www.sec.gov/news/press/2009/2009-31.htm>.

80. Tom McGinty, *SEC Lawyer One Day, Opponent the Next*, WALL ST. J., Apr. 5, 2010, <http://online.wsj.com/article/SB10001424052702303450704575160043010579272.html>.

81. *Id.*

82. Peter J. Henning, *The Revolving Door and S.E.C. Enforcement*, DEALBOOK, N.Y. TIMES (Apr. 8, 2010, 3:07 PM), <http://dealbook.nytimes.com/2010/04/08/the-revolving-door-and-s-e-c-enforcement/>.

83. Michael Lewis & David Einhorn, Op-Ed., *The End of the Financial World as We Know It*, N.Y. TIMES, Jan. 4, 2009, <http://www.nytimes.com/2009/01/04/opinion/04lewiseinhorn.html>.

B. *Why Do Defendants Agree?*

Consent judgments quite obviously require consent by the defendants; so for such a high proportion of cases to settle, these defendants must be reaping substantial benefits from settling, rather than litigating their cases. This is doubly true if we assume—as is consistently argued—that defendants in settlements pay roughly the equivalent in damages that they would likely pay if found liable at trial.⁸⁴ The benefits of consent judgments include: (a) the limitation of collateral risks; (b) the limitation of litigation risks and the concomitant hassles of litigation; and (c) perhaps most importantly, the mitigation of any real reputational harm for either the corporations or individuals involved.

No Collateral Risks.—As covered above, a consent judgment does not require the defendant to admit any wrongdoing.⁸⁵ While this feature may only be a moral victory for an individual defendant who wishes to maintain his innocence, this can be crucial for defendants seeking to limit their exposure to collateral estoppel.⁸⁶ Collateral estoppel, also known as issue preclusion, bars the ability of a losing defendant to litigate issues that were already determined on the merits by a case litigated by that defendant.⁸⁷ Historically, issue preclusion could only be used between the same plaintiff and defendant in a subsequent suit.⁸⁸ However, this mutuality doctrine has been abandoned in most jurisdictions.⁸⁹ Nonmutual collateral estoppel generally allows a plaintiff who was not party to a prior suit to rely on that suit's determination of (a) an essential issue of fact or law (b) litigated and determined by (c) a valid and final judgment against that particular defendant.⁹⁰ This is a rule of judicial economy, seeking to prevent plaintiffs who have lost from soldiering on in malice and defendants who have lost from ignoring the decided issues in earlier cases. While the Supreme Court has stated that plaintiffs may not use issue preclusion if they could have easily joined the earlier action,⁹¹ the Commission's policy of opposing joinder makes this an unlikely obstacle.⁹² Further, district courts have broad discretion in determining whether or not to allow even offensive, nonmutual

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

85. See *supra* notes 20–22 and accompanying text.

86. Johnson, *supra* note 57, at 668.

87. See *Parklane Hosiery Co. v. Shore*, 439 U.S. 322, 329 (1979) (distinguishing between collateral estoppel as applied to losing defendants and to losing plaintiffs); RESTATEMENT (SECOND) OF JUDGMENTS § 29 (1982).

88. Johnson, *supra* note 57, at 666.

89. *Id.* at 666–67.

90. RESTATEMENT (SECOND) OF JUDGMENTS §§ 27, 29 (1982).

91. *Parklane Hosiery Co.*, 439 U.S. at 331.

92. See, e.g., SEC v. Bear, Stearns & Co., No. 03 Civ. 2937, 2003 WL 2200340, at *3 (S.D.N.Y. Aug. 25, 2003) (rejecting intervention on the grounds that “[r]eflexive intervention by the public in SEC actions would undermine both the SEC’s ability to resolve cases by consent decree and the efficient management of those cases by courts”).

issue preclusion.⁹³ The three significant factors courts consider are “(1) whether the defendant had a full and fair opportunity to litigate the issue in the first proceeding, (2) whether the court was fair in its determination in the first proceeding, and (3) whether there have been changes in the law since the first proceeding.”⁹⁴ Therefore, should a defendant attempt to litigate a Commission action and lose, it could be precluded from relitigating that issue against private plaintiffs who are suing over the same alleged wrong. This could expose potential defendants to substantial—and almost automatic—liability. Not only does this partially explain corporations’ reluctance to go to trial, but it may also explain the inclusion of the “neither admit nor deny” settlement language. While settlements do not give rise to issue preclusion,⁹⁵ corporations fear that if they were forced to admit guilt as part of a settlement with the Commission, it could later be used against them by private plaintiffs.⁹⁶

No Litigation Risks or Hassles.—Litigation will always pose risks that settlement does not. Even if we assume that the average Commission settlement is equivalent to what an average jury would return for the alleged violations of securities laws, there is always the chance that the penalty could be more significant. Further, going to trial is typically expensive and time-consuming.⁹⁷ The revelation of electronic discovery has only compounded these costs and hassles.⁹⁸ Beyond just these direct costs, though, litigation imposes significant burdens on corporate defendants. Their computers get imaged, their e-mails reviewed, and their employees can be subjected to hours upon hours of depositions.⁹⁹ Settling—and settling quickly—allows a company to exchange these hassles and risks for the certainty of a set fee. The settlement process cabins in both the potential for extreme liability—

93. *Parklane Hosiery Co.*, 439 U.S. at 331.

94. Johnson, *supra* note 57, at 668.

95. See Michael W. Loudenslager, Note, *Erasing the Law: The Implications of Settlements Conditioned Upon Vacatur or Reversal of Judgments*, 50 WASH. & LEE L. REV. 1229, 1246 (1993) (arguing that avoiding issue preclusion is one of the incentives for defendants to settle).

96. See U.S. SEC v. Citigroup Global Mkts. Inc., 827 F. Supp. 2d 328, 334 (S.D.N.Y. 2011) (explaining how the neither-admit-nor-deny rule protected Citigroup against future civil liability). It is not clear how real this fear is, and it is possible that settlement language could be structured so as to force a defendant to admit some conduct without it having collateral effects. For instance, in *SEC v. Goldman, Sachs & Co.*, Goldman Sachs admitted that its materials had “contained incomplete information” and that “it was a mistake” not to have disclosed certain information. SEC v. Goldman, Sachs & Co., Litigation Release No. 21,592, 98 SEC Docket 3135, 3135 (July 15, 2010).

97. Cf. Mukesh Bajaj et al., *Empirical Analysis: Securities Class Action Settlements*, 43 SANTA CLARA L. REV. 1001, 1010 tbl.4 (2003) (providing statistical data showing that only a small fraction of securities class actions settle within one year).

98. See Milberg LLP & Hausfeld LLP, *E-Discovery Today: The Fault Lies Not in Our Rules . . .*, 4 FED. CTS. L. REV. 131, 133–34 (2011) (explaining the high cost and lengthy duration of discovery in litigation); see also FED. R. CIV. P. 26(b)(5)(B) advisory committee’s notes (acknowledging that electronic discovery exacerbates the costs and effort of discovery).

99. See Becker, *supra* note 57, at 1861 (detailing the intrusive nature of Commission investigations).

through a runaway jury or issue preclusion—and the potential for extreme inconvenience through litigation. It reduces the collateral effects of litigation into a set fine: a business cost that simply needs to be reflected on the books.

No Reputational Harm.—This idea of accepting a fine because it is a certain, concrete business cost, while litigation represents a series of amorphous risks, underlies the third primary rationale for settling: it severely mitigates any long-term reputational harm for a defendant. First, settling keeps much of the dirty laundry under wraps, and allows companies to avoid potentially devastating negative publicity.¹⁰⁰ Since most consent judgments are announced the same day as a complaint is filed,¹⁰¹ settling with the Commission at least mitigates the negative publicity and may even be a financial windfall. Goldman Sachs, after announcing its consent to one of the largest financial penalties in Commission history, had its stock price rise by more than 10%.¹⁰²

Not only does the corporation get to avoid reputational harm and any collateral risks that such harm may impose, but the individuals involved—charged or uncharged—are able to walk away virtually untarnished. After the Global Financial Crisis of 2008, the Commission filed a number of actions alleging negligence or fraud against most of the major investment and commercial banks operating in the United States. These suits included actions against Citigroup, Goldman Sachs, Wachovia, AIG, JPMorgan Chase, UBS, and Merrill Lynch, among others.¹⁰³ Most of these cases have reached a negotiated settlement—usually requiring the charged defendants to pay hundreds of millions of dollars in fines¹⁰⁴—and yet the major executives at each of those banks have remained employed.¹⁰⁵ Not only are the

100. See, e.g., Andria Cheng, *Avon Slumps on Result, SEC Investigations*, MARKETWATCH (Oct. 27, 2011), http://articles.marketwatch.com/2011-10-27/industries/30744456_1_shares-of-avon-products-beauty-sales-sec (explaining how the stock prices of Avon Products Inc. plummeted and management “lost investor credibility” after announcement of an SEC investigation against the company).

101. Roger Parloff, *The Judge Who Slapped Citi*, CNNMONEY (Nov. 30, 2011, 11:43 AM), <http://finance.fortune.cnn.com/2011/11/30/judge-jed-rakoff-citigroup-sec/>.

102. John Curran, *Goldman Sachs Settles with SEC, Stock Soars*, TIME (July 15, 2010), <http://business.time.com/2010/07/15/goldman-sachs-settles-with-sec-stock-soars/>.

103. See *Wall Street's Repeat Violations, Despite Repeated Promises*, N.Y. TIMES, Nov. 7, 2011, <http://www.nytimes.com/interactive/2011/11/08/business/Wall-Streets-Repeat-Violations-Despite-PromisesStsssss.html?ref=business> (showing that suits have been brought and settled against each of these companies, among others, since the advent of the GFC).

104. See, e.g., BUCKBERG ET AL., *supra* note 16, at 2 (listing a number of the settlements with these companies).

105. *Bank Execs Still Clock In Despite Failures*, CBS NEWS (Jan. 20, 2010, 12:39 PM), <http://www.cbsnews.com/stories/2009/01/27/business/main4755956.shtml>. While a few of these executives have recently left their former positions at these institutions, these departures represent the exceptions and not the rule. Further, those executives that have left have usually been replaced by individuals who occupied other, senior positions during the financial meltdown. See, e.g., Donal Griffin & Bradley Keoun, *Citigroup Board Said to Oust Pandit After Setbacks*, BLOOMBERG (Oct. 19, 2012, 11:22 AM), <http://www.bloomberg.com/news/2012-10-16/pandit-steps-down-as-citigroup-s-chief-as-corbat-takes-over-1-.html> (discussing former CEO Vikram Pandit's ouster as

executives unpunished by the imposition of sanctions, but as one commentator put it, Commission consent judgments “can be a boon to in-house lawyers,” the very people who advise their companies to settle or defend a litigation action.¹⁰⁶ By not admitting or denying allegations against them, in-house lawyers are able to avoid any collateral actions for disbarment or sanctions by state bar associations.¹⁰⁷ Over the past few years a number of notable general counsels have agreed to Commission settlements without admitting or denying the allegations. The General Counsel of Apple Inc., Nancy Heinen, paid \$2.2 million for backdating stock options in 2008.¹⁰⁸ Google Inc.’s David Drummond, in a classic case of fool me once—fool me twice, was twice able to settle allegations against him by simply agreeing to “cease and desist” from future violations of securities laws, both in 2005 and in 2007.¹⁰⁹ Enron Corp.’s Jordan Mintz accepted a two-year ban from appearing before the Commission for disclosure issues in 2009.¹¹⁰ Despite these actions, Drummond remains at Google.¹¹¹ Heinen is a partner and advisor at SV2 (Silicon Valley Social Venture) Fund and sits on the advisory board of the University of California’s Berkeley Center for Law, Business and the Economy.¹¹² Mintz is the Vice President and Chief Tax Officer of Kinder Morgan.¹¹³ In other words, by settling, these general counsels have been able to avoid any real reputational or professional harm to their careers.

By settling, without admitting to the allegations of wrongdoing, both the individuals and the corporations are able to avoid the all-important collateral consequences and risks of litigation, and instead transform the entire enforcement scheme from a true penalty into a defined and strictly cabined business cost. The benefits of settling, both on an institutional and an individual level, explain not only the high volume of settlements, but also the very ineffectiveness of settlements to achieve the principal deterrent goals of a regulatory system.

IV. These Settlements Are Not Deterring Wrongdoing

The principal goal of Commission sanctions is to deter future wrongdoing, thereby allowing for the proper functioning of the markets.¹¹⁴

Citigroup CEO and his replacement with former Citigroup division CEO for Europe, the Middle East, and Africa, Michael Corbat).

106. Sue Reisinger, *No Apologies: The Agency’s No-Guilt Deals Can Be a Boon to In-House Lawyers*, CORP. COUNS., Feb. 2012, at 72.

107. *Id.*

108. *Id.*

109. *Id.*

110. *Id.*

111. *Id.* at 74.

112. *Id.*

113. *Id.*

114. See Becker, *supra* note 57, at 1852 (explaining that the severity of sanctions has primarily driven the SEC’s deterrence strategy for the past twenty years).

Commission sanctions—as civil penalties—are not traditionally retributive, and while the Commission does have the power to recoup penalties on behalf of private parties, the purpose of the Commission is not primarily to remedy past wrongs.¹¹⁵ Therefore, the success or failure of the Commission's regulatory policy must be judged on how well it deters. By that standard, the Commission is failing miserably. Large settlements may stop the presses¹¹⁶—but they do not stop legal trespasses.

A. *Theoretical Failures*

There are two widely accepted yet contradictory truths about the SEC settlement policy. One is that almost every case settles.¹¹⁷ The second is that these settlements result in civil penalties that are roughly equivalent in scale to what a defendant would be exposed to if the defendant went to trial and lost.¹¹⁸ These truths add up to an unassailable conclusion: the current settlement policy is failing to adequately deter wrongdoing, or at least failing to deter as effectively as the threat of trial would. For the deterrent effect of a settlement-only regime to approach that of a regime in which trials are a cognizable enforcement mechanism, corporate defendants should choose settlement in only about half of the cases. If that were the case, we would know that the settlement was as powerful a disincentive to commit wrongdoing as civil prosecution. The fact that defendants choose settlement in such an outrageously high proportion of cases can mean only one thing: settlement is not nearly the punishment that a trial would be.

Criminology theory states that the effectiveness of a deterrent scheme turns on (a) the severity of the sanctions; (b) the perceived likelihood that sanctions will be imposed; (c) the amount of time that would theoretically lapse between the wrongful conduct and the imposition of sanctions; and (d) the extent to which there are extralegal consequences from the unlawful conduct.¹¹⁹ As discussed above, in a settlement action, the severity of sanctions—as defined by monetary damages—is essentially equivalent to the severity of what would be imposed at trial.¹²⁰ The likelihood of the imposition of such sanctions is higher in a settlement scheme than in one where more cases are brought to trial. When bringing a case to trial, there is always the possibility that a defendant could win. The timeliness of the

115. See *id.* at 1852–53 (stating that the main focus of early SEC sanctions was not penal).

116. See Johnson, *supra* note 57, at 665, 672 (discussing avoiding reputational harm as a reason that the Commission may settle cases).

117. See *supra* note 16 and accompanying text.

118. Becker, *supra* note 57, at 1860; Johnson, *supra* note 57, at 661.

119. See Raymond Paternoster, *How Much Do We Really Know About Criminal Deterrence?*, 100 J. CRIM. L. & CRIMINOLOGY 765, 781, 783, 815 (2010) (laying out the first three factors and discussing extralegal factors such as social censure that arise from being arrested, which contributes to deterrence).

120. See Becker, *supra* note 57, at 1860 (arguing that the SEC obtains roughly the same sanctions through settlement as it would through litigation); Johnson, *supra* note 57, at 661 (same).

imposition of sanctions is also improved under a settlement scheme versus one in which cases are brought to trial (albeit minimally).¹²¹ All of these factors seem to point to a settlement scheme as one that achieves superior deterrence. The problem is that the facts do not bear this theory out—defendants are not electing to go to trial—and this is particularly evident as applied to the largest banks and financial companies in the United States.

Monetary sanctions for corporate defendants rarely punish the wrongdoer. In the Citigroup case discussed above, the wrongdoers there were the executives and Citigroup officials that countenanced a malignant mortgage securities deal.¹²² In the settlement with Goldman Sachs, the wrongdoers were traders who developed and packaged the fraudulent marketing materials and the executives who allowed it to happen.¹²³ Yet by settling, the real violators of the securities laws passed the buck and the bill from themselves to their shareholders (and creditors).¹²⁴ The assets of the company took a hit while the job security and assets of the individuals responsible remained unchanged.¹²⁵ This strikes at the fundamental truth of settling with large investment banks and corporations—the sanction is really no sanction at all and has only the most limited deterrent effect—namely, that those responsible for the violations of the securities laws reap all of the benefits of those violations, in the form of promotions, job security, and bonuses during the good times, without bearing the costs of the harm.¹²⁶ This explains why defendants are so eager to settle with the Commission. They avoid the costs that litigation may impose on them (from having to undergo the hassles of the trial process to suffering real reputational harm) in

121. See Becker, *supra* note 57, at 1861 (stating that litigation takes years, contributing to uncertainty and instability in the market, which can be avoided by settling).

122. U.S. SEC v. Citigroup Global Mkts. Inc., 827 F. Supp. 2d 328, 329–30 (S.D.N.Y. 2011).

123. SEC v. Goldman, Sachs & Co., Litigation Release No. 21,592, 98 SEC Docket 3135, 3135–36 (July 15, 2010).

124. See Peter J. Henning, *Behind Rakoff's Rejection of Citigroup Settlement*, DEALBOOK, N.Y. TIMES (Nov. 28, 2011, 5:14 PM), <http://dealbook.nytimes.com/2011/11/28/behind-judge-rakoffs-rejection-of-s-e-c-citigroup-settlement/> (agreeing with Judge Rakoff that the civil penalties “effectively penalize[] the very shareholders who were misled” by the corporate wrongdoing). This is not the first time that a global financial crisis has followed financial panic caused by financial institutions playing fast and loose with other people’s money. See generally LOUIS D. BRANDEIS, OTHER PEOPLE’S MONEY AND HOW THE BANKERS USE IT (1914).

125. The executives in charge of Citigroup and JPMorgan Chase pre-financial crisis are largely still there. See, e.g., *Members of the Board*, JPMORGAN CHASE & CO., <http://www.jpmorganchase.com/corporate/About-JPMC/board-of-directors.htm> (showing that all but two members of the current executive officers were on the board prior to the financial crisis); *Operating Committee*, CITIGROUP, http://www.citigroup.com/citi/about/our_leaders.html (showing that more than half of the members of Citigroup’s current Operating Committee were also in high-ranking positions prior to 2008). Although Vikram Pandit, the CEO of Citigroup, was recently let go, he was replaced with Michael Corbat, another longtime Citigroup executive. Griffin & Keoun, *supra* note 105.

126. A few months after settling with the Commission, Goldman Sachs paid out \$15.3 billion in salary and bonuses. Jill Treanor, *Goldman Sachs Bankers to Receive \$15.3bn in Pay and Bonuses*, GUARDIAN, Jan. 19, 2011, <http://www.guardian.co.uk/business/2011/jan/19/goldman-sachs-bankers-pay-bonuses>.

exchange for promising the Commission other people's (their shareholders') money.¹²⁷

These officers are not even harmed tangentially, through a decrease in the value of their stock options or an increase in pressure on their jobs by the Board of Directors, as the announcement of a settlement with the Commission is often greeted with a rise, rather than a decline, in the stock price of a corporation.¹²⁸ There is a common saying that it is unlikely that a criminal would violate a law if the punishment were five years in prison, but would not if the punishment were ten.¹²⁹ This has been used as a critique, by apologists for the Commission, of the notion that Commission penalties need to be increased for greater deterrence.¹³⁰ While this statement is undoubtedly true, an even truer statement is that if a slap on the wrist is not going to deter a potential violator of the securities laws, two slaps on the wrist will not be any more effective.¹³¹

If the first half of a consent judgment is an ineffective monetary penalty, the second half is a joke. In addition to requiring the payment of civil penalties and a disgorgement of profits, the Commission typically exacts an injunction as part of any settlement agreement.¹³² These injunctions theoretically allow the Commission to intervene and penalize a defendant, without having to resort to lengthy settlement negotiations or a trial, if the

127. While it is true that shareholders could punish executives and officers for their malfeasance, in our era of staggered boards and widely disseminated voting power, this is unlikely. See Harwell Wells, *The Birth of Corporate Governance*, 33 SEATTLE U. L. REV. 1247, 1252–53 (2010) (“[I]n a world of dispersed and ‘largely inactive’ shareholders . . . the separation of ownership and control is ‘the central problem of corporate governance.’”).

128. See, e.g., Curran, *supra* note 102 (explaining a 10% increase in the price of Goldman Sachs's stock after news leaked of its settlement with the Commission); cf. Andrew T. Berry, *Comments on Aggregation: Some Unintended Consequences of Aggregative Disposition Procedures*, 31 SETON HALL L. REV. 920, 921 (2001) (“Certainty has value in the capital markets’ analysis of a company A company which plausibly consigns mass tort liabilities to the past by . . . announcing a ‘global settlement’ (no matter how expensive) may be rewarded by an increase in its market capitalization.”).

129. See Becker, *supra* note 57, at 1867 (arguing that while “some people who would commit fraud at the risk of five years in prison would not do so if the risk were incarceration for ten years” the difference in punishment would only operate at the margins).

130. See *id.* (arguing that increasing the amount of penalties does not increase the effectiveness of deterrence after a certain threshold).

131. See Francesco Guerrera, *Slapped Wrist and Back to Business for Goldman*, FIN. TIMES (July 23, 2010, 5:56 PM), <http://www.ft.com/intl/cms/s/0/91354f24-9676-11df-9caa-00144feab49a.html#axzz1owFih618> (claiming that the \$550 million penalty, the largest fine levied by the SEC, was a slap on the wrist relative to the bank's regular earnings); Matt Taibbi, *Federal Judge Pimp-Slaps the SEC Over Citigroup Settlement*, ROLLING STONE (Nov. 29, 2011, 10:10 AM), <http://www.rollingstone.com/politics/blogs/taibblog/federal-judge-pimp-slaps-the-sec-over-citigroup-settlement-20111129> (calling settlement penalties “payoffs to keep the SEC off the banks’ backs” and comparing them to “the pad that numbers-runners or drug dealers pay to urban precinct-houses every month to keep cops from making real arrests”).

132. See 15 U.S.C. § 78u(d)(1) (2006) (providing for the imposition of “a permanent or temporary injunction or restraining order” as punishment for a violation); SEC v. Quadrangle Grp. LLC, Litigation Release No. 21,487, 98 SEC 1088, 1089 (Apr. 15, 2010) (imposing a civil fine and an injunction for a violation of the Securities Act of 1933).

defendant continues to violate securities laws after the consent judgment is ordered. While such a power has been called “serious,”¹³³ potentially “disastrous,”¹³⁴ and a “drastic remedy,”¹³⁵ in reality it has little bite. As Judge Rakoff pointed out in his *Citigroup* opinion, the SEC has not sought to enforce one of its injunctions in over ten years.¹³⁶ This despite the fact that corporate defendants routinely violate these injunctions in an almost pedestrian fashion. First, after the consent judgment is entered—in flagrant disregard of the “neither admit nor deny” language—violators will continue to clandestinely (or even openly) refute those allegations they just settled to the rest of the financial community.¹³⁷ Second, and even more galling, the defendants will not just continue to deny the allegations that they have just settled, but will perpetuate the activity for which they were originally punished.¹³⁸ Imposing injunctions on Commission defendants is the equivalent of a parent making a rebellious child promise to stop stealing cookies from the cookie jar. Sure, the child may promise, but at best, the only effect that this is going to have is to make her more circumspect in her thievery the next time.

B. Empirical Failures

In sum, the harm of sanctions is borne, if at all, by shareholders and creditors. Increasing monetary sanctions, as each successive Congressional enactment has sought to do, does not increase the deterrent power of sanctions because those making the decisions never internalize the threat of sanctions. Further, injunctive relief, a bulwark of the overall settlement scheme, has become defunct through disuse. This analysis is borne out by the empirical data. Since 1996, there have been fifty-one repeated violations of securities laws by the largest Wall Street firms.¹³⁹ Including the initial violations during this period, the largest Wall Street banks violated the securities laws a whopping seventy-seven times over the course of this

133. Thomas J. André, Jr., *The Collateral Consequences of SEC Injunctive Relief: Mild Prophylactic or Perpetual Hazard?*, 1981 U. ILL. L. REV. 625, 670.

134. *Id.*

135. *Aaron v. SEC*, 446 U.S. 680, 703 (1980) (Burger, C.J., concurring).

136. *U.S. SEC v. Citigroup Global Mkts. Inc.*, 827 F. Supp. 2d 328, 333 (S.D.N.Y. 2011).

137. *E.g.*, Bruce Carton, *Settling SEC Defendants Never ‘Admit’ Wrongdoing But They Sometimes Later ‘Deny’ It*, COMPLIANCE WEEK (Dec. 7, 2011), <http://www.complianceweek.com/settling-sec-defendants-never-admit-wrongdoing-but-they-sometimes-later-deny-it/article/218356/> (giving examples of after-the-fact denials that went unpunished by the Commission). The Commission is at least aware of this problem, even if it has not moved to quash it. *See* Luis A. Aguilar, Comm’r, Sec. & Exch. Comm’n, *Setting Forth Aspirations for 2011: Address to Practising Law Institute’s SEC Speaks in 2011 Program* (Feb. 4, 2011), available at <http://www.sec.gov/news/speech/2011/spch020411laa.htm> (proclaiming that he hopes to bring an end to the practice of defendants issuing press releases, after a settlement is announced, downplaying their wrongdoing).

138. *See Wall Street’s Repeat Violations, Despite Repeated Promises*, *supra* note 103 (listing all of the firms with repeat violations over the last fifteen years); *see also* Wyatt, *supra* note 32 (explaining the data).

139. *Wall Street’s Repeat Violations, Despite Repeated Promises*, *supra* note 103.

fifteen-year period.¹⁴⁰ The list of repeat offenders is remarkable in how well it correlates with the banks that were the most financially unstable and exposed during the GFC. Bank of America violated and settled eight breaches of the securities laws during this period, including four violations of § 17(a) of the Securities Act of 1933 for purposeful or negligent fraud in interstate commerce, and four violations of § 15(c) of the Securities Exchange Act of 1934 for purposeful fraud by a securities firm.¹⁴¹ Citigroup also had eight violations.¹⁴² Bear Sterns had six;¹⁴³ Goldman Sachs had three;¹⁴⁴ Merrill Lynch had seven.¹⁴⁵ Despite this reckless disregard for the securities laws, as mentioned above, the Commission refrained entirely from using its contempt power under prior settlements to enforce the securities laws.¹⁴⁶

Amusingly, these firms cannot seem to stop violating the securities laws even though the Commission has routinely granted exemptions to those very firms from those securities laws. As evidence of further favorable treatment, the Commission has granted 350 waivers to Wall Street institutions and financial companies, “allow[ing] the biggest firms to avoid punishments specifically meant to apply to fraud cases.”¹⁴⁷ The Commission’s settlement policy then—especially as applied to the largest Wall Street Banks—is failing to adequately punish wrongdoers, and unsurprisingly, this is failing to deter them from committing wrongs again in the future. The current variables are not adding up: a change in the deterrent calculus is needed.

V. Toward a New Enforcement Paradigm

Virtually any human activity can be understood as a product of benefits and costs, and individual and corporate actors are assumed to be rational enough to weigh the benefits of any action against the costs of such action as well as those of any reasonable alternative.¹⁴⁸ Violations of securities laws can be explained in this very manner. The utility of a violation is equal to the sum of the benefits of the violation (earn more money, avoid a loss), the costs of the violation (getting sanctioned by the Commission), the benefits of not violating the law (no risk of sanction, stability, peace of mind), and the costs of not violating the law (less money, etc.).¹⁴⁹ Deterrence theory

140. *Id.*

141. *Id.*

142. *Id.*

143. *Id.*

144. *Id.*

145. *Id.*

146. Wyatt, *supra* note 32.

147. Edward Wyatt, *S.E.C. Is Avoiding Tough Sanctions for Large Banks*, N.Y. TIMES, Feb. 3, 2012, [www.nytimes.com/2012/02/03/business/sec-is-avoiding-tough-sanctions-for-large-banks.html?_r=2&sq=SEC exemption securities fraud&st=cse&scp=1&pagewanted=all](http://www.nytimes.com/2012/02/03/business/sec-is-avoiding-tough-sanctions-for-large-banks.html?_r=2&sq=SEC%20exemption%20securities%20fraud&st=cse&scp=1&pagewanted=all).

148. Paternoster, *supra* note 119, at 782.

149. *Id.* at 783.

presumes—quite rationally—that an increase in the cost of a violation will decrease the likelihood of that violation.¹⁵⁰ The costs of a legal punishment are traditionally assumed to be (a) the certainty; (b) the severity; (c) the celerity or swiftness of the punishment; and (d) the extent of extralegal consequences.¹⁵¹ The greater any of these variables is, the lower the rate of violations should be.¹⁵² Further, there are two levels for each of these punishment properties. There is an objective level—the actual likelihood and amount of punishment a wrongdoer will receive—and more importantly, a subjective level—the likelihood that a potential wrongdoer perceives he will be punished and what he perceives will be the extent of his punishment.¹⁵³

As the empirical evidence above demonstrates, the current Commission enforcement regime insufficiently deters wrongdoing.¹⁵⁴ Therefore, the costs of violating the securities laws are not sufficiently high in comparison to the benefits of violating those laws (again, the benefits can be multifarious: increased compensation, esteem of peers from performing well, and avoidance of loss). This is despite the fact that (a) settlements are relatively certain punishments, as 98% of cases settle,¹⁵⁵ and, in each of those cases, the defendant is paying a not-insubstantial monetary penalty;¹⁵⁶ (b) settlements are cost-effective as an enforcement mechanism, allowing the Commission to presumably bring more cases than it otherwise would;¹⁵⁷ and (c) settlements by their very nature are imposed more swiftly than any punishment from litigation.¹⁵⁸

If this current regime has been unsuccessful, what are the potential solutions? One solution is the same one that the Commission and Congress have relied on in the past: greater monetary sanctions.¹⁵⁹ However, the problem with the current enforcement regime is not the amount of monetary sanctions imposed, but rather the fact that monetary sanctions against corporate defendants do not punish the individual decision makers but the shareholders. The bankers are playing with “house money.” Unless they

150. JACK P. GIBBS, CRIME, PUNISHMENT, AND DETERRENCE 5–6 (1975); see FRANKLIN E. ZIMRING & GORDON J. HAWKINS, DETERRENCE: THE LEGAL THREAT IN CRIME CONTROL 3 (1973) (“[T]here is the potent, ubiquitous, seemingly irrefutable thesis that attaching unpleasant consequences to behavior will reduce the tendency of people to engage in that behavior.”).

151. Paternoster, *supra* note 119, at 783, 815.

152. *Id.* at 784.

153. See GIBBS, *supra* note 150, at 5 (distinguishing between objective and subjective perceptions of potential costs, indicating that “prescribed or ‘threatened’ punishments . . . do not deter individuals unless they perceive some risk”).

154. See *supra* Part IV.

155. See *supra* note 16 and accompanying text.

156. See BUCKBERG ET AL., *supra* note 16, at 1 (indicating that the median settlement in 2011 was \$1.47 million).

157. See Johnson, *supra* note 57, at 663 (“Generally, the expense, risk, and delay that frequently attend formal adjudication explain, at least in part, a party’s preference for, and the rising incidence of, settlement.”).

158. *Id.*

159. See *supra* Part II.

reach absolutely crippling levels—such that directors and officers of large financial institutions are routinely fired for their failure to prevent frauds or such that the companies actually fail as a result of the penalties—higher monetary sanctions are never going to adequately deter individual wrongdoers. In other words, the only way to efficiently deter wrongdoing is for wrongdoers to internalize the costs of their wrongdoing. The solution is not to increase the monetary sanctions imposed on the corporations, but rather to initiate a sanctions regime that imposes pain—either directly or collaterally—on the decision makers and managers, so that they will internalize the costs of violations and be deterred from authorizing or engaging in them.

More Trials.—One method to achieve this goal is to bring more cases to trial, by either refusing to settle or by putting a halt to the practice of settling cases without requiring an admission of wrongdoing. The Commission may already be moving in this direction. In its settlement, Goldman Sachs acknowledged that its marketing materials for the disputed collateralized debt obligation contained “incomplete information” and that it was a “mistake.”¹⁶⁰ Further, the Commission recently announced a policy of requiring admissions of guilt when the defendant has already been convicted in a parallel criminal proceeding.¹⁶¹ While these steps are still long strides away from actually requiring admissions of guilt in most settlements, they are signs of progress.¹⁶² Demanding actual admissions of guilt in all settlements would recalibrate the costs and benefits of settlement, as such admissions could open up the defendants to *res judicata* liability.¹⁶³ This would further the real goal of such a policy change, which is not extracting nominal admissions of guilt out of defendants but bringing more of them to trial. On just a basic theoretical level, it is clear that trials would be a more effective deterrent than settlements in their current form. If trials and settlements were equally distasteful to corporate defendants, theory would dictate that only

160. SEC v. Goldman, Sachs & Co., Litigation Release No. 21,592, 98 SEC Docket 3135, 3135 (July 15, 2010).

161. See Steve Schaefer, *SEC Rule Change Doesn't Mean Much For Wall Street Settlements*, FORBES (Jan. 06, 2012, 3:38 PM), <http://www.forbes.com/sites/steveschaefer/2012/01/06/sec-rule-change-wont-have-wall-street-admitting-guilt/> (explaining the rule change and pointing out just how limited its effect will be).

162. Even aside from the collateral benefits of requiring admissions of wrongdoing, there is something anathema to the broader conception of justice to allow wrongdoers to acknowledge a modicum of guilt by settling but not have them actually admit that guilt or the harm they have wrought. It is for these reasons that the U.S. Department of Justice largely prevents a defendant from pleading *nolo contendere* (accepting a guilty plea without admitting or denying the allegations). U.S. DEP'T OF JUSTICE, U.S. ATTORNEYS' MANUAL § 9-16.010 (2008), http://www.justice.gov/usao/eousa/foia_reading_room/usam/title9/16mcrm.htm#9-16.010 (“United States Attorneys may not consent to a plea of *nolo contendere* except in the most unusual circumstances . . .”).

163. See *supra* Part III.

about 50% of cases would be settled. However, 98% of cases settle.¹⁶⁴ Defendants are avoiding trial for a reason.

As mentioned above, the costs of violating securities laws are a function of a sanction's swiftness, severity, and probability.¹⁶⁵ While the swiftness of punishment is relevant, and it is admitted that settlement is swifter as a general rule than litigation, this factor will be ignored as it is unlikely that the length of a trial will significantly alter a defendant's risk calculus. The severity of the sanctions here are a combination of monetary sanctions, the injunctive relief typically sought in settlements, and the extralegal consequences of a Commission action.¹⁶⁶ Given the assumptions stated at the outset of this Note—including that the monetary penalties imposed in settlements approach the level which would be achieved at trial—trials are not going to increase, in the aggregate, the amount of monetary sanctions imposed on a given defendant.¹⁶⁷ While a trial will force the defendant to expend not-insubstantial amounts of money on litigation costs, the expected sanction has to be discounted by the chance that the Commission will lose its case at trial. Over the run of cases, we can assume that this will largely be a wash. Even if it is not, given that increases in the monetary sanctions imposed on corporate defendants have proven relatively ineffective in deterring wrongful behavior, a relatively small change in the average amount of sanctions imposed can be assumed to have only a *de minimis* effect on defendant behavior.

Importantly, however, trials impose substantial costs that settlements do not. Because monetary sanctions against a publicly traded corporate defendant are essentially benign—they are penalties against the shareholders, not the decision makers—an effective regulatory policy needs to focus on the fourth factor of deterrence: extralegal harm. Trials wreak extralegal harm. Trials vilify, they expose, and they punish—all on an individual level. As outlined above, the current settlement regime is so favored by corporate defendants because it allows them to avoid (a) the hassles of litigation, (b) any *res judicata* or collateral effects, and (c) any real reputational or long-lasting harm.¹⁶⁸ Corporate defendants desperately want to avoid trial for all of these reasons—why else would they settle in such high numbers?—and so a more effective enforcement paradigm must focus on them in order to deter wrongdoing. If there were real reputational costs, on both a corporate and an individual level, to violating the securities laws, these costs would be internalized and according to the deterrence model laid out above, this would decrease the likelihood of such violations. More trials would impose such harm.

164. See *supra* note 16 and accompanying text.

165. Pasternoster, *supra* note 119, at 783.

166. See *supra* notes 20–22 and accompanying text.

167. See *supra* notes 57–60 and accompanying text.

168. See *supra* subpart III(B).

A potential counterargument to this is that while trials may increase the sanctions of Commission enforcement, due to the SEC's limited resources, bringing more trials will necessarily decrease the total number of cases brought, and therefore will not improve deterrence because the likelihood that sanctions will be imposed will fall as the number of trials brought increases.¹⁶⁹ This fear is overblown. First, an ideal solution would be to increase the Commission's budget proportionally to the increase in litigation costs.¹⁷⁰ However, even if we assume that the Commission's budget remains stable in the face of more trials, and even if this means that fewer investigations can be launched, this should not greatly impinge on the overall deterrent effects of the new policy. For one matter, the Commission could reprioritize its budget so that the cuts to pay for these trials would come from areas where securities violations are minor and do not fundamentally affect the health and stability of financial markets, such as stock-option backdating.¹⁷¹ Under this scenario, while it is true that fewer total cases would be brought, the same number of cases would be brought against the biggest violators of securities laws and those who perpetrated more egregious crimes, like outright fraud. More importantly, however, the effect of a sanctions regime depends not on the *objective* likelihood of sanction but on the *subjective*, or perceived, likelihood of sanction. While it is true that there is likely a correlation between these two levels,¹⁷² they are not identical, and the very public and excruciating nature of trials may actually increase the perceived likelihood of getting caught, even if it does not increase the objective likelihood of getting caught.

169. Director Khuzami argued this after Judge Rakoff handed down his opinion in the *Citigroup* case. Press Release, Robert Khuzami, Dir., Div. of Enforcement, U.S. Sec. & Exch. Comm'n, SEC Enforcement Director's Statement on Citigroup Case (Dec. 15, 2011), <http://www.sec.gov/news/press/2011/2011-265.htm> ("In contrast, the new standard adopted by the court could in practical terms press the SEC to trial in many more instances, likely resulting in fewer cases overall and less money being returned to investors.").

170. Recently, the Commission has highlighted its need for more resources. See, e.g., *Financial Services and General Government Appropriations for 2011: Hearings Before the Subcomm. on Fin. Servs. & Gen. Gov't Appropriations of the H. Comm. on Appropriations*, 111th Cong. 75–79 (2010) (statement of Mary Schapiro, Chairman, U.S. Sec. & Exch. Comm'n) (claiming that the SEC required additional resources to successfully adapt to the growing size and complexity of financial markets). An increase in the size of the Commission's budget should, by itself, have a positive effect on the overall effectiveness of the enforcement regime. See generally Howell E. Jackson & Mark J. Roe, *Public and Private Enforcement of Securities Laws: Resource-Based Evidence*, 93 J. FIN. ECON. 207 (2009) (arguing that public enforcement is more effective than private enforcement for ensuring financial markets' health).

171. For a discussion of the controversial practice of penalizing "backdating," see generally Jeffrey Benner et al., Moody's Investor Servs., *Stock Option "Backdating,"* MOODY'S SPECIAL COMMENT, June 2006, at 1 (explaining the uptick in prosecutions and the legal gray area of backdating).

172. See Paternoster, *supra* note 119, at 785 (noting a presumption among deterrence theorists that "there is a strong positive correlation between objective and subjective (perceptual) properties of punishment," but arguing that such correlation cannot be taken for granted).

In sum, the severity of sanctions imposed by the Commission's current settlement policy is minimal. Especially when the Commission brings suit against a corporation, rather than an individual, (a) the perpetrators of the harm typically keep their jobs or find lucrative jobs elsewhere in the field; (b) the share price usually goes up; (c) the executives still earn bonuses; and (d) the cost of the settlement is borne almost entirely by others (shareholders and creditors). Further, the injunctive relief obtained by the Commission is even less effective than these monetary sanctions because it is entirely ignored, even in the face of blatant repeat offenders.¹⁷³ As such, there is only minimal internalization of the harms of violating securities laws by the decision makers in an organization. Altering this calculus will not only affect individual decision making, but will likely affect institutional decision making—ideally leading to the adoption of stricter internal controls to avoid the negative collateral costs of trial.

Bringing more cases to trial by requiring admissions of guilt in settlements imposes real costs on these individuals—costs that the defendants are stridently seeking to avoid when they settle all of their cases. Finally, the increase in the costs internalized by defendants through taking more cases to trial will not be fully offset by an inability to bring as many cases, even if the Commission's budget is not concomitantly increased. This is because a spate of trials will drastically increase the *perception* that the securities laws are being enforced. Setting examples, not settling, is the key to this calculus.

Individual Liability.—An alternative method to achieve this same result—the internalization of the costs of violating securities laws by the individuals deciding to violate securities laws—is to impose direct and personal liability on those perpetrators. There are two ways to do this. One is to bring more trials against individuals, rather than corporations. As mentioned above, the constant criticism of the Commission that it does not bring cases against individuals is largely untrue, except when the Commission is prosecuting the largest financial institutions.¹⁷⁴ Bringing suits against individuals in these cases could have a substantial effect on the criminal calculus in these institutions. Second, Congress could statutorily inflict personal liability on the officers and executives of companies that are found liable for fraud by the Commission. The Dodd-Frank Act took a step in this direction by allowing clawbacks from executives who were “erroneously awarded compensation.”¹⁷⁵ This provision allows recovery, from any current or former executive officer, of any incentive-based compensation awarded during a three-year period preceding any reporting error that is in excess of what the executive would have otherwise received

173. See *supra* notes 132–38 and accompanying text.

174. See *supra* notes 30–32 and accompanying text.

175. Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub. L. No. 111-203, § 954, 124 Stat. 1376, 1904 (2010) (codified at 15 U.S.C. § 78j-4 (Supp. IV 2011)).

absent said reporting error.¹⁷⁶ While a step in the right direction, this solution suffers from the “house money effect.”¹⁷⁷ That is, in the wake of a prior gain—say winning a few hundred dollars on a hand of blackjack—people are far more willing to make a risky bet—risking the prior gain—than they would be with what they might regard as their own money.¹⁷⁸ This casino concept can be applied to the theory of clawing back compensation. The only money being risked for clawbacks is compensation that the officer would not have gained but for the reporting error. Because of this, there is little incentive not to misreport. All that is being risked is house money, and there is always the chance that the misrepresentation will not be caught.

This house money effect can further explain why paying officers in equity stakes—stock options—also does not make them more cautious when considering whether to take huge risks or violate securities laws. While common sense suggests that such equity stakes would make officers proceed cautiously before exposing themselves and their companies to huge risks—especially systemic risks, like those during the GFC, that financial institutions may not even have fully understood—equity stakes have proven ineffective at creating appropriate, risk-neutral incentives.¹⁷⁹ At the end of the day, the fraudsters are still gambling with anyone’s money but their own—be it from their prior gains or worse: from their investors or commercial account holders.

Instead of threatening officers with recouping their ill-gotten gains, a more effective deterrence regime would threaten an officer with losing some of his pre-reporting-error personal assets. While doing so may seem like a radical suggestion, it is not. In fact, for most of Wall Street’s history, bankers were personally (and collectively) liable for their actions. It was only in the 1970s, when investment banks began moving from a partnership to a corporate model, that bankers began being shielded from liability for their actions.¹⁸⁰ Perhaps not surprisingly, this switch coincided with Wall Street firms becoming much more risk prone in their activities, as well as with an uptick in the number and extent of securities laws violations.¹⁸¹ A

176. *Id.*

177. For a broader description of this premise, see generally RICHARD H. THALER, *QUASI RATIONAL ECONOMICS* (1991).

178. *See id.* at 49 (observing that “under some circumstances a prior gain can increase subjects’ willingness to accept gambles”).

179. *See* Claire Hill & Richard Painter, *Berle’s Vision Beyond Shareholder Interests: Why Investment Bankers Should Have (Some) Personal Liability*, 33 *SEATTLE U. L. REV.* 1173, 1173 (2010) (suggesting that investment bankers should be personally liable even for legal risks that they take with other people’s money).

180. *See id.* at 1177–78 (discussing the switch from partnership to corporate form).

181. *See id.* at 1181–82 (detailing how the switch from the partnership model affected firm profits and risk taking); *see also supra* Part II (describing the frauds and corresponding securities legislation that began becoming more and more frequent in the 1980s). *See generally* MICHAEL LEWIS, *LIAR’S POKER: RISING THROUGH THE WRECKAGE ON WALL STREET* (1989) (giving a first-hand account of the culture at Solomon Brothers shortly after it became a corporation).

return to something approaching this partnership model, by making officers of companies found liable for fraud by the Commission personally liable, would require these officers to internalize the full costs of their actions. This should have a profound effect, even if the extent of this personal liability is limited relative to each officer's total assets, and would likely encourage the adoption of stricter internal controls, and less risky behavior generally, by these regulated institutions. While it is true that such a regime would at times be unfair, with individuals who had no knowledge or ability to stop the misconduct being punished for it nonetheless, it is even more unfair to impose the costs of risky and illegal behavior on stockholders, creditors, and ultimately, society writ large.

Disregarding the corporate form to punish villainy is not an entirely novel idea. The common law has long permitted a wronged party to "pierce the corporate veil" in limited circumstances, as when it appears that the corporate form has been used fraudulently to conceal assets or illicitly screen an individual from liability.¹⁸² Moreover, in recent years there have been a few voices in the academy calling for the corporate shield to be tossed aside to prevent the insolvency of commercial and investment banks that are deemed "too big to fail." Peter Conti-Brown has suggested imposing shareholder liability to bail out systemically important financial institutions in case they become insolvent.¹⁸³ Claire Hill and Richard Painter sought to tackle this same problem of insolvency by advocating the imposition of limited personal liability on investment bankers for the debts of the bank.¹⁸⁴

However, in none of the above situations is the case for individual liability as strong as it is in the prosecution of individuals who have perpetrated securities fraud or other financial wrongdoing. First, the risk calculus here is more fundamentally askew. The rewards of lawbreaking for an individual grossly outstrip the threatened pain of an SEC fine on the employer and its shareholders. Second, unlike in the context of insolvency, the nexus between the punishment and the crime is tighter when dealing with illegality. Punishment is more justified when it comes in the wake of an actual crime—an actual transgression that society, through its legislature, has deemed worthy of prohibition—as opposed to simply nearsighted investment strategies or gross underfunding.

In the end, the goal of both of the approaches suggested by this Note is to force the individual decision makers in corporations to internalize the costs

182. See generally David H. Barber, *Piercing the Corporate Veil*, 17 WILLAMETTE L. REV. 371 (1981) (discussing piercing the corporate veil).

183. See generally Peter Conti-Brown, *Elective Shareholder Liability*, 64 STAN. L. REV. 409, 409 (2012) (proposing "elective shareholder liability," which requires that bank shareholders either structure the bank's capital to include less debt or have shareholders (as opposed to taxpayers) "cover the ultimate costs of the bank's failure").

184. See Hill & Painter, *supra* note 179, at 1173–74 (suggesting methods of making investment bankers partially liable in situations where the bank becomes insolvent).

of violating the securities laws by imposing unwanted liability on them—either extralegally through the reputational harm of trials or statutorily through the imposition of some form of personal liability. The American taxpayer has issued a blanket insurance policy to systemically important financial institutions. Through bailouts and lax enforcement of the securities laws, we have built our financial system in a complex hall of mirrors, not only tacitly allowing but functionally incentivizing fraud. It is plain that an investment banker who stands to gain millions from misreporting earnings or from marketing faulty, subprime mortgages will engage in that activity when neither he nor his company truly bear the risk of the downside of that gambit. If he is caught, the company is slapped on the wrist and perhaps he misses out on a bonus. If he is not, the potential reward is massive. With such a rigged game, we should have been less than surprised when the Global Financial Crisis hit in 2008. Without action, we should be less than surprised when it happens again.

VI. Conclusion

This Note set out to prove one central point: that the Securities and Exchange Commission's policy of settling virtually every case with a consent judgment is firmly—as Judge Rakoff himself stated—against the public interest of the United States. By settling, the Commission is undercutting the very securities laws it is tasked with enforcing. Both theory and empirical evidence convincingly demonstrate that settlements are not adequately deterring the violations of securities laws, and that instead, the largest financial institutions view them merely as a cost of doing business. Instead, the Commission—along with Congress, if need be—should adopt measures to make the individual decision makers at financial institutions fully internalize the costs of their behaviors, through bringing more cases to trial or, alternatively, through imposing personal liability on the executives of corporate defendants.

A former chairman of the Commission once famously threatened to leave a defendant “naked, homeless, and without wheels.”¹⁸⁵ While that may be hyperbole, the current system—which allows defendants to remain fully clothed, housed, and lets them keep their cars and executive parking spaces to boot—is not an acceptable alternative. Adjusting the lawbreaking calculus so that these officers transition from acting risk prone with house money to acting risk neutral will do volumes for the health and stability of the financial markets as a whole.

—Ross MacDonald

185. Jonathan Eisenberg, *Enforcement Issues and Litigation: Litigating With the SEC—A Reasonable Alternative to Settlement*, 21 SEC. REG. L.J. 421, 421 (1994).

Blowing the Whistle on Civil Rights: Analyzing the False Claims Act as an Alternative Enforcement Method for Civil Rights Laws*

Traditional antidiscrimination laws do not effectively deter or remedy civil rights violations by local governments and related entities. The federal government lacks the resources to litigate more than a limited number of large discrimination cases at one time.¹ Individually injured civil rights litigants face significant statutory and court-imposed limitations on making discrimination claims against local governments.² Public interest litigators advocating for institutional change face problems of standing³ and the challenge of pushing for broader social change while remaining loyal to individual clients.⁴

As a few commentators and litigators have recognized, albeit in limited circumstances,⁵ the False Claims Act (FCA)⁶ offers antidiscrimination civil rights litigators a powerful alternative path for litigating against discrimination, and possibly other civil rights violations, by local

* Thank you to the Assistant U.S. Attorneys in the Houston Civil Division who introduced me to the False Claims Act, to Professor Jim Harrington and the 2011 Constitutional Litigation seminar students who showed me good use for the FCA, and to the *Texas Law Review* editors who made my idea presentable. I want to thank my parents, Alan and Cathie Mayrell, for teaching me the importance of speaking up for those who do not have a voice, and Ms. Toby Nix for first showing me how. And most of all thank you to Veronika Bordás, who makes every day bright and keeps every *id.*'s period italicized.

1. The Justice Department's Civil Rights Division lost nearly 70% of its lawyers during the Bush Administration, and even with its recent move to restore staffing, the Division had only filed twenty-nine civil employment discrimination cases as of fourteen months after President Obama took office. Joel Wm. Friedman, *The Impact of the Obama Presidency on Civil Rights Enforcement in the United States*, 87 IND. L.J. 349, 358–59 (2012); see also Kitty Calavita, *The Struggle for Racial Justice: The Personal, the Political, and . . . the Economic*, 44 LAW & SOC'Y REV. 495, 499 (2010) (observing that the Department of Justice prosecuted very few criminal civil rights cases under Presidents Clinton and Bush); Rachel A. Harmon, *Promoting Civil Rights Through Proactive Policing Reform*, 62 STAN. L. REV. 1, 3 (2009) (discussing how the Department of Justice devotes only limited resources to police department institutional change litigation under 42 U.S.C. § 14141 (2006)).

2. See *infra* Part I.

3. See *infra* Part I.

4. Note, *The Plaintiff as Person: Cause Lawyering, Human Subject Research, and the Secret Agent Problem*, 119 HARV. L. REV. 1510, 1513–16 (2006) (contrasting “traditional” lawyering, which discourages lawyers from “serving two masters,” with “cause lawyering,” which encourages lawyers to advocate for “a more just society” in their litigation); see also Derrick A. Bell, Jr., *Serving Two Masters: Integration Ideals and Client Interests in School Desegregation Litigation*, 85 YALE L.J. 470, 482–86 (1976) (discussing cases where the objectives of the lawyers and their clients in school segregation cases were not aligned).

5. See *infra* note 21.

6. 31 U.S.C. §§ 3729–33 (2006 & Supp. IV 2011).

governments. The FCA provides for a private claim against federal government contractors and grant recipients, including local governments, when those entities violate the terms of their contracts or grants.⁷ Under the FCA, a private person can bring a claim on behalf of the United States and, in exchange, receive a large fraction of treble contract damages plus civil monetary penalties and attorneys' fees.⁸ Civil rights litigants can use the FCA because counties, cities, police departments, local hospitals, and public schools receive federal grant money.⁹ As a condition for receiving and keeping those grants, local governments agree to comply with civil rights laws, including laws against employment discrimination and discrimination in how the entities provide services paid for with federal grant money.¹⁰ FCA claims arguably exist against local governments when they fail to comply with antidiscrimination grant terms while either continuing to request disbursements of federal grants or while retaining the federal money they have already received.¹¹

The FCA offers significant benefits to civil rights plaintiffs. Plaintiffs' (called "relators" in the FCA) damages, can be quite large—up to 30% of a maximum of triple the value of the contract or grant—and plaintiffs can also take away per-claim civil penalties as well as ask for attorneys' fees.¹² These significant damages should incentivize private attorneys to litigate these claims as well as disincentivize government entities from violating antidiscrimination statutes.¹³ And because the injured party under the FCA is

7. 31 U.S.C. § 3729(a)–(b); see also Patricia Meador & Elizabeth S. Warren, *The False Claims Act: A Civil War Relic Evolves into a Modern Weapon*, 65 TENN. L. REV. 455, 458–61 (1998) (explaining the FCA's history).

8. See 31 U.S.C. § 3729(a)(1) (providing for treble damages and civil monetary penalties); *id.* § 3730 (providing that private persons can sue on behalf of the United States). In 2011, the civil monetary penalties were increased to a maximum of \$11,000 per false claim filed. 28 C.F.R. § 85.3(9) (2011).

9. *Who Is Eligible for a Grant?*, GRANTS.GOV, <http://www.grants.gov/aboutgrants/eligibility.jsp>. Local governments received approximately \$68 billion from intergovernmental funds provided by the federal government according to a 2009–2010 Census Bureau survey of local governments. BUREAU OF THE CENSUS, U.S. DEP'T OF COMMERCE, STATE AND LOCAL GOVERNMENT FINANCES BY LEVEL OF GOVERNMENT AND BY STATE: 2009–10 (2012), available at <http://www2.census.gov/govs/estimate/10slsstab1a.xls>.

10. See, e.g., *infra* subpart II(A); see also Memorandum from the Att'y Gen. to Heads of Exec. Dep'ts & Agencies Providing Federal Financial Assistance, Enforcement of Nondiscrimination Laws in Programs and Activities that Receive American Recovery and Reinvestment Act Funding (Sept. 27, 2010), available at http://www.justice.gov/crt/about/cor/arra_memo.pdf (reinforcing to federal agencies the importance of enforcing antidiscrimination laws in American Recovery and Reinvestment Act (ARRA) grants).

11. See *infra* subpart II(C).

12. See *infra* section II(B)(2).

13. This Note's argument rests on the assumption that civil rights damages incentivize changes in defendant behavior. See generally, e.g., Myriam E. Gilles, *In Defense of Making Government Pay: The Deterrent Effect of Constitutional Tort Remedies*, 35 GA. L. REV. 845 (2001) (arguing that civil rights damages have a deterrent effect). But see, e.g., Joanna C. Schwartz, *Myths and*

the United States,¹⁴ institutional-change litigants do not face the standing problems they otherwise have to overcome under laws based on remedying individual injuries.¹⁵

The FCA provides civil rights litigators with another avenue for enforcing antidiscrimination laws, but it also comes with risks. This Note argues for using the FCA to defend civil rights to the benefit of discrimination victims.¹⁶ It also argues that agencies should use their flexibility in contracting to expand the civil rights requirements of contractors to include requirements of compliance with constitutional norms appropriate for the recipient agency.¹⁷ Despite this Note's optimistic view of increasing damages against civil rights violators, there are risks to increasing the size of damages. Greater damages hurt local government coffers despite the Court's and Congress's professed desire to protect local governments from punitive damages in the civil rights context.¹⁸ Larger damages for relators also could discourage the worthy goal of reconciliation between the injured party and the local government.¹⁹ Furthermore, these penalties could decrease the local government's willingness to admit wrongdoing in traditional civil rights disputes because they will know that reconciliation and settlement would not bar future FCA claims by third parties based on those admissions.²⁰ This Note argues that despite these risks, when used judiciously by litigants, the FCA can play a useful role where individual remedies do not suffice or institutional-change litigants lack standing to apply pressure.

Some commentators in academia have recognized the potential to use the FCA to defend civil rights.²¹ Of those few pieces, no article has stated a

Mechanics of Deterrence: The Role of Lawsuits in Law Enforcement Decisionmaking, 57 UCLA L. REV. 1023, 1028, 1045–52 (2010) (presenting evidence that many law enforcement officials lack sufficient information to connect civil rights lawsuits to behaviors and policies).

14. 31 U.S.C. § 3730(b) (2006).

15. See *Lujan v. Defenders of Wildlife*, 504 U.S. 555, 560–61 (1992) (explaining the constitutional standing requirements for individually injured plaintiffs).

16. See *infra* Part II.

17. See *infra* Part III.

18. See *infra* subpart IV(A).

19. See *infra* subpart IV(B).

20. See *infra* subpart IV(B).

21. See generally Stephen F. Hayes, *Enforcing Civil Rights Obligations Through the False Claims Act*, 1 COLUM. J. RACE & L. 29 (2011) (discussing theoretical issues about the application of the FCA in civil rights and institutional change litigation); Dayna Bowen Matthew, *A New Strategy to Combat Racial Inequality in American Health Care Delivery*, 9 DEPAUL J. HEALTH CARE L. 793 (2005) [hereinafter Matthew, *Health Care*] (proposing using the FCA to protect civil rights in health care); Dayna Bowen Matthew, *Disastrous Disasters: Restoring Civil Rights Protections for Victims of the State in Natural Disasters*, 2 J. HEALTH & BIOMEDICAL L. 213 (2006) [hereinafter Matthew, *Public Health*] (proposing using the FCA to protect rights in public health); Jan P. Mensz, *Citizen Police: Using the Qui Tam Provision of the False Claims Act to Promote Racial and Economic Integration in Housing*, 43 U. MICH. J.L. REFORM 1137 (2010) (discussing using the FCA to enforce

complete legal theory of FCA liability based on violations of antidiscrimination laws broadly defined. Nor has the literature thoroughly addressed the hazards of using the FCA to increase the damages against local governments in civil rights litigation. This Note is the first to present a detailed analysis of FCA-specific legal issues. In the current literature, only one recent article addresses the theoretical benefits of using the FCA to litigate for institutional change, and it does not address the difficulties involved in bringing a civil rights FCA claim.²² Other earlier works propose a very limited application of the FCA to enforce Title VI disparate impact claims in the contexts of health care,²³ public health,²⁴ and housing discrimination as well as the use of the FCA in the context of First Amendment Establishment Clause claims.²⁵ They do not expand into antidiscrimination laws more broadly or address in detail the legal issues involved in using the FCA in the civil rights context.²⁶ Rather than looking at the theoretical problems of institutional-change litigation or narrow cases where the FCA could be applied, this Note addresses the legal difficulties involved in bringing an FCA claim based upon a broad swath of antidiscrimination laws in a wide range of settings.²⁷ Furthermore, this Note proposes that executive agencies expand the scope of contractual protections for civil rights to include relevant constitutional protections.²⁸ Finally, this

the Fair Housing Act); Matthew J. Termine, *Promoting Residential Integration Through the Fair Housing Act: Are Qui Tam Actions a Viable Method of Enforcing "Affirmatively Furthering Fair Housing" Violations?*, 79 *FORDHAM L. REV.* 1367 (2010) (discussing using the FCA to enforce the Fair Housing Act); Randall M. Levine, Note, *Enforced Separation: Utilizing the False Claims Act to Prosecute Government Contractors Spending Federal Funds in Violation of Church/State Regulations*, 35 *PUB. CONT. L.J.* 155 (2005) (proposing using the FCA to protect First Amendment rights).

22. See generally Hayes, *supra* note 21 (discussing the theoretical arguments surrounding the application of the FCA in civil rights and institutional-change litigation).

23. Matthew, *Health Care*, *supra* note 21, at 822.

24. Matthew, *Public Health*, *supra* note 21, at 234.

25. Levine, *supra* note 21, at 157–58; Mensz, *supra* note 21, at 1139.

26. See generally Mensz, *supra* note 21 (expounding upon the use of the FCA to enforce racial and economic integration policy); Termine, *supra* note 21 (discussing the use of *qui tam* actions to enforce FHA duties under the FCA).

27. See *infra* Part II.

28. See *infra* Part III. Articles have suggested the narrow use of conditioned grants from the Department of Justice to states and local police departments to incentivize state and local regulations to encourage compliance with constitutional rights. See, e.g., Kami Chavis Simmons, *Cooperative Federalism and Police Reform: Using Congressional Spending Power to Promote Police Accountability*, 62 *ALA. L. REV.* 351, 383–85 (2011) (proposing a program to partially condition Department of Justice Community Oriented Policing funds upon entities improving accountability for police misconduct). These articles have not broadly proposed conditioning funding upon complete compliance with appropriate constitutional rights requirements on the assumption that private individuals would enforce these protections via the FCA.

Note uniquely addresses the hazards of using the FCA to increase civil rights liability.²⁹

In a few limited situations, litigators have already attempted to use the FCA, with varying degrees of success, to sue for violations of laws such as the Fair Housing Act, a case in which the defendants settled for \$52 million,³⁰ and the Rehabilitation Act, a case which, as of the time of writing, remains in a district court battle of the pleadings.³¹ These narrowly focused cases have demonstrated that the idea of using the FCA as a mechanism to defend civil rights can work. Because one case settled and the other is ongoing, however, these cases do not offer prospective litigators insights into the legal difficulties of civil rights FCA actions.

This Note sets out to explain the legal theory through which civil rights litigators can effectively litigate claims against local government discriminators using the FCA. Part I briefly outlines the scheme of antidiscrimination laws and regulations that are potentially enforceable under the FCA and their limitations. Part II lays out the legal theory of how an antidiscrimination action could form the basis of an FCA claim and provides recent examples of courts favorably reacting to plaintiffs' use of the FCA in civil rights suits. Part III briefly proposes that agencies use their contracting flexibility to add relevant constitutional requirements. Part IV discusses the potential legal and policy hazards of using the FCA to increase the liability of local governments for civil rights violations.

I. Current Civil Rights Schemes Are Limited by Restrictions on Damages and Standing

Many types of antidiscrimination schemes suffer from significant constraints when used as the basis for litigation against local governments and similar entities. Limitations on damages generally,³² specific bars

29. See *infra* Part IV.

30. Mensz, *supra* note 21, at 1148 & n.69.

31. See *United States ex rel. Gillespie v. Kaplan Univ.*, No. 09-20756-CIV, 2012 WL 1852085 (S.D. Fla. May 21, 2012) (order denying motion to dismiss and motion for judgment on the pleadings); *United States ex rel. Gillespie v. Kaplan Univ.*, No. 09-20756-CIV, 2012 WL 1852159 (S.D. Fla. May 21, 2012) (order granting in part motion for leave to amend); *United States ex rel. Diaz v. Kaplan Univ.*, No. 09-20756-CIV, 2011 WL 3627285 (S.D. Fla. Aug. 17, 2011) (order granting in part and denying in part a motion to dismiss).

32. 42 U.S.C. § 1981a(b)(1) (2006) (limiting punitive damages available under Title VII, the ADA, and the Rehabilitation Act to situations in which the defendant acted "with malice or with reckless indifference to the federally protected rights of an aggrieved individual"); *id.* § 1981a(b)(3) (limiting the sum of compensatory and punitive damages to \$300,000 for employers of 501 or more employees and lower amounts for smaller employers). Damages are also small on average. Among ADA claims brought to the EEOC, the average payout in benefits per "merit resolution[]" through the EEOC process, including unsuccessful though meritorious claims but excluding money recovered in litigation, is about \$14,525. See *Americans with Disabilities Act of 1990 (ADA) Charges (Includes Concurrent Charges with Title VII, ADEA, and EPA): FY 1997 - FY 2011*, U.S.

against punitive damages in claims against local governments,³³ problems of standing for non-injured institutional change litigants,³⁴ theoretical problems measuring actual damages,³⁵ and the inability to sue for violations of many civil rights regulations³⁶ all restrict the ability of civil rights litigants to effectively deter behaviors that harm rights. This Part will briefly address the limitations of claims against local governments under employment antidiscrimination laws, service discrimination laws, and constitutional torts.

A. *Employment Antidiscrimination Laws Restrict Damages Against Local Governments*

Several laws prohibit discrimination in employment by private employers and governments. Title VII forbids employment discrimination based on “race, color, religion, sex, or national origin,”³⁷ the Americans with Disabilities Act (ADA) bans discrimination based upon disability,³⁸ and the Age Discrimination in Employment Act (ADEA) limits the freedom of employers to discriminate based on age.³⁹ The ADA’s precursor, the Rehabilitation Act (also known as Section 504), requires that recipients of federal money provide equal treatment in employment to the disabled.⁴⁰ Together, these laws cover a range of class-based discrimination.

Each of these antidiscrimination statutory schemes provides for damages remedies,⁴¹ but these remedies suffer from significant limitations. Many of these schemes limit the sum of punitive and compensatory damages

EQUAL EMP’T OPPORTUNITY COMM’N, <http://www.eeoc.gov/eeoc/statistics/enforcement/ada-charges.cfm> (reporting 5,239 “merit resolutions” and \$76.1 million in monetary benefits for fiscal year 2010).

33. See 42 U.S.C. § 1981a(b)(1) (prohibiting punitive damages against “a government, government agency or political subdivision”).

34. See *Alexander v. Sandoval*, 532 U.S. 275, 285–86, 293 (2001) (holding that Title VI regulations banning activity beyond that explicitly limited by Title VI itself do not create private causes of action); *id.* at 280, 292 (criticizing the effort to say that Congress ratified private causes of action based on regulations by pointing to language in the Rehabilitation Act Amendments of 1986 § 1003, 42 U.S.C. § 2000d-7(a)(2) (2006), which the Court emphasized refers to foreclosing state immunity for violations of *statutes*).

35. See, e.g., *Carey v. Piphus*, 435 U.S. 247, 266–67 (1978) (stating that, on remand, respondents could not recover damages of more than one dollar for deprivation of a constitutional right). See generally John G. Niles, Comment, *Civil Actions for Damages Under the Federal Civil Rights Statutes*, 45 TEXAS L. REV. 1015 (1967) (exploring the problems inherent in assessing monetary damages for infringement of civil rights).

36. See, e.g., *Alexander*, 532 U.S. at 285–86, 293 n.8 (questioning “whether authorization of a private right of action to enforce a statute constitutes authorization of a private right of action to enforce regulations that go beyond what the statute itself requires”).

37. 42 U.S.C. § 2000e-2(a).

38. *Id.* § 12112(a) (2006 & Supp. IV 2011).

39. 29 U.S.C. § 623 (2006).

40. *Id.* § 794(a).

41. E.g., 42 U.S.C. § 1981a (2006) (providing that damages are available in cases of intentional employment discrimination).

against private employers⁴² and expressly forbid punitive damages against “a government, government agency or political subdivision.”⁴³ A discriminatory act that does not result in actual damages that can be characterized as compensatory will not lead to damages even in cases where measurable harm occurs.⁴⁴

Employment discrimination laws also come with procedural hurdles. Perhaps most important is the requirement of individual injury, which translates to standing.⁴⁵ As a result, groups interested in changing employers’ behaviors need individually injured plaintiffs willing to litigate against the employer. Outside organizations often cannot initiate an action on their own without an employee who wants to play ball, risking his career and livelihood.

B. Service Antidiscrimination Laws Similarly Limit Damages Against Municipalities

Service antidiscrimination laws forbid discrimination in the provision of services funded by federal government grants and contracts. Title VI forbids federal recipients of grant money from discriminating based on race or national origin when they provide services using those funds.⁴⁶ The Rehabilitation Act likewise bans this type of discrimination based on disability,⁴⁷ and the Age Discrimination Act (distinct from the ADEA) addresses service discrimination based on age.⁴⁸ Title IX addresses service discrimination based on gender, but only by higher educational institutions receiving federal grant money.⁴⁹ Several other specific statutes also contain specific prohibitions on sex discrimination.⁵⁰

42. *See supra* note 32.

43. 42 U.S.C. § 1981a(b)(1).

44. *See, e.g., Kerr-Selgas v. Am. Airlines, Inc.*, 69 F.3d 1205, 1214 (1st Cir. 1995) (“[G]enerally a claimant may not recover punitive damages without establishing liability for either compensatory or nominal damages.”); *see also* RESTATEMENT (SECOND) OF TORTS § 908 cmt. b (1979) (stating that an independent cause of action must exist in order to justify punitive damage awards).

45. *E.g., Thompson v. N. Am. Stainless*, 131 S. Ct. 863, 869–70 (2011) (limiting Title VII standing to injured parties within “the zone of interests” of Title VII’s statutory goals of preventing employment discrimination).

46. 42 U.S.C. § 2000d.

47. 29 U.S.C. § 794(a) (2006).

48. 42 U.S.C. § 6102.

49. 20 U.S.C. § 1681 (2006).

50. *See, e.g., 31 U.S.C. § 6711* (2006) (banning the full gamut of class-based discrimination by local governments receiving money from the Local Government Fiscal Assistance Fund). A search for (“be denied the benefits of” /25 sex) in WestlawNext’s United States Code Annotated database returns thirty statutes that operate similarly.

Title VI, the Rehabilitation Act, the Age Discrimination Act, and Title IX do not permit punitive damages.⁵¹ Like employment antidiscrimination laws, institutional-change litigators need individually injured plaintiffs or an injured class of plaintiffs in order to litigate under these laws.⁵² The Court has imposed another roadblock in the context of service antidiscrimination, blocking many suits based upon related regulations.⁵³ This can make it impossible to litigate claims deriving from violations of regulations created to enforce these statutes.

C. General Civil Rights Laws Like Section 1983 Impose Significant Barriers for Plaintiffs

Section 1983 provides the primary pathway for suing local government employees as well as local government entities for constitutional and statutory violations done under the color of law.⁵⁴ While Section 1983 serves as the broadest sweeping civil rights scheme, extensive Supreme Court doctrine has weakened its power by disallowing punitive damages against local governments,⁵⁵ by requiring plaintiffs to overcome the high hurdle of qualified immunity,⁵⁶ and by other means.⁵⁷

51. See *Barnes v. Gorman*, 536 U.S. 181, 188–90 (2002) (holding that punitive damages cannot be read into Title VI and stating that the Age Discrimination Act and the Rehabilitation Act also do not include punitive damages). Courts currently disagree about whether *City of Newport v. Fact Concerts, Inc.*, 453 U.S. 247, 271 (1981), which limits punitive damages against municipalities in the context of Section 1983, applies to Title IX claims against local governments. For a more extensive discussion of the confusion surrounding this question, see Katrina A. Pohlman, Note, *Have We Forgotten K-12? The Need for Punitive Damages to Improve Title IX Enforcement*, 71 U. PITT. L. REV. 167, 172–78 (2009).

52. Eric J. Kuhn, *Standing: Stood Up at the Courthouse Door*, 63 GEO. WASH. L. REV. 886, 891 (1995).

53. See *supra* note 34.

54. 42 U.S.C. § 1983 (2006); see also Richard Frankel, *Regulating Privatized Government Through § 1983*, 76 U. CHI. L. REV. 1449, 1452 (2006) (describing Section 1983 as “the primary vehicle for protecting individuals from violations of their constitutional and federal statutory rights by state actors”).

55. *Fact Concerts*, 453 U.S. at 27.

56. See *Pierson v. Ray*, 386 U.S. 547, 554–55 (1967) (reading a qualified immunity defense into Section 1983).

57. See, e.g., *Connick v. Thompson*, 131 S. Ct. 1350, 1360, 1362–64 (2011) (making it more difficult under Section 1983 to show a pattern or practice of rights-violating behavior and demonstrate a failure to train); *Ashcroft v. Iqbal*, 129 S. Ct. 1937, 1950–52 (2009) (raising the pleading standard for Section 1983 claims); *Gonzaga Univ. v. Doe*, 536 U.S. 273, 284–85 (2002) (explaining that plaintiffs only can use Section 1983 to enforce clearly established rights intended to have remedies rather than laws more generally); *Fact Concerts*, 453 U.S. at 271 (holding that municipalities are generally not liable for punitive damages under Section 1983); *Edelman v. Jordan*, 415 U.S. 651, 674–77 (1974) (restating that states have sovereign immunity from Section 1983 claims for damages); see generally Ivan E. Bodensteiner, *Congress Needs to Repair the Court's Damage to § 1983*, 16 TEX. J. C.L. & C.R. 29 (2010) (identifying the disabling limits placed upon Section 1983). Circuit courts have also made it more difficult for plaintiffs to succeed on Section 1983 claims. See, e.g., *Meadours v. Ermel*, 483 F.3d 417, 422–23 (5th Cir. 2007)

Local government employers often insure their individual employees against Section 1983 liability,⁵⁸ passing the cost of civil rights violations indirectly to the local government through increased liability insurance premiums and possibly surplus liability beyond coverage maximums.⁵⁹ Separately, the Court's highly restrictive doctrine of qualified immunity, which places extremely high procedural and factual hurdles in front of plaintiffs, makes suing individual violators very difficult as this standard requires everything from heightened pleading to complex showings of knowledge and responsibility by the officer.⁶⁰

Section 1983 also permits direct suits against the local government and supervisors, but aside from the limitations on punitive liability,⁶¹ the Court has raised the hurdle even higher than for claims against individuals. The plaintiff must demonstrate that the supervisors or supervising entity had a policy or effective policy of violating a constitutional right before they can be liable, a standard that recent opinions have made next to insurmountable.⁶²

II. Traditional Antidiscrimination Laws Can Form the Basis of an FCA Claim as Federal Grant Recipient Contracts Are Already Written

Violations of employment and service antidiscrimination laws and regulations have the potential to also breach federal grant and contract requirements, and this breach, in turn, creates FCA liability. Federal contractual terms expressly condition receiving and retaining federal grants and contracts upon compliance with a host of antidiscrimination laws as well as sometimes upon additional civil rights terms included in the contract.⁶³ Section 1983, however, cannot form the basis of an FCA claim. Section 1983 provides a mechanism to sue for the denial of a constitutional or statutory right, but creates no rights in and of itself.⁶⁴ This Part will provide an example of federal contracts with civil rights terms and lay out the

(describing the two-step analysis used to determine whether state officials receive qualified immunity), *overruled on other grounds by* *Bustos v. Martini Club Inc.*, 599 F.3d 458, 463 (5th Cir. 2010) (addressing a separate Texas tort law issue).

58. See generally Martin A. Schwartz, *Should Juries Be Informed that Municipality Will Indemnify Officer's § 1983 Liability for Constitutional Wrongdoing?*, 86 IOWA L. REV. 1209 (2001) (discussing local government indemnification of employees for Section 1983 violations).

59. See Theodore Eisenberg & Stewart Schwab, *The Reality of Constitutional Tort Litigation*, 72 CORNELL L. REV. 641, 651 (1987) (attributing increased liability insurance premiums to constitutional tort claims).

60. See *supra* note 56–57.

61. See *supra* note 55.

62. E.g., *Connick*, 131 S. Ct. at 1359–60, 1362–64 (making it more difficult under Section 1983 to show a pattern or practice of rights-violating behavior and demonstrate a failure to train by the municipality).

63. See, e.g., *infra* subpart II(A).

64. *Chapman v. Hous. Welfare Rights Org.*, 441 U.S. 600, 617 (1979) (“[Section] 1983 by itself does not protect anyone against anything.”).

structure of an FCA claim based on the violation of an antidiscrimination law. The last subpart will discuss the few major examples of FCA civil rights claims.

A. COPS—An Example of a Federal Grant with Antidiscrimination Terms

As an example contract and program, this Note will use the Department of Justice's Community Oriented Policing Services (COPS) grants.⁶⁵ These grants fund the hiring of police officers and the creation of crime-prevention programs.⁶⁶ The program has provided⁶⁷ and continues to deliver⁶⁸ grants to a large fraction of American cities, and the grants are often substantial in size.⁶⁹ While COPS serves as a useful, simple example of a federal contract with antidiscrimination terms, bear in mind that huge numbers of local governments and related entities like hospitals, public schools, community organizations, local commerce organizations, and regional transportation agencies also receive federal funding.⁷⁰ Many of those grants require that the grant recipient certify nondiscrimination in employment and the provision of services. These requirements can be stated expressly within the contract or by reference to the relevant laws and regulations.⁷¹ They can also be implied

65. See generally OFFICE OF CMTY. ORIENTED POLICING SERVS., U.S. DEP'T OF JUSTICE, 2009 COPS HIRING RECOVERY PROGRAM GRANT OWNER'S MANUAL (2009) [hereinafter COPS GRANT MANUAL], available at http://www.cops.usdoj.gov/pdf/chrp_gom.pdf (explaining the COPS grant program).

66. *Id.* at 5.

67. Forty-four percent of the U.S. population received police protection in 2000 from a department receiving a COPS grant in the period between 1994 and 2001. William N. Evans & Emily G. Owens, *COPS and Crime*, 91 J. PUB. ECON. 181, 186–87 (2007). At least 98% of cities larger than 250,000 people received a COPS officer-hiring grant in that period. *Id.* at 188. The average total of COPS hiring grants in the period given to cities with populations over 250,000 people equaled, in round numbers, about \$21 million. See *id.* at 188 (multiplying \$165 million per year by eight years for the period of the program studied and dividing by the fraction of cities receiving grants in the period, which equals about 98.4% of 61 cities).

68. Press Release, U.S. Dep't of Justice, US Department of Justice COPS Office Awards over \$243 Million to Hire New Officers (Sept. 28, 2011), available at <http://www.cops.usdoj.gov/Default.asp?Item=2600> (reporting that 238 law enforcement agencies and municipalities received COPS hiring grants worth a total of \$243 million).

69. See *id.* (dividing \$243 million by 238 grant recipients equals an average grant of slightly less than \$1 million).

70. See *supra* note 9; Edward T. Canuel, *Supporting Smart Growth Legislation and Audits: An Analysis of U.S. and Canadian Land Planning Theories and Tools*, 13 MICH. ST. J. INT'L L. 309, 317 (2005) (discussing federal funding of transportation projects); *The Greater New York Chamber of Commerce Along with the Business and Labor Coalition of New York (BALCONY) and the New York Hispanic Chamber of Commerce Have Been Awarded a Federal Grant to Help Small Businesses with Health Care and Health Insurance Issues*, GREATER N.Y. CHAMBER COM. (Apr. 12, 2012), http://www.ny-chamber.com/news_detail.asp?id=61 (describing a federal grant given to the Greater New York Chamber of Commerce to conduct health care insurance workshops).

71. See, e.g., COPS GRANT MANUAL, *supra* note 65, at 63 (delimiting antidiscrimination requirements in text).

from federal statutes and regulations that condition receipt of funding upon compliance with the law.⁷²

Many programs include express requirements of compliance with nondiscrimination provisions. The COPS Grant Manual includes the following requirement for grant recipients:

[The recipient] will not, on the ground of race, color, religion, national origin, gender, disability or age, unlawfully exclude any person from participation in, deny the benefits of or employment to any person, or subject any person to discrimination in connection with any programs or activities funded in whole or in part with federal funds.⁷³

The COPS's Grant Manual also asserts that "[t]he COPS Office has the right to sanction or terminate your agency's project when there is reason to believe that your agency[, for example,] [i]s not substantially complying with the grant requirements or other applicable provisions of federal law."⁷⁴

Agency-specific regulations can also create requirements for grant recipients.⁷⁵ "[S]uch assurance [of compliance with nondiscrimination requirements] shall obligate the recipient for the period during which Federal financial assistance is extended pursuant to the application."⁷⁶ DOJ regulations include local government activities of divisions not receiving the grant "if the policies of such other department, agency, or office will substantially affect the project for which Federal financial assistance is requested."⁷⁷ Importantly, funding is conditioned on compliance:

If there appears to be a failure or threatened failure to comply with this subpart and if the noncompliance or threatened noncompliance cannot be corrected by informal means, the responsible Department official may suspend or terminate, or refuse to grant or continue, Federal financial assistance, or use any other means authorized by law, to induce compliance with this subpart.⁷⁸

The takeaway: statutes, as discussed in this subpart, as well as general contracting regulations, agency-specific regulations, and the grant contracts themselves all contain language directly conditioning grant funding upon compliance with various antidiscrimination requirements.

72. See, e.g., 29 U.S.C. § 794(a) (2006) (banning discrimination based on disability by recipients of federal funds); 41 C.F.R. § 60-1.4(a)(1) (2011) (requiring federal contractors to comply with employment antidiscrimination laws).

73. COPS GRANT MANUAL, *supra* note 65, at 63.

74. *Id.* at 26.

75. E.g., 28 C.F.R. § 42.105 (2012) (listing requirements specific to the Department of Justice).

76. *Id.* § 42.105(a)(1).

77. *Id.* § 42.105(b).

78. *Id.* § 42.108(a).

B. FCA Qui Tam Procedure and Damages

1. *FCA Procedure.*—The FCA creates a private cause of action against contractors and grant recipients, including local governments, hospitals, and businesses (but not states)⁷⁹ that have violated the terms of a federal contract.⁸⁰ In the typical FCA claim, a whistle-blower with specific insider knowledge of a fraud on the government, called a “relator[],”⁸¹ files a claim, called a “*qui tam*,”⁸² in federal court against the defendant on behalf of the United States.⁸³ The relator serves the claim on the United States under seal.⁸⁴ The law gives the government time to investigate the claim.⁸⁵ The government then decides whether to intervene and take over the litigation, leave the litigation to the relator, or ask the court to dismiss the relator’s claim.⁸⁶ If the government intervenes, then it and the relator negotiate their respective roles in the litigation.⁸⁷ The court then unseals the complaint and the relator serves it on the defendant.⁸⁸ At this point, the case resembles a fraud claim and uses the heightened Rule 9 pleading standard of fraud.⁸⁹

In a civil rights FCA claim, the relator would provide the government information collected either as an insider at the defendant organization or, in light of recent amendments to the FCA, as an outsider with insights into the defendant organization garnered through other litigation.⁹⁰ It is possible that even information obtained by private auditors and investigators in their professions, so long as it is neither stored in a federal database nor reported on the news, would now suffice as the basis for a claim.⁹¹ The relator, for example, could provide evidence that the local government had taken federal money to provide enhanced community policing, but the program funded by

79. *Vt. Agency of Natural Res. v. United States ex rel. Stevens*, 529 U.S. 765, 787–88 (2000) (holding that states and state agencies are not “person[s]” liable under the FCA).

80. 31 U.S.C. § 3729(a)(1) (2006 & Supp. IV 2011).

81. Thomas R. Lee, Comment, *The Standing of Qui Tam Relators Under the False Claims Act*, 57 U. CHI. L. REV. 543, 543 (1990).

82. “‘Qui tam’ is short for ‘qui tam pro domino rege quam pro se imposito sequitur,’ meaning ‘who brings the action as well for the king as for himself.’” *Id.* at 543 n.4.

83. 31 U.S.C. § 3730(b)(1) (2006).

84. *Id.* § 3730(b)(2).

85. *Id.* § 3730(a), (b)(3).

86. *Id.* § 3730(b)(4), (c)(2)(A).

87. *Cf. id.* § 3730(c)(2)(C), (D) (providing that limitations may be imposed on the relator’s participation).

88. *Id.* § 3730(b)(2), (3).

89. *United States ex rel. Thompson v. Columbia/HCA Healthcare Corp.*, 125 F.3d 899, 903 (5th Cir. 1997).

90. Chris S. Stewart, Note, *Resourceful Relators: The Rise of Qui Tam Suits Under the False Claims Act Based on Information Obtained in Civil Litigation*, 89 TEXAS L. REV. SEE ALSO 169, 169 (2010).

91. Beverly Cohen, *KABOOM! The Explosion of Qui Tam False Claims Under the Health Reform Law*, 116 PENN ST. L. REV. 77, 99–100 (2011).

the federal government had systematically engaged in racial discrimination. It would serve this information in a sealed document on the government and try to persuade the Civil Division of the appropriate U.S. Attorney's Office to intervene, as cases with government intervention have a far higher success rate for the relator if also a lower maximum fraction of the winnings for the relator.⁹² In all likelihood, the U.S. Attorney's Office will decline to intervene because the dollars at stake in a civil rights FCA claim, while large to the local government entity and the relator, would not approach the scale of a Medicaid or Medicare FCA claim.⁹³ If the claim seems at all reasonable, however, it is unlikely the government will ask the court to dismiss the relator's claim.⁹⁴ If the government intervenes, the relator's participation will vary depending upon the plaintiff's and plaintiff's counsel's willingness to take on greater expense to participate in the litigation as well as the value of the original information to the government's case. If the government does not intervene, the relator will take the case through pretrial to settlement or appeals.

2. *Damages.*—Damages under the FCA can tally up to a large number. Potentially, the value of the contract or grant multiplied by up to a factor of three is at stake for the defendant, not to mention a civil monetary penalty of up to around \$11,000 for each false claim or statement, as well as attorneys' fees.⁹⁵ The relator's share, apart from attorneys' fees, can go as high as 30% if the United States does not intervene, between 15% and 20% if the United States does intervene,⁹⁶ but no more than 10% if the relator's claim is largely based on a public disclosure.⁹⁷ The amount of the contract damages component can, depending upon circuit law, equal the full value of the claim or actual damages.⁹⁸ Most successful FCA claims end in settlement;

92. See generally CIVIL DIV., U.S. DEP'T OF JUSTICE, FRAUD STATISTICS - OVERVIEW (2010), available at [http://www.fcaalert.com/uploads/file/Stats\(1\).pdf](http://www.fcaalert.com/uploads/file/Stats(1).pdf) (reporting data permitting comparison of *qui tam* cases from 1987 through 2010 where the government intervened and did not intervene with respect to relator shares and dismissals).

93. See U.S. ATTORNEY'S OFFICE, E. DIST. PA., FALSE CLAIMS ACT CASES: GOVERNMENT INTERVENTION IN QUI TAM (WHISTLEBLOWER) SUITS 2, available at <http://www.justice.gov/usao/pac/Documents/fcprocess2.pdf> (reporting that fewer than 25% of *qui tam* actions lead to government intervention). Others have agreed with the assumption that the government is more likely to intervene in high-dollar cases. See, e.g., Ben Depoorter & Jef De Mot, *Whistle Blowing: An Economic Analysis of the False Claims Act*, 14 SUP. CT. ECON. REV. 135, 149 (creating a game-theoretic model of the behavior of FCA litigators).

94. See Mike Scarcella, *Taking the Whistle Out of Her Hand*, CORP. COUNSEL, Mar. 1, 2012, <http://www.law.com/jsp/cc/PubArticleCC.jsp?id=1202541469671> (explaining that it is "rare for the government to ask a judge to dismiss a suit" brought under the FCA).

95. 31 U.S.C. § 3729(a)(1) (2006); *id.* § 3730(d)(2); 28 C.F.R. § 85.3(9) (2011).

96. 31 U.S.C. § 3730(d)(1)–(2).

97. *Id.* § 3730(d)(1).

98. Compare, e.g., *Harrison v. Westinghouse Savannah River Co.*, 176 F.3d 776, 785 n.7 (4th Cir. 1999) (holding the United States can recover the full value paid on a false claim), with, e.g.,

therefore, the relator and Department of Justice have practical discretion to determine whether to push for a larger damages multiplier and civil monetary penalty.⁹⁹

The primary differences in the civil rights context are that the contract values subject to forfeiture will be smaller than, say, hospital Medicare contracts, and the civil monetary penalties fewer because there will be fewer individual claims for payment by the defendant to base the penalties upon. That said, the grants are still large.

C. *Legally False Certification of Compliance with Civil Rights Laws*

In a civil rights FCA claim, the relator's claim will usually rest on a claim that the defendant local government expressly or impliedly certified compliance with an antidiscrimination law upon which the grant is conditioned.¹⁰⁰ The defendant's certification can either be false at the time of the certification or disbursement of funds, or it can become false during the period after disbursement while the defendant continues to use the funds.¹⁰¹ This subpart will explore the most likely manner in which a local government might make implied certifications of compliance with antidiscrimination laws to receive funding and then render those certifications false by noncompliance.

1. Implied Certification.—A local government recipient of a federal grant or contract funds, like the community policing funds, can certify its compliance with antidiscrimination requirements in several ways. The most obvious form of certification occurs when the representative of the local government signs a contract guaranteeing its previous, current, or intended future compliance with contractual terms included expressly or by reference in the contract.¹⁰² Some circuits recognize that the certification can also arise from an express statutory requirement that funding under a grant or contract

Young-Montenay, Inc. v. United States, 15 F.3d 1040, 1043 (Fed. Cir. 1994) (holding the government is entitled to actual damages).

99. Cf. Thomas H. Stanton, *Fraud-and-Abuse Enforcement in Medicare: Finding Middle Ground*, HEALTH AFFAIRS, July–Aug. 2001, at 28, 36 (“In its reliance on settlement rather than adjudication, the False Claims Act resembles many other areas where prosecutorial discretion is perhaps more important than the letter of the law.”).

100. See *infra* section II(C)(1).

101. See *infra* section II(C)(1).

102. See, e.g., United States *ex rel.* Compton v. Midwest Specialties, Inc., 142 F.3d 296, 302 & n.4 (6th Cir. 1998) (holding that failure to comply with product testing requirements specified in express contract terms caused liability under the FCA); COPS GRANT MANUAL, *supra* note 65, at 63 (stating that “[b]y the applicant’s authorized representative’s signature, the applicant assures that it will comply with all legal and administrative requirements that govern the applicant for acceptance and use of federal grant funds”).

be conditioned upon compliance with that statute.¹⁰³ Yet another basis recognized in some circuits finding certification is when a reasonable connection can be implied between a statutory requirement and the contract.¹⁰⁴

While these categories of implied certification distinguish the source of the certification, another dimension is the relation in time of the certification to its falsification. In many cases, the grant recipient will not begin to fall afoul of the contractual terms until after certification.¹⁰⁵ For example, an enhanced community policing program cannot discriminate in the provision of community policing services that depend on federal funding until after the federal funding permitted the creation of those services. Courts have recognized that certification can be implied from an earlier express certification.¹⁰⁶ This can happen in two ways. First, every time the defendant pulls additional funds from the grant, even if those requests do not expressly state the full extent of the antidiscrimination requirements of the program, courts can imply those certifications into the disbursement claims.¹⁰⁷ Thus, each claim for disbursement impliedly includes a false certification. Alternatively, the defendant could receive only a one-time disbursement and then fall out of compliance after that disbursement while keeping the money. This would constitute retaining an overpayment if done knowingly and violate the FCA.¹⁰⁸

103. See, e.g., *Mikes v. Straus*, 274 F.3d 687, 700 (2d Cir. 2001) (limiting implied certification based upon Medicare statutes to cases where the language expressly conditions payment upon compliance).

104. See, e.g., *United States ex rel. Hutcheson v. Blackstone Med., Inc.*, 647 F.3d 377, 388 (1st Cir. 2011) (disagreeing with *Mikes* and declining to limit implied certification based on statutes to cases where payment is explicitly conditioned on compliance); *Ebeid ex rel. United States v. Lungwitz*, 616 F.3d 993, 998–99 (9th Cir. 2010) (constraining the *Mikes* holding to Medicare cases).

105. For example, when the University of Phoenix executed an agreement with the Department of Education (DOE) to receive Higher Education Act (HEA) funds, it agreed to not give incentive payments to its recruiters based upon financial aid enrollment figures while it participated in the program. *United States ex rel. Hendow v. Univ. of Phoenix*, 461 F.3d 1166, 1168–69 (9th Cir. 2006). The University then later submitted claims for payment that did not require express certification of compliance with the HEA requirements. *Id.* at 1169–70, 1176–77.

106. See, e.g., *Ebeid*, 616 F.3d at 998 (“Implied false certification occurs when an entity has previously undertaken to expressly comply with a law, rule, or regulation, and that obligation is implicated by submitting a claim for payment even though a certification of compliance is not required in the process of submitting the claim.”); *Hendow*, 461 F.3d at 1175–77 (“The execution of this Agreement [which references program requirements] by the Institution and the Secretary is a prerequisite to the Institution’s initial or continued participation in [the] program.”).

107. See *supra* note 105.

108. 31 U.S.C. § 3729(a)(1)(G) (2006 & Supp. IV 2011); *id.* § 3729(b)(3).

2. *Legally False*.—A certification is or becomes false when it is either factually or legally untrue.¹⁰⁹ A defendant makes a factually false claim or certification when it makes a statement that defies reality—i.e., it claims that it delivered a product when it did not.¹¹⁰ A legally false claim or certification happens when a contractor or grantee says it has met a legal standard but has not actually met the requirements of that law.¹¹¹ For example, the Anti-Kickback Statute (AKS) bans hospitals that receive Medicare payments from paying kickbacks to physicians who refer patients to the hospital.¹¹² Any claims for payment from Medicare, made after a hospital has paid a physician something that would be legally classified as a kickback, would constitute a legally false certification of compliance with AKS.¹¹³

The antidiscrimination relator will frame the argument around a theory of a legally false certification. The relator will set out to show that the defendant's behavior did not comply with the antidiscrimination contractual terms or laws, and the claim will operate within all of the familiar antidiscrimination statutory frameworks, burden shifting, and related legal standards. If, for example, the community policing effort runs afoul of Title VII by engaging in racial discrimination, then, applying the *McDonnell Douglas* burden-shifting standard, the relator will need to first show a prima facie case of individual discrimination,¹¹⁴ a pattern or practice of discrimination,¹¹⁵ or disparate impact.¹¹⁶ At this point, the burden shifts to the defendant, who can provide a nondiscriminatory reason for the apparent discrimination, after which the burden shifts once more onto the relator to prove this reason merely a pretext.¹¹⁷ To win an FCA claim based on

109. *Mikes v. Straus*, 274 F.3d 687, 696–97 (2d Cir. 2001); John T. Brennan, Jr. & Michael W. Paddock, *Limitations on the Use of the False Claims Act to Enforce Quality of Care Standards*, 2 J. HEALTH & LIFE SCI. L. 37, 48 (2008).

110. *United States v. Rivera*, 55 F.3d 703, 709 (1st Cir. 1995) (“The paradigmatic example of a false claim under the FCA is a false invoice or bill for goods or services.”).

111. *United States v. Sci. Applications Int’l Corp.*, 626 F.3d 1257, 1266 (D.C. Cir. 2010).

112. Anti-Kickback Statute, 42 U.S.C. § 1320a-7b(b), (g) (2006 & Supp. IV 2011) (forbidding the payment of kickbacks to physicians for referrals to provide services to patients that Medicare will pay for, and linking violations to the FCA).

113. *Id.*

114. *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802–07 (1973) (defining burden-shifting under Title VII); see also *Connick v. Thompson*, 131 S. Ct. 1350, 1360, 1362–64 (2011) (raising the burden of showing pattern and practice under Section 1983).

115. *Int’l Bhd. of Teamsters v. United States*, 431 U.S. 324, 336 (1977); see *Connick*, 131 S. Ct. at 1360, 1362–64 (raising the burden of showing pattern and practice under Section 1983).

116. 28 C.F.R. § 42.104(b)(2) (2012) (forbidding grant recipients from using “criteria or methods of administration which have the effect of subjecting individuals to discrimination because of their race, color, or national origin, or have the effect of defeating or substantially impairing accomplishment of the objectives of the program as respects individuals of a particular race, color, or national origin.”). But see *Alexander v. Sandoval*, 532 U.S. 275, 285–86, 293 (2001) (barring direct, private enforcement of Title VI disparate impact regulations).

117. *McDonnell Douglas*, 411 U.S. at 802–04.

noncompliance with an antidiscrimination statute, the relator will have to win on all aspects of the traditional antidiscrimination claim, except he will not have to prove individual injury or damages.

D. Corporate Knowledge of the False Claim or Certification

Under the FCA, a defendant must knowingly make a false statement.¹¹⁸ “[K]nowingly” means that the entity submitting the claim “(i) has actual knowledge of the information; (ii) acts in deliberate ignorance of the truth or falsity of the information; or (iii) acts in reckless disregard of the truth or falsity of the information.”¹¹⁹ The relator does not have to show “proof of specific intent to defraud.”¹²⁰ “[M]ere negligence or even gross negligence” does not create liability,¹²¹ nor does mismanagement not rising to the level of deliberate indifference or reckless disregard.¹²² However, for example, when a medical doctor gave full control of Medicare billing to an individual lacking experience in Medicare billing, a court found the doctor to have acted with reckless disregard with respect to filing his Medicare claims.¹²³ Similar to negligent mismanagement, negligent or mistaken understandings of “opaque” regulations resulting in “minor or technical” errors lacking evidence of “a sinister shadow” are not sufficient.¹²⁴

The circuits disagree about whether the sum of the different pieces of information known by different employees can meet the knowledge requirement in a corporate setting. The D.C. Circuit has said that the relator cannot prove knowledge by showing that “the ‘collective knowledge’” of the defendant’s employees met the knowledge standard.¹²⁵ Instead, knowledge requires showing that specific employees independently had sufficient knowledge.¹²⁶ The Fourth Circuit, however, has held that the relator does not need to prove that a single actor knew both of the requisite false certification and the activities making that certification false.¹²⁷ The Supreme Court has recognized the disagreement but has not reached a conclusion on the question

118. 31 U.S.C. § 3729(a) (2006 & Supp. IV 2011).

119. *Id.* § 3729(b)(1)(A).

120. *Id.* § 3729(b)(1)(B).

121. *United States ex rel. Longhi v. United States*, 575 F.3d 458, 468 (5th Cir. 2009); *see also United States ex rel. Aakhus v. Dyncorp, Inc.*, 136 F.3d 676, 682 (10th Cir. 1998) (requiring “an aggravated form of gross negligence, or ‘gross negligence-plus’”); *Hagood v. Sonoma Cnty. Water Agency*, 81 F.3d 1465, 1478 (9th Cir. 1996) (“The requisite intent is the knowing presentation of what is known to be false, as opposed to innocent mistake or mere negligence.” (internal quotation marks omitted)).

122. *United States ex rel. Farmer v. City of Hous.*, 523 F.3d 333, 339 (5th Cir. 2008).

123. *United States v. Stevens*, 605 F. Supp. 2d 863, 868–69 (W.D. Ky. 2008).

124. *Farmer*, 523 F.3d at 339–41.

125. *United States v. Sci. Applications Int’l Corp.*, 626 F.3d 1257, 1275 (D.C. Cir. 2010).

126. *Id.* at 1275–76.

127. *United States ex rel. Harrison v. Westinghouse Savannah River Co.*, 352 F.3d 908, 918–19 (4th Cir. 2003).

of collective knowledge.¹²⁸ A related but limited dispute looks at the question of whether liability for the false claims of an employee passes vicariously onto the employer. Liability does accrue, in general, so long as the actor had apparent authority.¹²⁹

If collective knowledge is sufficient, then the civil rights relator's job is relatively easy: he needs only to prove that individual employees knew or were deliberately indifferent toward finding out whether discrimination occurred, and that at the same time other employees certified that no discrimination occurred and were at least deliberately indifferent toward finding out whether this was true. In circuits where collective knowledge cannot show corporate knowledge, a relator will need to try to show knowledge or deliberate ignorance by a supervisor with responsibility for nondiscrimination compliance.

E. Proving the False Claim Was Material to Payment

If the government might have paid the claim or permitted the grant recipient to keep the claim had it known of the false certification, the false certification might not be material. A false certification is "material" to a claim when it has "a natural tendency to influence, or be capable of influencing, the payment or receipt of money or property."¹³⁰ In several circuits, materiality has a broad meaning that includes any activities that "'could have' . . . or had the 'potential' to influence the government's decision" to pay, but does not require "that the false statements actually did so."¹³¹ The Eighth Circuit adopted a narrower, outcome-focused test and requires that the relator show that the statement had "the purpose and effect of causing the United States to pay out money it is not obligated to pay."¹³² Three circuits have a strict requirement requiring the relator to show that, but for the false statement, the government would not have paid.¹³³

128. *Staub v. Proctor Hosp.*, 131 S. Ct. 1186, 1191–92 (2011).

129. *United States v. O'Connell*, 890 F.2d 563, 569 (1st Cir. 1989). *But see* *United States v. Ridgley State Bank*, 357 F.2d 495, 500 (5th Cir. 1966) (requiring that the actor make the false claim with the purpose of benefiting his employer for liability to accrue).

130. 31 U.S.C. § 3729(b)(4) (Supp. IV 2011); *see also* *United States ex rel. Hutcheson v. Blackstone Med., Inc.*, 647 F.3d 377, 385–86 (1st Cir. 2011) (discussing the development of the court-created doctrine of factual versus legal falsity under the FCA and opting not to use the formal categories).

131. *United States ex rel. Longhi v. United States*, 575 F.3d 458, 469 (5th Cir. 2009). The Fifth Circuit also indicated that it felt that the amendments to the FCA in the Fraud Enforcement and Recovery Act of 2009 § 4, 31 U.S.C. § 3729(b)(4) (Supp. IV 2011), reflected in the language quoted above, reaffirm the approach taken by the Fourth, Sixth, and Ninth Circuits. *Longhi*, 573 F.3d at 470.

132. *Costner v. URS Consultants, Inc.*, 153 F.3d 667, 677 (8th Cir. 1998).

133. *See id.*; *see also* *United States v. First Nat'l Bank of Cicero*, 957 F.2d 1362, 1373–74 (7th Cir. 1992) (holding that the false statement need not have been the event causing loss, so long as the government would not have paid but for reliance on the false statement); *United States v. Hibbs*,

At least one court has held that when certifications are conditions for participation in a program but not payment, they are not material to payment.¹³⁴ Another court distinguished that case and held that when a court can equate participation in a program with payment, then false certification of conditions of participation creates liability under the FCA.¹³⁵ Language describing the relevant requirements as “condition[ing] . . . [payment] upon compliance,” permitting participation “only if” conditions are met, or describing the conditions as “prerequisite[s]” to participation all can indicate that a false statement about those conditions creates a false claim.¹³⁶

The element of materiality has the potential to create significant problems for an antidiscrimination FCA relator. A defendant could argue that, based on a federal agency’s previous behavior, even if the agency had known of the defendant’s breach of its antidiscrimination contract terms, it would not have ended the contract or grant. The Department of Justice, the defendant could argue, would not terminate a COPS grant merely because of one accusation of a systematic disparate impact in a program receiving COPS funding.¹³⁷ In the stricter circuits,¹³⁸ demonstration of agency indifference could ring the death knell for an antidiscrimination FCA claim. Many other circuits, however, conclude that the mere fact that the agency reserved the right to act in response to discrimination would be sufficient.¹³⁹

The agency’s statutes and regulations also often create a right to voluntarily remedy violations.¹⁴⁰ Instead of directly linking noncompliance to breach of contract, the agency could provide a path for dispute resolution and reconciliation between the injured and the grant recipient.¹⁴¹ The

568 F.2d 347, 350–51 (3d Cir. 1977) (taking an even narrower stance than the Eighth Circuit by requiring the payment be made “by reason of” a false claim, in that the false claim must have been the actual source of the government’s loss).

134. See *Mikes v. Straus*, 274 F.3d 687, 701–02 (2d Cir. 2001) (drawing a distinction between compliance as a condition of participation and compliance as a condition of reimbursement).

135. *United States ex rel. Hendow v. Univ. of Phoenix*, 461 F.3d 1166, 1176–77 (9th Cir. 2006) (“The University argues that the ban is merely a condition of *participation*, not a condition of *payment*. But in this case, that is a distinction without a difference. . . . [I]f we held that conditions of participation were not conditions of payment, there would be no conditions of payment at all . . .”).

136. *Id.* at 1176.

137. *Cf.* COPS GRANT MANUAL, *supra* note 65, at 19 (“Remedies for noncompliance *may* include, but are not limited to: suspending grant funding, repaying misused grant funds, voluntary withdrawal from or involuntary termination of remaining grant funds, and bars from receiving future COPS grants.” (emphasis added)).

138. See *supra* notes 132–34 and accompanying text.

139. See *supra* notes 131, 135–36 and accompanying text.

140. *E.g.*, 42 U.S.C. § 2000d-1(2) (2006) (providing that “no such action shall be taken until the department or agency concerned has advised the appropriate person or persons of the failure to comply with the requirement and has determined that compliance cannot be secured by voluntary means”).

141. *Id.*

regulation would not permit the agency merely to end the grant or contract.¹⁴² In these situations, a defendant would argue that a breach by violation of antidiscrimination laws could not directly lead the agency to end the contract or decline to pay a claim. This assessment, however, would depend upon an incomplete reading of the agency's remedies for these sorts of breaches. In most cases, if an entity fails to comply with the agency's attempts to remedy the breach in cooperation with the entity, the agency can then terminate the contract or grant.¹⁴³ To simplify, a first breach effectively puts the defendant on probation, and a second pushes the defendant into termination. The defendant may argue that the relator's claim reflects only the first breach. A relator would respond that a false certification covering up the first breach prevented a second breach from leading to termination. As a result, the first false certification materially affected the agency's decision to honor the contract in the case of a second breach.

F. Answering the Public Disclosure Bar Under Modern FCA Law

The public disclosure bar presents the most significant hurdle for relators, but recent amendments have lowered the bar relators must jump to maintain their claims against this defense. The public disclosure bar purports to block *qui tam* claims based off of information in the public record unless the relator was the original source of the information.¹⁴⁴ Under earlier versions of the FCA, many courts had interpreted the sweep of this restriction broadly, defining public information to include partial disclosures as well as disclosures in state and local courts.¹⁴⁵ Recent amendments limit the reach of this restriction to records from hearings in which the federal government participated as a party, federal reports, and the news.¹⁴⁶ The amendments also make clear that to assert this defense, defendants must show that the public disclosures form substantially the same allegations that the relator has claimed rather than showing that only a single part overlaps.¹⁴⁷ This means that plaintiffs could use information found through discovery for a traditional

142. See, e.g., 34 C.F.R. § 104.6 (2011) (omitting the option for the United States to cancel the grant or contract based on disability-based discrimination); *id.* § 106.3 (requiring remedial and affirmative action but not the cancellation of a grant or contract for sex-based discrimination).

143. 42 U.S.C. § 2000d-1(2).

144. 31 U.S.C. § 3730(e)(4) (2006 & Supp. IV 2011).

145. See, e.g., *Fed. Recovery Servs., Inc. v. United States*, 72 F.3d 447, 450–51 (5th Cir. 1995) (holding that filings in state court constitute a public disclosure and that allegations even partly based upon public disclosures are barred).

146. 31 U.S.C. § 3730(e)(4) (Supp. IV 2011).

147. *Id.*; see also Stewart, *supra* note 90, at 179–80 (recognizing that the amended law limits the public disclosure bar to information based on federal reports or hearings in which the federal government participated and which more than slightly overlap with the relator's claims).

civil rights claim to litigate an FCA claim even without a whistle-blower,¹⁴⁸ or perhaps could use data from local and state governments to show disparate impact.

Congress has also broadened the original source exception. In the case of traditional whistle-blowers,¹⁴⁹ if the relator acted as the original source of the information behind the public disclosures, then the relator escapes the public disclosure bar.¹⁵⁰ Under the new law, if the relator contributes materially to the information publicly disclosed and makes it available to the government prior to filing a *qui tam*, then the relator bypasses the bar.¹⁵¹ This means that if a relator tries to resolve a dispute through, for example, the Equal Employment Opportunity Commission disclosure process, which would inform the federal government of a claim, the relator can overcome the bar.

Despite the relatively relator-friendly law under the amended FCA, civil rights relators still face a difficult challenge. Many antidiscrimination schemes require significant reporting by grant recipients, including through equal employment opportunity plans.¹⁵² They further require that a grant recipient give notice to the funding agency when a court renders a judgment for discrimination against the defendant.¹⁵³ These measures, intended to encourage the disclosure of illicit discrimination to the government, present significant obstructions to FCA claims because they create public disclosure defenses even under amended law. If the recipient entity complies with all of its reporting requirements, then those reports could preclude any FCA claim based upon them as they are federal government documents.¹⁵⁴

Fortunately, relators can maintain their claims despite these in-depth reporting requirements. First, relators should bear in mind that even when a defendant reports its behavior to the federal government, anything in state and local records remains fair game. Second, and very importantly, the relator could expose false reporting to the federal government as part of these antidiscrimination remediation efforts. Just because there are reporting requirements does not mean that reporting has properly occurred. Third, if

148. Cf. Stewart, *supra* note 90, at 169–70, 172, 179 (claiming the amendments make it more plausible that disinterested relators who discover information through discovery might survive the public disclosure bar).

149. *Id.* at 169.

150. 31 U.S.C. § 3730(e)(4) (2006).

151. *Id.* § 3730(e)(4) (Supp. IV 2011).

152. See, e.g., COPS GRANT MANUAL, *supra* note 65, at 19–20 (requiring the creation of, and compliance with, Equal Employment Opportunity Plans by COPS grantees of certain types and sizes receiving large grants).

153. See, e.g., *id.* at 63 (“In the event that any court or administrative agency makes a finding of discrimination on grounds of race, color, religion, national origin, gender, disability or age against the applicant after a due process hearing, it agrees to forward a copy of the finding . . .”).

154. See *supra* note 147 and accompanying text.

the civil rights advocate has a traditional, individually injured plaintiff, and can ethically and economically afford to bring the traditional claim together with the FCA claim, then the relator will serve as the original source of the public disclosure and will avoid this defense.

Finally, under the recent amendments to the FCA, the relator can simply add to the information publicly available and maintain a claim as long as the additional information is not “substantially the same” as the publicly disclosed allegations.¹⁵⁵ Therefore, the partial reporting of an incident of discrimination by the entity or the reporting of one incident but not a second similar incident will not necessarily block the relator’s FCA claim based on the second incident.

G. *At Least One Civil Rights FCA Claim Has Been Successfully Litigated*

A broad search reveals only a few incidents of civil rights FCA claims,¹⁵⁶ but those cases often failed due to insufficient pleading¹⁵⁷ and other preliminary statutory bars¹⁵⁸ before the court could reach the merits of FCA liability.

Some relators, however, have gotten past these initial bars. In one Eighth Circuit FCA suit, a relator accused Arkansas of misrepresenting

155. 31 U.S.C. § 3730(e)(4)(A) (2006); *see also* Stewart, *supra* note 90, at 179 (observing that the 2010 amendments to the FCA “ensur[e] that only cases where ‘substantially the same’ allegations were publicly [disclosed] will be barred,” which will help to “assur[e] relators that ‘where only one element of the fraudulent transaction is in the public domain (e.g., X), [they] may mount a case by coming forward with either the additional elements necessary to state a case of fraud (e.g., Y) or allegations of fraud itself (e.g., Z)’” (quoting *United States ex rel. Springfield Terminal Ry. Co. v. Quinn*, 14 F.3d 645, 655 (D.C. Cir. 1994))).

156. The author searched Westlaw’s ALLFEDS database of all federal court cases using the following broad query: (“false claims act” “31 u.s.c. s 3729” “31 u.s.c. s 3730”) & (“title ix” “title vii” “title vi” “rehabilitation act” “americans with disabilities act” “individuals with disabilities education act” “42 u.s.c. s 12112” “20 u.s.c. s 1681” “42 u.s.c. s 2000e” “42 u.s.c. s 2000d” “29 u.s.c. s 704” “20 u.s.c. s 1400”). This query resulted in 479 federal court opinions as of December 2011, of which the author determined that only nine pertained to the topic of this Note.

157. *See, e.g., United States ex rel. Bly-Magee v. Premo*, 333 Fed. App’x 169, 170 (9th Cir. 2009) (unpublished opinion) (holding that alleging false certification of compliance with the Rehabilitation Act without alleging which specific provision was violated failed under Federal Rule of Civil Procedure 9(b)); *Raghavendra v. Trs. of Columbia Univ.*, No. 06 Civ 6841(PAC), 2008 WL 2696226, at *10 (S.D.N.Y. July 7, 2008) (dismissing FCA claim based upon general allegations that Columbia mistreated minorities without any specific factual details).

158. *See, e.g., Stoner v. Santa Clara Cnty. Office of Educ.*, 502 F.3d 1116, 1123, 1127 (9th Cir. 2007) (dismissing a false certification claim arising as a result of noncompliance with Individuals with Disabilities Education Act (IDEA) because a state could not be sued as a person under the FCA and a relator cannot sue pro se); *United States ex rel. Westerfield v. Univ. of S.F.*, No. C 04-03440 JSW, 2006 WL 335316, at *5 (N.D. Cal. Feb. 14, 2006) (dismissing a claim for failing to allege that the relator was the original source of information disclosed in a state law claim), *abrogated by* 31 U.S.C. § 3730(e)(4) (Supp. IV 2011) (considering only federal court claims, federal documents, and news reports as public disclosures).

compliance with the Rehabilitation Act in claims for education funding.¹⁵⁹ After surviving appeal on a jurisdictional issue and returning to district court, the main arguments raised by the defendants at summary judgment were those addressed earlier in this Note, including whether the claim for funds included certification, whether those certifications were material to the government's decision to pay, and whether the defendants made those false certifications knowingly.¹⁶⁰ Unfortunately, the case ended without a written opinion,¹⁶¹ and we cannot tell how the court would have decided it on the FCA merits.

Plaintiffs in a New York district court also successfully settled an FCA case addressing compliance with Federal Fair Housing Act (FFHA) grants.¹⁶² A local municipality accepted block grants from the Department of Housing and Urban Development (HUD), which required recipients to take race into account when developing housing programs.¹⁶³ A civil rights organization alleged that the municipality ignored race in its programs but certified compliance with those requirements.¹⁶⁴ The court held that the false certifications were material as a matter of law even though HUD knew about the noncompliance, did not act on that information, and did not have any legal obligation to cease funding.¹⁶⁵ The court also said that because of the explicit relationship between the grants and the conditions, the implied certifications through grant draw downs (requests for payment) on grants survived even the narrower Second Circuit implied certification rule.¹⁶⁶ However, the court left to the jury the question of whether HUD's failure to act resulted in a lack of knowledge by the municipality of its noncompliance,¹⁶⁷ an issue which did not reach the jury because the parties settled for \$52 million.¹⁶⁸

159. *United States ex rel. Rodgers v. Arkansas*, 154 F.3d 865, 866–67 (8th Cir. 1998) (affirming a district court holding that the state can be sued under the FCA), *cert. dismissed*, 527 U.S. 1018 (1999), *overruled by* *Vt. Agency of Natural Res. v. United States ex rel. Stevens*, 529 U.S. 765, 787 (2000) (holding that states are not “person[s]” under the FCA).

160. Memorandum Supporting Motion for Summary Judgment at 5, *United States ex rel. Rodgers v. Arkansas*, No. LR-C-96-195, 1999 WL 33997233 (E.D. Ark. Jan. 5, 1999).

161. The docket list shows that the court granted summary judgment for party school districts, though without a written opinion, and that the state and the relators stipulated a dismissal. Docket Sheet, *Rodgers*, No. LR-C-96-195 (located through LexisAdvance docket search).

162. *Menz*, *supra* note 21, at 1147–48 & n.69.

163. *United States ex rel. Anti-Discrimination Ctr. of Metro N.Y., Inc. v. Westchester Cnty.*, 495 F. Supp. 2d 375, 377 (S.D.N.Y. 2007).

164. *Id.* at 376, 387.

165. *United States ex rel. Anti-Discrimination Ctr. of Metro N.Y., Inc. v. Westchester Cnty.*, 668 F. Supp. 2d 548, 569–70 (S.D.N.Y. 2009).

166. *Id.* at 566–67.

167. *Id.* at 567.

168. *Menz*, *supra* note 21, at 1148 & n.69.

Another positive case for coping with both public disclosure and stating a claim based upon earlier express certification recently received a pass in Florida, where a district court refused to dismiss an FCA claim alleging that Kaplan University failed to comply with the Rehabilitation Act and relevant DOE regulations despite its certifications to that effect.¹⁶⁹ In order to receive funds, the University had previously agreed to comply with those regulations.¹⁷⁰ The DOE's Office of Civil Rights (OCR) later found that Kaplan had been out of compliance because it had failed to address numerous requirements for disability nondiscrimination.¹⁷¹ The court held that the relator had sufficiently pled an FCA claim for the period where OCR found Kaplan out of compliance.¹⁷² As the original source of information to the government for the OCR investigation,¹⁷³ the relator will likely survive any future effort to use the public disclosure bar, and this case is a positive sign that civil rights FCA claims will survive the significant reporting requirements under traditional civil rights laws and contracts.

Courts have also dismissed civil rights FCA claims based upon noncompliance. When a relator alleged violations of the Individuals with Disabilities Education Act (IDEA),¹⁷⁴ the Ninth Circuit dismissed the claim, choosing not to imply certification of the IDEA laws into the claims for payment because the court felt the certification was not material to the payment.¹⁷⁵ Recently, the Ninth Circuit made this opinion's limited view on implied certification less clear, and now the Circuit might view an implied certification claim more favorably.¹⁷⁶

III. Agencies Should Add Grant-Appropriate Constitutional Standards to Contracts

The President and the executive branch agencies arguably have the power to impose contractual terms upon grant recipients, demanding they certify compliance with relevant constitutional requirements.¹⁷⁷ While

169. *United States ex rel. Diaz v. Kaplan Univ.*, No. 09-20756-CIV, 2011 WL 3627285, at *1-2, *6 (S.D. Fla. Aug. 17, 2011).

170. *Id.* at *1-2.

171. *Id.* at *2 & n.7.

172. *Id.* at *6.

173. *See id.* at *2 (discussing OCR's investigation of Kaplan's failure to accommodate the co-relator's bipolar disorder).

174. 20 U.S.C. §§ 1400-82 (2006).

175. *United States ex rel. Hopper v. Anton*, 91 F.3d 1261, 1266-67 (9th Cir. 1996); *see also United States ex rel. Westerfield v. Univ. of S.F.*, No. C 04-03440 JSW, 2006 WL 2884331, at *3-4 (N.D. Cal. Oct. 10, 2006) (following *Hopper*).

176. *See Ebeid ex rel. United States v. Lungwitz*, 616 F.3d 993, 998 (9th Cir. 2010) (recognizing the theory of false certification from a prior express certification of compliance with a law).

177. One note has argued that the First Amendment of the Constitution, like the Anti-Kickback Statute, can be implied as a condition to receive federal funding. *Levine, supra* note 21, at 156-57,

Congress could act more freely to set such requirements,¹⁷⁸ Congress is effectively and unfortunately a broken branch of government.¹⁷⁹ Therefore the executive branch would have to act alone. For example, the COPS program could require police departments' compliance with the Supreme Court and relevant lower courts' standards of Fourth and Fifth Amendment law. This would allow the FCA to cover at least some of the constitutional rights territory Section 1983 would address if not for the Supreme Court's limiting doctrines.

Presidential power to enact such rules through executive orders must derive from a constitutional power or congressional grant of authority.¹⁸⁰ There is no obvious example of the Congress expressly granting the President the power to regulate government procurement for the protection of constitutional rights.¹⁸¹ Nor does the general procurement power to protect efficiency and economy seem sufficient because constitutional rights are unlikely to be part of a "nexus" related to those goals, though that assumption is contentious.¹⁸² Alternatively, the argument would either have to be that objectives stated in specific spending statutes implicitly could be better served by setting constitutional goals or that the President can act within Justice Jackson's "zone of twilight" where Congress has not spoken.¹⁸³ The current administration has already considered doing something similar by banning discrimination based on sexual orientation, though in the end President Obama declined to sign an executive order to that effect.¹⁸⁴

167–68. The argument is that the government cannot do anything unconstitutional, and funding the probable establishment or the suppression of free exercise of religion would amount to an unconstitutional act. *Id.* at 168. If this is the case, then the President needs to do nothing. However, no court has validated implied certification from the Constitution, and the stricter courts discussed in subpart II(E) of this Note would likely hold that the Constitution does not expressly condition, for example, COPS payment upon compliance with the Fifth Amendment.

178. See *South Dakota v. Dole*, 483 U.S. 203, 207 (1987) (defining the sweeping congressional spending power).

179. Ben Pershing, *In 2011, Fewer Bills, Fewer Laws and Plenty of Blame*, WASH. POST, Dec. 5, 2011, http://www.washingtonpost.com/politics/in-2011-fewer-bills-fewer-laws-and-plenty-of-blame/2011/12/05/gIQA566iXO_story.html.

180. *Youngstown Sheet & Tube Co. v. Sawyer*, 343 U.S. 579, 585 (1952).

181. *Cf.* Simmons, *supra* note 28, at 383 (proposing that Congress condition COPS funding on states establishing programs to promote police accountability).

182. See VANESSA K. BURROWS & KATE M. MANUEL, CONG. RESEARCH SERV., R41866, PRESIDENTIAL AUTHORITY TO IMPOSE REQUIREMENTS ON FEDERAL CONTRACTORS 2–3, 22–24 (2011), available at <http://www.fas.org/sgp/crs/misc/R41866.pdf> (discussing the Presidential procurement power).

183. *Youngstown*, 343 U.S. at 637 (Jackson, J., concurring).

184. Jackie Calmes, *Obama Won't Order Ban on Gay Bias by Employers*, N.Y. TIMES, Apr. 11, 2012, <http://www.nytimes.com/2012/04/12/us/politics/obama-wont-order-ban-on-gay-bias-by-employers.html>.

In the case of COPS, the President could argue that policing in America means policing constitutionally, and therefore that fulfilling the goal of community policing requires constitutional community policing. A failure to comply with those requirements could lead to popular backlash against police departments, limiting their ability to effectively police communities as well as to a practical failure to police within the appropriate definition of policing. Therefore, the President can and should require compliance with those requirements in order to ensure the success of the program.

IV. Weighing the Benefits and Hazards of Using the FCA in Civil Rights Claims

FCA liability based on violations of antidiscrimination laws gives another tool to litigators for both individually injured plaintiffs as well as groups interested in institutional-change litigation. FCA litigation increases the potential liability for discriminatory behavior. And, perhaps most usefully, courts that have addressed similar types of litigation in the Fair Housing Act and Rehabilitation Act contexts have not rejected FCA liability for these types of laws.¹⁸⁵ The increased liability, however, comes with hazards, such as pushing against Congress's intended municipal liability and creating perverse incentives against reconciliation and disclosure.¹⁸⁶

A. *Increased Civil Rights Liability Under the FCA Runs Contrary to Congress's and the Court's Protection of Local Government Coffers*

The policy underlying the Court's and Congress's decision to prevent punitive damages against localities under the civil rights laws rests primarily on a conception that the public coffers should not be drained because of the poor decisions of agents of the government.¹⁸⁷ The Court has decided that the FCA's multiplied contract damages and civil monetary penalties are not purely punitive damages and are permitted against local governments.¹⁸⁸ That said, the use of the FCA to exact larger damages from municipalities runs afoul of the underlying policy rationale of the ban on those punitive damages even if the Court, for various reasons, has decided to treat FCA damages differently than punitive damages. The public coffers of local governments would shrink because of discrimination claims framed under the FCA, and this could trigger a response by the Court or Congress to clamp down on what might be framed as abuse of the FCA. These policies could also increase the incentives not to take government funding and could pull

185. See *supra* notes 162–73 and accompanying text.

186. One article also argues that the FCA offers the benefit of not requiring courts to enforce complex consent decrees by limiting remedies to damages. Hayes, *supra* note 21, at 54–56.

187. *City of Newport v. Fact Concerts, Inc.*, 453 U.S. 247, 261–64 (1981).

188. *Cook Cnty., Ill. v. United States ex rel. Chandler*, 538 U.S. 119, 129–34 (2003).

funding away from local governments that, despite their discrimination, typically provide important services. Both of these concerns caution against excessive use of the FCA to attack one-time civil rights violations when the discriminatory acts do not represent a broader and more significant policy or practice.

The law as it stands does not block these abuses except in as much as in certain circuits a defendant could argue that the certification of nondiscrimination did not materially affect payment by the government.¹⁸⁹ Many courts of appeals, however, would quite possibly permit an FCA claim based off of a single, innocuous violation without any hard legal bar.¹⁹⁰ Therefore, civil rights litigators, if wanting to keep the FCA available and not risk a pushback from the courts, should exercise care in which cases they choose to use the FCA. Arguably, litigators should reserve the FCA for cases where there is an individually aggrieved plaintiff for whom the traditional laws cannot provide a remedy or where there is a policy or practice needing institutional change.

B. FCA Liability Could Discourage Reconciliation Between Defendants and Plaintiffs

The bigger the pot of gold the plaintiff sees, the more likely the plaintiff will push to get a money award close in size to that hypothetical amount. FCA liability potentially adds a second pot. Savvy plaintiffs aware of the potential to use the FCA will have an incentive to choose not to usefully participate in efforts to reconcile their differences with the defendant, even when the defendant acts in good faith and the parties could otherwise work out their disagreement. Less savvy, naive plaintiffs will face less of an incentive to not negotiate at the pre-litigation stage because they will likely know less about the FCA than traditional civil rights laws. Once they enter litigation and their attorney informs them of the FCA possibility, their desired settlement value will likely increase. These naive plaintiffs may not recognize the likely higher probabilities of achieving a jointly more desirable outcome through reconciliation as compared to complex, slow, arguably highly unpredictable FCA litigation. As a result, the amount of litigation could increase at the expense of simpler, more effective efforts by both sides to reach a mutually positive outcome. This risks decreasing the overall utility of the litigation from a societal standpoint and possibly from an individual standpoint as well in the many cases where the relator loses.

The very availability of FCA-based remedies could disincentivize the admissions of wrongdoing typically necessary for reconciliation to occur because the potential defendant will recognize that every admission could

189. See *supra* notes 132–36 and accompanying text.

190. See *supra* notes 131 and accompanying text.

serve as the basis of a future FCA claim.¹⁹¹ Even if the injured party or institutional change litigator agrees not to sue under the FCA using this evidence, another litigant could. Admissions for reconciliation would have to occur in private if at all, which could lessen the impact of reconciliation in the context of institutional change.

Civil rights litigators working with individually injured plaintiffs have a duty to inform their prospective relators that the process of using the FCA is long and complex, and that while it offers significant procedural and damages advantages over traditional civil rights litigation in some cases, it is not a perfect solution for all claims. Furthermore, just as under traditional civil rights claims, both institutional change and traditional advocates should recognize that in many cases the defendant has not acted in bad faith and wants to reach a compromise. If, however, a defendant has not acted in good faith towards reaching a compromise, and if the problem is more systemic than a single injured plaintiff, the FCA would be a powerful lever.

C. FCA Liability Has an Ambiguous but Probably Positive Effect on the Truthful Disclosure of Discrimination to the Federal Government

The impact of expanded FCA liability pushes against discriminators in two directions with respect to encouraging proper reporting to the government. First, it would encourage extensive federal reporting of discrimination in order to avoid the hazard of an FCA claim by a civil rights advocate.¹⁹² Of course, this reporting would mean little without someone monitoring, but that someone need not be the agency if the information is made publicly available online or through Freedom of Information Act requests. This would permit traditional, individualized civil rights advocates access to greater information about unconstitutional behaviors by local government entities. Cutting in the other direction, if a local government knowingly fails to report honestly and fails to honestly certify compliance with training and avoidance of violations, then it opens the door to FCA claims, with far greater risks for lying than for coming clean.¹⁹³ Local governments might consider such a scheme a “damned if you do and damned if you don’t” result, but when constitutional rights are at stake, perhaps such a result is not so unfair.

Despite these pressures to encourage quality disclosure of discriminatory acts, the relatively lower risk of an outsider discovering subtle acts of discrimination, or the possibility that the discrimination might in fact

191. See Stewart, *supra* note 90, at 177–80 (explaining both that the 2010 amendments to § 3730 allow the relator to include public admissions in an FCA claim as long as the relator contributes his or her own information and that state and local records containing admissions may be used as the basis for an FCA claim).

192. See *supra* notes 152–54 and accompanying text.

193. See *supra* notes 95–97, 119–23 and accompanying text.

be unknowing, could press local governments not to disclose. An entity that, in light of the legal and factual burdens involved, calculates the probability of suit under the FCA to be sufficiently low to wash out the higher possible damages might opt not to report. And if the entity believes that the public disclosure bar will not protect it from independent sources of information despite its disclosures, it might disclose even fewer instances of discrimination in order to reduce the probability of tipping off a potential relator.

Altogether, the former pressures seem more probable than the latter because the latter antidisclosure effect requires assuming the public disclosure bar will not be effective if the entity discloses its wrongdoing to the federal government. Considering that the discriminator—even a discriminator only through deliberate indifference or recklessness—has more access to its own records and information than a relator will, the discriminator can decide to relieve itself of its burdens under the FCA by painting complete pictures of its discrimination to the federal government. Only a fraction of those instances will result in government action or in providing additional, unique information to traditional individually injured plaintiffs, and it would effectively bar FCA claims. The use of FCA liability is likely to increase disclosure of discrimination to the government.

V. Conclusion

The False Claims Act offers institutional change and traditional civil rights lawyers a tool to exact significant damages from federal contractors and grantees that violate current nondiscrimination conditions imposed by the federal government. While the requirements of knowledge, materiality, and the public disclosure bar all present potential obstacles to a civil rights litigant, they in many cases can be overcome as matters of law. Courts in a few limited cases have demonstrated that there are judges who are willing to listen to these types of claims and will not permit pure matters of law to stand in the way. In order to succeed, the relator needs to bring facts to the table as a whistle-blower, an individually injured plaintiff inside of the organization, or an informed outsider. Civil rights litigators should exercise judgment in order to avoid pushback. The litigator should also recognize the need to encourage reconciliation between the aggrieved and the defendant when possible. While there are hazards to increased liability, agencies should seriously consider using their power to set the terms of contracts to enforce further constitutional requirements on grant recipients. Responsible *qui tam* litigation by civil rights advocates and agency changes have the potential to make the FCA an effective remedy for victims of civil rights abuses.

—Ralph C. Mayrell



JAMAIL CENTER FOR LEGAL RESEARCH
TARLTON LAW LIBRARY
THE UNIVERSITY OF TEXAS SCHOOL OF LAW

The Tarlton Law Library Oral History Series features interviews with outstanding alumni and faculty of The University of Texas School of Law.

Oral History Series

- | | |
|---|--|
| No. 1 - <i>Joseph D. Jamail, Jr.</i> 2005. \$20 | No. 6 - <i>James DeAnda</i> 2006. \$20 |
| No. 2 - <i>Harry M. Reasoner</i> 2005. \$20 | No. 7 - <i>Russell J. Weintraub</i> 2007. \$20 |
| No. 3 - <i>Robert O. Dawson</i> 2006. \$20 | No. 8 - <i>Oscar H. Mauzy</i> 2007. \$20 |
| No. 4 - <i>J. Leon Lebowitz</i> 2006. \$20 | No. 9 - <i>Roy M. Mersky</i> 2008. \$25 |
| No. 5 - <i>Hans W. Baade</i> 2006. \$20 | |

Forthcoming:

Gloria Bradford, Patrick Hazel, James W. McCartney,
Michael Sharlot, Ernest E. Smith, John F. Sutton, Jr.

*Other Oral Histories Published by the
Jamail Center for Legal Research*

- Robert W. Calvert* (Texas Supreme Court Trilogy, Vol. 1). 1998. \$20
Joe R. Greenhill, Sr. (Texas Supreme Court Trilogy, Vol. 2). 1998. \$20
Gus M. Hodges (Tarlton Law Library Legal History Series, No. 3). 2002. \$20
Corwin Johnson (Tarlton Law Library Legal History Series, No. 4). 2003. \$20
W. Page Keeton (Tarlton Legal Bibliography Series, No. 36). 1992. \$25
Jack Pope (Texas Supreme Court Trilogy, Vol. 3). 1998. \$20

Order online at <http://tarlton.law.utexas.edu/> click on Publications
or contact Publications Coordinator,
Tarlton Law Library, UT School of Law,
727 E. Dean Keeton St., Austin, TX 78705

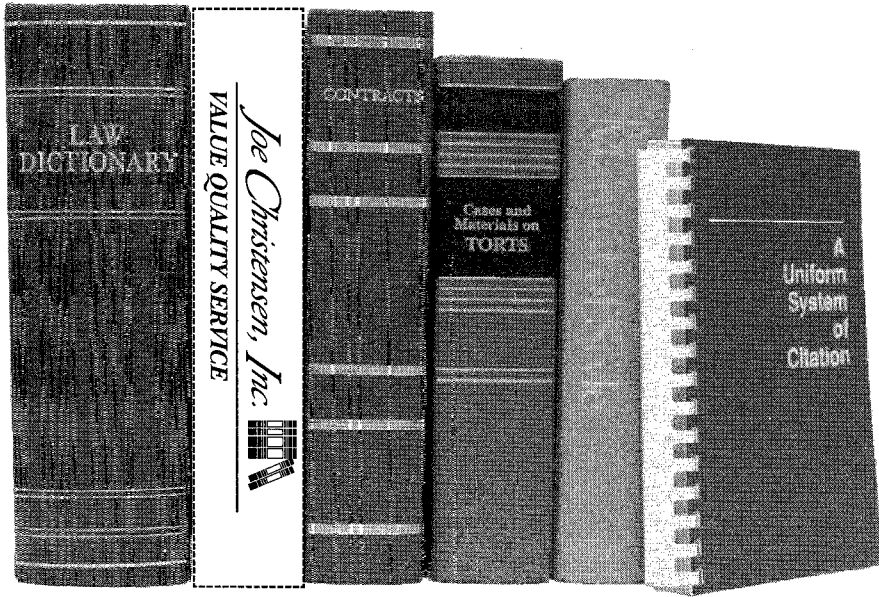
phone (512) 471-6228; fax (512) 471-0243;
email tarltonbooks@law.utexas.edu

THE UNIVERSITY OF TEXAS SCHOOL OF LAW PUBLICATIONS
 What the students print here changes the world

Journal	domestic/foreign
Texas Law Review http://www.TexasLRev.com	\$47.00 / \$55.00
Texas International Law Journal http://www.tilj.org	\$45.00 / \$50.00
Texas Environmental Law Journal http://www.texenrls.org/publications_journal.cfm	\$40.00 / \$50.00
American Journal of Criminal Law http://www.ajcl.org	\$30.00 / \$35.00
The Review of Litigation http://www.thereviewoflitigation.org	\$30.00 / \$35.00
Texas Journal of Women and the Law http://www.tjwl.org	\$40.00 / \$45.00
Texas Intellectual Property Law Journal http://www.tiplj.org	\$25.00 / \$30.00
Texas Hispanic Journal of Law & Policy http://www.thjlp.org	\$30.00 / \$40.00
Texas Journal On Civil Liberties & Civil Rights http://www.txjclcr.org	\$40.00 / \$50.00
Texas Review of Law & Politics http://www.trolp.org	\$30.00 / \$35.00
Texas Review of Entertainment & Sports Law http://www.tresl.net	\$40.00 / \$45.00
Texas Journal of Oil, Gas & Energy Law http://www.tjogel.org	\$30.00 / \$40.00
Manuals:	
<i>The Greenbook: Texas Rules of Form</i> 12th ed. ISBN 1-878674-08-0	
<i>Manual on Usage & Style</i> 11th ed. ISBN 1-878674-55-2	

To order, please contact:
 The University of Texas School of Law Publications
 727 E. Dean Keeton St.
 Austin, TX 78705 U.S.A.
 Publications@law.utexas.edu


ORDER ONLINE AT:
<http://www.texaslawpublications.com>



We Complete the Picture.

In 1932, Joe Christensen founded a company based on Value, Quality and Service. Joe Christensen, Inc. remains the most experienced Law Review printer in the country.

Our printing services bridge the gap between your editorial skills and the production of a high-quality publication. We ease the demands of your assignment by offering you the basis of our business—customer service.

Joe Christensen, Inc. 

1540 Adams Street
Lincoln, Nebraska 68521-1819
Phone: 1-800-228-5030
FAX: 402-476-3094
email: sales@christensen.com

Value

Quality

Service

Your Service Specialists

* * *

Texas Law Review

The Greenbook: Texas Rules of Form *Twelfth Edition*

A comprehensive guide for Texas citation, newly revised in 2010.

Texas Law Review Manual on Usage & Style *Twelfth Edition*

A pocket reference guide on style for all legal writing.

Newly revised and released in Fall 2011

**School of Law Publications
University of Texas at Austin
727 East Dean Keeton Street
Austin, Texas USA 78705
Fax: (512) 471-6988 Tel: (512) 232-1149
Order online: <http://www.utexas.edu/law/publications>**

